Claudia Angelini
Erik Bongcam-Rudloff
Adriano Decarli
Paola MV Rancoita
Stefano Rovetta (Eds)

Computational Intelligence Methods for Bioinformatics and Biostatistics

12th International Meeting on
Computational Intelligence Methods for Bioinformatics and Biostatistics

CNR Research Area, Naples, Italy
September 10-12, 2015

Conference Proceedings

CiBB2015

This volume contains the papers presented at CIBB 2015:
Twelfth international meeting on Computational Intelligence methods for Bioinformatics and Biostatistics

held on September 10-12, 2012 at CNR Research Area, Naples, Italy

This volume was edited by Claudia Angelini, Erik Bongcam-Rudloff, Adriano Decarli, Paola MV Rancoita, Stefano Rovetta (Eds)

ISBN: 9788890643798

Published by Department of Informatics, University of Salerno and Istituto pe r le Applicazioni del Calcolo CNR

Published on September 12, 2015, in Naples, Italy.

This conference was organized with the support of EasyChair conference system.

Our Main sponsors:

Istituto per le Applicazioni del Calcolo IAC-CNR
Istituto di Genetica e Biofisica IGB-CNR
Department of Clinical Sciences and Community Health, University of Milan
Gruppo Nazionale di Calcolo Scientifico GNCS-INDAM

Other Sponsors

Genomix4life S.r.l.
BMR Genomics S.r.l.
MM Biotech S.c.a.r.l.

Patronages

Bioinformatics ITalian Society (BITS)

CIBB HISTORY

From 2004 to 2007, CIBB had the format of a special session of larger conferences, namely, WIRN 2004 in Perugia, WILF 2005 in Crema, FLINS 2006 in Genoa, and WILF 2007 in Camogli. Given the great success of the special session at WILF 2007 that included 26 strongly rated papers, the Steering Committee decided to turn CIBB into an autonomous conference starting with the 2008 edition in Vietri. The following editions in Italian venues were held in Genoa (2009), Palermo (2010) and Gargnano (2011). Until 2012, CIBB meetings were held annually in Italy with an increasing number of participants. CIBB 2012 was
the first edition organized outside Italy, in Houston, while the 2013 edition was organized in Nice, France and 2014 was organized in Cambridge, United Kingdom .

A rigorous peer-review selection process has been applied every time to ultimately select the papers included in the program of the conference, in the post-conference Proceedings published by LNBI-LNCS book series by Springer Verlag, and in some cases, selected papers were published in special issues of well established international journals, such as BMC Bioinformatics.

CIBB 2014, Cambridge, United Kingdom
CIBB 2013, Nice, Italy
CIBB 2012, Huston, Italy
CIBB 2011, Gargnano sul Garda, Italy
CIBB 2010, Palermo, ItalyP
CIBB 2009, Genoa, Italy
CIBB 2008, Vietri sul Mare, Italy
CIBB 2007, Portofino, Italy
CIBB 2006, Genova, Italy
CIBB 2005, Crema, Italy
CIBB 2004, Perugia, Italy

CIBB 2015 OVERVIEW

CIBB (Computational Intelligence methods for Bioinformatics and Biostatistics) is a meeting with more than 10-year of history. Its main goal is to provide a forum open to researchers from different disciplines to present problems concerning computational techniques in bioinformatics, systems biology and medical informatics, to discuss cutting edge methodologies and accelerate life science discoveries. Following this tradition and roots, this year's meeting will bring together researchers from the international scientific community interested in this field to discuss the advancements and the future perspectives in bioinformatics and biostatistics. Applied biologists are also invited to participate in order to propose novel challenges aimed to have high impact for molecular biology and translational medicine. Location is the CNR Research Area Napoli 1, in Naples Italy. CNR Research Area Napoli 1 is the Biomedical and Biotechnological of the National Research Council in Naples.

Technical areas addressed by CIBB 2015 include, but are not limited to: High dimensional statistical analysis of omic data; Next generation sequencing bioinformatics; Multi-omic data integration; Methods for supervised and unsupervised learning; Prediction of protein structures; Methods for comparative genomics; Algorithms for molecular evolution and phylogenetic analysis; Mathematical modelling and simulation of biological systems; Systems and synthetic biology; Bio-molecular databases and data mining; Bio-medical text mining and imaging; Statistical methods for the analysis of clinical data; Methods for the visualization of high dimensional complex omic data; Software tools for bioinformatics.

The scientific program includes, besides some plenary talks, contributed papers that will be presented in plenary oral and poster sessions. Accepted papers will be published in the conference proceedings. A selection of papers presented at CIBB 2015 will be published a in post conference volume printed by an international publisher. Moreover, we are planning to publish the best papers in extended form in a special issue of BMC Bioinformatics.

Five special sessions are organized at CIBB 2015:

- *The EDGE, enhanced definition of genomic entities for systems biomedicine in oncology,*
- *New knowledge from old data: power of data analysis and integration methods*
- *Regularization methods for genomic data analysis*
- *Large-Scale and HPC data analysis in bioinformatics: intelligent methods for computational, systems and synthetic biology*
- *Multi Omic metabolic models and statistical Bioinformatics of adaptations and biological associations*

ISBN: 9788890643798

Published by Department of Informatics, University of Salerno and IAC-CNR.
Published on September 12, 2015, in Naples, Italy.

ORGANIZING INSTITUTIONS

CIBB 2015 is jointly organized by:
Istituto per le Applicazioni del Calcolo “Mauro Picone”, CNR
Istituto di Genetica e Biofisica, “Adriano Buzzati Traverso”, CNR

INNS International Neural Network Society SIG Bioinformatics
INNS International Neural Network Society SIG Bio-pattern
IEEE-CIS-BBCT Task forces on Neural Networks and Evolutionary Computation

General Chairs

Claudia Angelini, Istituto per le Applicazioni del Calcolo, Italy
Adriano Decarli, University of Milan, Italy
Erik Bongcam-Rudloff, Swedish University of Agricultural Sciences, Sweden.

Biostatistics Technical Chair

Paola MV Rancoita, Vita-Salute San Raffaele University, Italy

Bioinformatics Technical Chair

Stefano Rovetta, University of Genova, Italy

Special Sessions Chair

Franck Picard, CNRS LBBE, Lyon 1, France

Local Organizing Committee

Valerio Costa, Institute of Genetics and Biophysics, Italy
Italia De Feis, Istituto per le Applicazioni del Calcolo, Italy
Angelo Facchiano, Istituto di Scienze dell'Alimentazione, Italy

Publicity Chair

Francesco Masulli, University of Genova, Italy & Temple University, USA

Publication Chair

Riccardo Rizzo, Istituto di Calcolo e Reti ad Alte Prestazioni, Italy

Finance Chair

Elia M. Biganzoli, University of Milan

Administrative Secretary

Patrizia Montanaro, Istituto per le Applicazioni del Calcolo, Italy

STEERING COMMITTEE

Pierre Baldi, University of California, Irvine, CA, USA
Elia Biganzoli, University of Milan, Italy
Mariaclelia Di Serio, University Vita-Salute San Raffaele, Italy
Alexandru Floares, Oncological Institute Cluj-Napoca, Romania
Jon Garibaldi, University of Nottingham, United Kingdom
Nikola Kasabov, Auckland University of Technology, New Zealand
Francesco Masulli, University of Genova, Italy and Temple University, PA, USA
Leif Peterson, TMHRI, Houston, Texas, USA
Roberto Tagliaferri, University of Salerno, Italy

PROGRAM COMMITTEE

Fentaw Abegaz, University of Groningen, The Netherlands
Federico Ambrogi, University of Milano, Italy
Sansanee Auephanwiriyakul, Chiang Mai University, Thailand
Mario Cannataro, University "Magna Grecia" of Catanzaro, Italy
Hailin Chen, Qiagen, Inc., US
Davide Chicco, University of Toronto, Canada
Federica Cugnata, University Vita-Salute San Raffaele, Italy
Antonio Eleuteri, The Royal Liverpool and Broadgreen University Hospitals, UK
Enrico Formenti, University of Nice-Sophia Antipolis, France
Arief Gusnanto, University of Leeds, UK
Raffaele Giancarlo, University of Palermo, Italy
Javier Gonzalez, University of Sheffield, UK
Yin Hu, Sage Bionetworks, US
Pawel Labaj, BOKU Vienna, Austria
Paulo Lisboa, Liverpool John Moores University, UK
Giosue' Lo Bosco, University of Palermo, Italy
Anna Marabotti, Uviversity of Salerno, Italy
Giancarlo Mauri, University "Bicocca" of Milano, Italy
Luciano Milanesi, ITB-CNR, Italy
Marta Milo, University of Sheffield, UK
Paola Paci, IASI-CNR, Italy
Marianna Pensky, University of Central Florida, US
Davide Risso, University of California, Berkeley, US
Riccardo Rizzo, ICAR-CNR, Italy
Paolo Tieri, IAC-CNR, Italy
Maurizio Urso, ICAR-CNR, Palermo, Italy
Veronica Vinciotti, Brunel University, UK
Blaz Zupan, University of Ljubljana, Slovenia

CiBB2015

Invited Speakers

Michele Ceccarelli
Quatar Computing Research Institute, HBKU, Quatar
University of Sannio, Italy

Dario Greco
Finnish Institute of Occupational Health Systems, Finland

Dirk Husmeier
Glasgow University, UK

Wessel Van Wieringen
Vrije Universiteis, The Netherland

Cinzia Viroli
University of Bologna, Italy

Daniel Yekutieli
Tel Aviv University, Israel

Invited Speaker for the General Chairs

Erik Bongcam-Rudloff, SLU Swedish University of Agricultural Sciences, Sweden

Conference program
September 10, 2015

8:30-9:20 Registration
9:20-9:30: Welcome
9: 30-10:10 Invited Speaker Talk

Dario Greco,
Computational challenges in systems nanotoxicology

10:10-10:50 Contributed Regular Talks

David Causeur, Emeline Perthame and Ching-Fan Sheu.
Signal identification in ERP data by decorrelated Higher Criticism Thresholding

Pınar Kavak, Bekir Ergüner, Duran Üstek, Bayram Yüksel, Mahmut Şamil Sağıroğlu, Tunga Güngör and Can Alkan..
Improving genome assemblies using multi-platform sequence data

10:50-11:30 Coffee Break
11:30-11:50 Contributed Short Talks

Reem Alsrraj, Bassam Alkindy, Christophe Guyeux, Laurent Philippe and Jean-François Couchot.
Well-supported phylogenies using largest subsets of core-genes by discrete particle swarm optimization.

Daniela Evangelista, Mariano Avino, Kumar P. Tripathi and Mario R. Guarracino.
PrimatesDB: a functional resource on skeletal muscle tissue specific transcriptome of the Pan troglodytes.

11:50-12:00 Introduction to the Special Guest

12:00-13:00 Special Guest talk

Ada Yonath (Nobel Laureate in Chemistry)
Species-specific antibiotics and the microbiome

13:00-14:30 Lunch

Conference program
September 10, 2015

14: 30-15:10 Invited Speaker Talk

Daniel Yekutieli,
Bayesian FDR controlling test procedures.

15:10-15:50 Contributed Regular Talks

Annamaria Carissimo, Luisa Cutillo and Italia De Feis.
Validation Of Community Robustness

Monika Kurpas and Krzysztof Puszynski.
A simulation study of the relationship between replication stress detection pathway and the cell cycle.

15:50-16:10 Contributed Short Talks

Antonio Eleuteri.
A commentary on a censored regression estimator.

Kumar Parijat Tripathi, Sonali Chavan, Seetharaman Parashuraman, Marina Piccirillo, Sara Magliocca and Mario Guarracino.
Comparison of gene expression signature using rank based statistical inference

16:10-16:40 Coffee Break

16:40-18:20 Special Session *The EDGE, enhanced definition of genomic entities for systems biomedicine in oncology*

Christine Desmedt.
Dissecting the biological complexity of breast cancer.

Emanuela Guerra, Rossano Lattanzio, Marco Trerotola, Pasquale Simeone, Valeria Relli, Patrizia Querzoli, Enzo Bianchini, Domenico Angelucci, Giuseppe Pizzicannella, Laura Antolini, Andrea Telatin, Barbara Simionati, Mauro Piantelli and Saverio Alberti. *Transcriptomic analysis of the Trop-2 metastatic program.*

Romano Demicheli.
Cancer Images: from invading hordes to pseudo-organ structures.

Giuseppe Marano, Patrizia Boracchi, Elia M. Biganzoli.
Assessment of the robustness of Bayesian P-spline estimation techniques for prognostic assessment and prediction.

Michele Libutti.
New Techs and oncology in System Medicine.

18:20 Closing day

Conference program
September 11, 2015

8:30-9:00 Registration
9:00-9:40 Invited Speaker Talk

Wessel van Wieringen
Ridge estimation of multiple Gaussian graphical models: individually, simultaneously, and integratively.

9:40-10:40 Contributed Regular Talks

Davide Chicco and Marco Masseroli.
Validation Procedures for Predicted Gene Ontology Annotations

Riccardo Rizzo, Antonino Fiannaca, Massimo La Rosa and Alfonso Urso.
A deep learning approach to DNA sequence classification: first results.

Francesco Russo, Dario Righelli and Claudia Angelini.
A walking tour in Reproducible Research and Big Data Management with RNASeqGUI and R.

10:40-11:20 Coffee Break

11:30-13:00 Special Session New knowledge from old data: power of data analysis a nd integration methods

Guillaume Devailly and Anagha Joshi.
T*ranscription control in human cell types by systematic analysis of ChIP sequencing data from the ENCODE*.

Leen De Baets, Sofie Van Gassen, Tom Dhaene and Yvan Saeys.
Novel unsupervised learning methods for single cell data visualization and trajectory inference.

Panayotis Vlastaridis, Stephen G. Oliver, Yves Van de Peer and Grigorios Amoutzias.
Phosphoproteomics: A critical view through the bioinformatics lens.

Bruno Giotti and Tom Freeman.
Meta-analysis of human cell cycle-associated transcripts using published data.

Wilbert Sibanda.
D-Optimal Designs: Differences in HIV risk profiles between Gen X black women and entire population of black women attending antenatal clinics in South Africa.

13:00-14:30 Lunch

Conference program
September 11, 2015

14: 30-15:10 Invited Speaker Talk

Cinzia Viroli
Modeling overdispersion heterogeneity in differential expression analysis using mixtures

15:10-15:50 Special Session Regularization methods for genomic data analysis

Julien Chiquet (Invited speaker)
Fast tree inference with weighted fusion penalties

15:50-16:00 External short talk

Asli Ismihan Ozen
NGS data analysis with XploreRNA.

16:00-16:30 Coffee Break

16:30-17:50 Special Session Large-Scale and HPC data analysis in bioinformatics: intelligent methods for computational, systems and synthetic biology

Fabio Tordini, Ivan Merelli, Pietro Liò, Marco Aldinucci and Luciano Milanesi.
nuChaRt: embedding High Performance Computing in R for augmented DNA Exploration

Abhinandan Khan, Rajat Kumar Pal and Goutam Saha.
A novel Technique for reduction of false positives in predicted regulatory networks.

Andrea Tangherloni, Paolo Cazzaniga, Marco Nobile, Daniela Besozzi and Giancarlo Mauri. *Deterministic simulations of large-scale models of cellular processes accelerated on graphics processing units.*

Giacomo Paschina, Daniele D'Agostino, Federica Chiappori, Luca Ravelli and Ivan Merelli.
An Hadoop-based algorithm for clustering Protein Structures

17:50-18:30 Round table and discussion about next CIBB edition

STEERING COMMITTEE

18:30 Closing day

20:00-22:30 Social Dinner at Renzo e Lucia Restaurant

Conference program
September 12, 2015

8:30-9:00 Registration
9:00-9:40 Invited Speaker Talk

Dirk Husmeier
Bayesian inference of antigenic sites in viral evolution

9:40-10:40 Contributed Regular Talks

Paola Lecca and Angela Re.
Identifying modules in Biological network with WG-cluster.

Arianna Consiglio, Corrado Mencar, Giorgio Grillo and Sabino Liuni.
Managing NGS differential expression uncertainty with fuzzy sets.

Mounia Haddoud, Aïcha Mokhtari, Thierry Lecroq and Said Abdeddaim.
Supervised term weights biomedical text classification.

10:40-11:20 Coffee Break
11: 20-12:00 Invited Speaker Talk

Michele Ceccarelli
Integrative Molecular Analysis Across Adult Glioma: Novel Relationships Between Histological Subtypes and Molecular Signatures.

12: 00-12:40 Invited General Chair Talk

Erik Bongcam-Rudloff
Challenges in the annotation of NGS data.

12:40-14:10 Lunch

Conference program
September 12, 2015

14:10-15:00 Contributed short Talks

Emanuel Weitschek, Giulia Fiscon, Valerio Cestarelli, Paola Bertolazzi and Giovanni Felici.
LAF Barcoding: classifying DNA Barcode multi-locus sequences with feature vectors and supervised approaches.

Giosue' Lo Bosco and Dario La Neve.
Alignment free Dissimilarities for sequence classification.

Eugenio Del Prete, Diego d'Esposito, Maria Fiorella Mazzeo, Rosa Anna Siciliano and Angelo Facchiano.
A *workflow for the comparative analysis of MALDI-TOF mass spectrometric data in proteomics.*

Amit Dubey, Anna Marabotti, Pramod W. Ramteke and Angelo Facchiano.
Ethno-pharmacology Based In Silico Approach Tracing Chymase Inhibitors from Herbal Nutraceutical Resources.

Sebastian Daberdaku and Carlo Ferrari.
A voxel-based tool for protein surface representation.

15:00-15:50 Special Session Multi Omic metabolic models and statistical Bioinformatics of adaptations and biological associations (part I)

Claudio Angione, Sandra Pucciarelli, Barbara Simionati and Pietro Liò.
Measuring adaptation to extreme environments with a multi-omic approach.

Antonio Starcevic.
Proteome semantics – application of natural language processing to peptide mass fingerprinting.

Marco Fondi, Emanuele Bosi, Luana Presta, Pietro Liò and Renato Fani.
Modeling metabolic adaptation to cold shock and substrates switching.

15:50-16:20 Coffee Break

Conference program
September 12, 2015

16:20-17:10 Special Session Multi Omic metabolic models and statistical Bioinformatics of adaptations and biological associations (part II)

Başarbatu Can, Arda Durmaz and Osman Uğur Sezerman.
Comparative Analysis of Differentially Expressed Pathways in Mouse with ALS

Ivano Zara, Ilena Li Mura, Andrea Telatin and Barbara Simionati.
SNP-SHOT: integrating annotation sources for target enrichment experiments.

Federica Chiappori.
Psychrophilic protein modeling.

Pietro Tedesco, Fortunato Palma Esposito, Antonio Mondini, Glen Brodie, Renato Fani, Marcel Jaspars and Donatella de Pascale.
Antimicrobial compounds from Antarctic bacteria.

17:10-17:40 Closing Remarks

17:40 Closing CIBB2015

CIBB 2015 Invited talks

GENERAL CHAIRS

Claudia Angelini, IAC-CNR, Italy

Adriano Decarli, University of Milan, Italy

Erik Bongcam-Rudloff, SLU Swedish University of Agricultural Sciences, Sweden

Integrative Molecular Analysis Across Adult Glioma: Novel Relationships Between Histological Subtypes and Molecular Signatures.

Michele Ceccarelli[1,2,6], Floris P. Barthel[3,6], Tathiane M. Malta[4,6], Thais S. Sabedot[4,6], Stefano M. Pagnotta[1], Samreen Anjum[6],, Houtan Noushmehr[4], Antonio Iavarone[5], Roel G.W. Verhaak3[3] and LGG-GBM TCGA Analysis Working Group on behalf of TCGA Research Network

(1) Department of Science and Technology, University of Sannio, Benevento, Italy

(2) Qatar Computing Research Institute, HBKU, Qatar

(3) Department of Genomic Medicine, Department of Bioinformatics and Computational Biology, & Department of Neuro-Oncology, Department of Neurosurgery, Department of Pathology,
University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

(4) Department of Genetics, Ribeiro Preto Medical School - FMRP, University of So Paulo, Brazil

(5) Department of Neurology, Department of Pathology, Institute for Cancer Genetics, Department of Systems Biology and Biomedical Informatics,
Columbia University Medical Center, New York, NY, USA

(6) The authors contributed equally to this work

Keywords:

Abstract.

Gliomas represent approximately 30% of central nervous system tumors and 80An unbiased analysis across glioma grades and histologies that integrates all the possible molecular and genetic information has never been attempted. Therefore, the characterization of the molecular features that mark each of the specific Low Grade Glioma (LGG) and Glioblastoma (GBM) subgroup remains elusive. Most importantly, current analyses have not yet clarified the relationships between LGGs and highly malignant GBMs that share common genetic hallmarks such as IDH mutation or TERT promoter mutation status. Understanding these relationships is of critical importance in clinical management of gliomas and will be necessary to evolve to an objective genome-based clinical classification of these tumors. To address the above questions, we assembled a dataset comprising all TCGA newly diagnosed glioma consisting of 1,122 patients (516 LGG and 606 GBM), which have been analyzed using multiple molecular platforms including mRNA sequencing and expression arrays, DNA methylation arrays, exome sequencing, DNA copy number profiling arrays and targeted proteomics using reverse phase protein arrays. We address crucial technical challenges in analyzing this comprehensive dataset, including the integration of multiple available platforms and sources of data (e.g. multiple methylation and gene expression platforms) and have extended our analysis to pediatric gliomas and pilocytic astrocytoma to span the broader spectrum of glioma. We identified new glioma subgroups with distinct molecular and clinical features and may shed light on the mechanisms driving progression of LGG (WHO grades II and III) into full-blown GBM (WHO grade IV).

Challenges in the annotation of NGS data

Erik Bongcam-Rudloff[1]

(1) Swedish University of Agricultural Sciences, Sweden.

Keywords: NGS, ,

Abstract.

Computational Challenges in Systems Nanotoxicology

Dario Greco[(1)]

(1) Unit of Systems Toxicology and Nanosafety Research Centre
Finnish Institute of Occupational Health Systems, Finland

Keywords: Engineered nanomaterials, Nanotoxicology, Data integration, Feature selection

Abstract. Engineered nanomaterials (ENM) are incorporated in many consumer products and human exposure increases as the development of new ENM proceeds. However, the features that make ENM desirable in various applications have also the potential to alter the biological properties impacting their safety. The novel field of systems nanotoxicology aims at studying the nano-bio interactions at multiple levels by comprehensive molecular profiling of the exposed cells, tissues and organisms. The aim is to model the effect of ENMs taking into account the intrinsic physico-chemical characteristics of the materials in order to help the development of new safe-by-design ENMs.

In the context of the EU FP7 project NANOSOLUTIONS, we are coordinating the systems biology work package with the aim of developing a computational classifier able to predict the safety of ENMs. In order to succeed in this task, we need to address multiple computational challenges related to data integration feature selection, exploration of the solution space, and optimization of the predictive model. In the context of other projects, we are also inferring the gene regulatory networks that underlie the specific responses to ENM exposure, both in vitro and in vivo. In addition, we try to bridge the fields of nanotoxicology and nanomedicine, by systematically comparing the complex molecular responses to ENM exposure with those specific to drug treatments in vitro.

Bayesian inference of antigenic sites in viral evolution

Dirk Husmeier[(1)]

(1) University of Glasgow, United Kingdom

Keywords: Bayesian inference, MCMC, Vaccines

Abstract. Understanding how closely related viruses offer protection against emerging strains is vital for creating effective vaccines. For many viruses, in particular Foot-and-Mouth disease virus (FMDV) where multiple serotypes often co-circulate, testing large numbers of vaccines can be infeasible. Therefore the development of an in silico predictor of cross-protection between strains is important to help optimise vaccine choice. This is especially the case in sub-Saharan Africa where several South African Territories (SAT) serotypes are endemic. I will present a sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution (SABRE), which can account for the experimental variability in the data. The method uses spike and slab priors to allow the model to predict antigenic variability and identify sites in the viral protein which are important for the neutralisation of the virus. Using the SABRE method we are able to identify a number of key antigenic sites within some of the SAT serotypes, as well as provide estimates of significant changes in the evolutionary history of the serotypes. I will show how our method outperforms state-of-the-art mixed effects models and demonstrate how changing the Markov chain Monte Carlo (MCMC) proposal method used for the inclusion of variables can offer significant reductions in computational requirements. This is joint work with Vinny Davies, Richard Reeves, William Harvey and Francois Maree.

Ridge estimation of multiple Gaussian graphical models: individually, simultaneously, and integratively.

Wessel Van Wieringen[(1)]

(1) Vrije Universiteit, The Netherland

Keywords: Gaussian Graphical models, gene-gene interaction network, penalized regression

Abstract.

Molecular biology aims to understand the molecular processes that occur in the cell. That is, which molecules present in the cell interact, and how is this coordinated? For many cellular process, it is unknown which genes play what role. A valuable source of information to uncover gene-gene interactions are (onco)genomics studies. Such studies comprise samples from n individuals with, e.g., cancer of the same tissue. Each sample is interrogated molecularly and the expression levels of many (p) genes are measured simultaneously. From these high-dimensional omics data the gene-gene interaction network may be unravelled when the presence (absence) of a gene-gene interaction is operationalized as a conditional (in)dependency between the corresponding gene pair. Then, under the assumption of multivariate normality, the gene-gene interactions correspond to zero's in the precision matrix (which are proportional to the partial correlations).

When dealing with high-dimensional data, the sample covariance matrix is singular and the sample precision matrix is not defined. But even if $p < n$ and p approaches n, the sample precision matrix yields inflated partial correlations. Both situations require a form of regularization to obtain a well-behaved estimate of the precision matrix, and consequently of the gene-gene interaction network. To this end we study ridge estimation of the precision matrix in the high-dimensional setting. We illustrate its use on the reconstruction of the gene-gene interaction network from oncogenomics data.

Often the samples included in oncogenomics studies originate from different clinical groups. Interest then concentrates on differences in the gene-gene interaction network among the groups. To identify those the aforementioned ridge estimation procedure is extended to the multi-group case. Its estimation employs a fused ridge penalty, which penalizes not only the absolute size of the precision elements but also the difference among the group precisions.

Time allowing, we point out how the proposed ridge estimation framework may learn dynamic networks from time-course genomics experiments.

Modeling overdispersion heterogeneity in differential expression analysis using mixtures.

Cinzia Viroli[(1)]

(1) University of Bologna, Italy

Keywords: RNA-seq, Mixture models, Differential gene expression

Abstract. In the last 15 years, the development of massively parallel sequencing platforms for mapping the genome has completely revolutionized the way of thinking and studying gene expression patterns. The recent Next-Generation Sequencing (NGS) technologies allow to simultaneously investigate thousands of features within a single reliable and cost-effective experiment, thus offering a challenging way to enhance our understanding of how genetic differences affect health and disease. The NGS data are read counts and they are commonly analyzed by the Negative Binomial probabilistic model. A relevant issue associated with this probabilistic framework is the reliable estimation of the overdispersion parameter, reinforced by the limited number of replicates generally observable for each gene. Many strategies have been proposed to estimate this parameter, but when differential analysis is the purpose, they often result in procedures based on plug-in estimates, and the discrepancy between the estimation framework and the testing framework can lead to uncontrolled type-I errors. In this talk, a mixture model framework is presented. It allows each gene to share information with other genes that exhibit similar variability. Then a consistent statistical test is developed for differential expression analysis. It is shown through a wide simulation study that the proposed method improves the sensitivity of detecting differentially expressed genes with respect to the common procedures, since it reaches the nominal value for the type-I error, while keeping elevate discriminative power between differentially and not differentially expressed genes.

Bayesian FDR controlling test procedures.

Daniel Yekutieli[(1)]

(1) Tel Aviv University, Israel.

Keywords: FDR, Bayesian approaches,

Abstract. Bayesian FDR controlling procedures for the two-group model yield almost identical results to the adaptive Benjamini-Hochberg procedure. In my talk I show that this is not necessarily the case in more complicated testing problems. I will discuss joint work with Ruth Heller on establishing replicability in multiple genomewide association studies, a difficult testing problem that involves testing complex null hypotheses and for which there is no natural test statistic. I will present our empirical Bayes methodology. I will explain the relation between the Bayes approach and the BH procedure and why, in this example, Bayesian FDR controlling methods offer substantially more power (at the same FDR level) than the BH procedure.

CIBB 2015 Regular contributed talks

BIOSTATISTICS TECHNICAL CHAIR

Paola MV Rancoita, University Vita-Salute San Raffaele, Italy

BIOINFORMATICS TECHNICAL CHAIR

Stefano Rovetta, University of Genova, Italy

Validation Of Community Robustness

Annamaria Carissimo(1), Luisa Cutillo(2), Italia De Feis(3)

(1) Telethon Institute of Genetics and Medicine, Pozzuoli, Italy
bioinformatics core, carissimo@tigem.it

(2)University of Naples "Parthenope"
DISAQ, luisa.curtillo@uniparthenope.it

(2)Consiglio Nazionale delle Ricerche
IAC, Napoli, i.defeis@iac.cnr.it

Keywords: Community Detection, Networks, Variation of Information, Multiple Testing.

Abstract.

The large amount of work on community detection and its applications leaves unaddressed one important question: the statistical validation of the results. In this paper we present a methodology able to clearly detect the truly significance of the communities identified by some technique, permitting to discard those that could be merely the consequence of edge positions in the network. Given a community detection method and a network of interest, our procedure examines the stability of the partition recovered against random perturbations of the original graph structure. To address this issue, we specify a perturbation strategy and a null model to build a stringent statistical test on a special measure of clustering distance, namely Variation of Information. The test determines if the obtained clustering departs significantly from the null model, hence strongly supporting the robustness against perturbation of the algorithm that identified the community structure. We show the results obtained with the proposed technique on simulated and real datasets.

1 Scientific Background

Networks are mathematical representation of interactions among the components of a system and can be modelled by graphs. A graph G=(V,E) consists of a collection of vertices V, corresponding to the individual units of the observed system, and a collection of edges E, indicating some relation between pairs of vertices.

Graphs modelling real systems, i.e. social, biological, and technological networks, display non trivial topological features. Indeed they present big inhomogeneities, have a broad degree distribution, with a tail that often follows a power law, i.e. many vertices with low degree coexist with some vertices with large degree, and the distribution of edges is locally inhomogeneous, with high concentrations of edges within special groups of vertices and low concentrations between these groups. These properties define a complex network. In the study of complex networks, a network is said to have a community structure if the vertices can be divided in g groups (potentially overlapping), such that nodes belonging to the same group are densely connected and the number of edges between nodes of different groups is minimal.

The problem of community detection (graph partitioning) consists of finding the community structure and has been widely studied by researchers in a variety of fields, including statistics, physics, biology, social and computer science in the last 15 years. Finding communities within an arbitrary complex network can be a computationally difficult task. The number of communities, if any, within the network is typically unknown

and the communities are often of unequal size and/or density. Despite these difficulties, however, several methods for community finding have been developed and employed with varying levels of success, see [3], [5], [6], [7], [12] and [14] for reviews.

Although the huge work developed for community detection and its applications, the question of the significance of results still remains open. The problem is the robustness of the recovered partition and its validation against randomness.

In this paper we present a methodology able to clearly detect the truly significance of the communities identified by some technique, permitting to discard those that could be merely the consequence of edge positions in the network. Given a community detection method and a network of interest, our procedure examines the stability of the partition recovered against random perturbations of the original graph structure.

2 Materials and Methods

Variation of Information (VI) is an information theoretic criterion for comparing two partitions, or clusterings, of the same data set [13]. It is a metric and measures the amount of information lost and gained in changing from clustering $\mathcal{C}$ to clustering $\mathcal{C}'$. The criterion makes no assumptions about how the clusterings were generated and applies to both soft and hard clusterings.

Given a dataset D and two clusterings $\mathcal{C}$ and $\mathcal{C}'$ of D, with K and K' non empty clusters, respectively, VI is defined as

$$VI\left(\mathcal{C},\mathcal{C}'\right) = H\left(\mathcal{C}\right) + H\left(\mathcal{C}'\right) - 2I\left(\mathcal{C},\mathcal{C}'\right) \tag{1}$$

where $H\left(\mathcal{C}\right)$ is the entropy associated with clustering $\mathcal{C}$

$$H\left(\mathcal{C}\right) = -\sum_{k=1}^{K} P(k)\log P(k), \tag{2}$$

and $I\left(\mathcal{C},\mathcal{C}'\right)$ is the mutual information between $\mathcal{C}$ and $\mathcal{C}'$, i.e the information that one clustering has about the other

$$I\left(\mathcal{C},\mathcal{C}'\right) = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P\left(k,k'\right)\log\frac{P\left(k,k'\right)}{P\left(k\right)P\left(k'\right)}. \tag{3}$$

$P\left(k\right)$ is the probability of a point being in cluster C_k and $P\left(k,k'\right)$ is the probability that a point belongs to C_k in clustering $\mathcal{C}$ and to $C_{k'}$ in $\mathcal{C}'$.

Another equivalent expression for VI is

$$VI\left(\mathcal{C},\mathcal{C}'\right) = H\left(\mathcal{C}|\mathcal{C}'\right) + H\left(\mathcal{C}'|\mathcal{C}\right). \tag{4}$$

The first term measures the amount of information about $\mathcal{C}$ that we loose, while the second measures the amount of information about $\mathcal{C}'$ that we have to gain, when going from clustering $\mathcal{C}$ to clustering $\mathcal{C}'$.

VI metric is the basis of the hypothesis testing procedure we propose to establish the statistical significance of a recovered community structure in a complex network. Our original idea is to generate two different curves based on the VI measure and to statistically test their difference. The first curve VIc would be obtained computing VI between the partition of our original network and the partition of different perturbed version of our original network. The second curve VIc_{random} would be obtained computing VI between the partition of a null random network and the partition of different perturbed version of such null network. Mimicking the approach proposed by [11] and [9], we restrict our perturbed networks to having the same numbers of vertices and edges as the original unperturbed network, hence only the positions of the edges are

perturbed. Moreover, we expect that a network perturbed only a small amount has just a few edges moved in different communities, while a maximally perturbed network produces completely random clusters. Our simplified version of the perturbation strategy consists in randomly permuting a percentage p of edge from the original graph. Again a null percentage of permutation $p = 0$ corresponds to the original unperturbed graph, while $p = 1$ corresponds to the maximal perturbation level. The VIc and $VIc_{random}$are considered curves as functions of the percentage of perturbation p statistically and a statistical test is performed to asses the difference between the two curves. This step is achieved by a time series approach, considering the percentage of perturbation p as time point; infact its variation from 0 to 1 induces an intrinsic order to the data structure as in temporal data, indeed we expect that the VI of the perturbed network grows with the perturbation level. Moreover we generate many perturbed graphs (i.e. 10) for each different level of p and these are considered as replicates per time points in our strategy. The null model, that is the starting Network related to the VIc_{random} curve, is generated via the vl method, a method to generate random graphs having a prescribed degree sequence [16].

We reformulate the testing problem

$$H_0 : \mathrm{VIc} = \mathrm{VIc}_{\mathrm{random}}$$

as

$$H_0 : \log_2 \frac{\mathrm{VI}}{\mathrm{VI}_{\mathrm{random}}} = 0$$

and this permits us to take advantage of two analysis tools set up for time course microarray data, namely Bayesian Analysis for Time Series (BATS) [1], [2] and Gaussian Process regression (GP) [10], whose aim is to identify differentially expressed genes in a one-sample time-course microarray experiment.

3 Results

In order to provide an example of our analysis workflow, we selected a publicly available biological dataset. The biological dataset considered is a protein- protein interaction network. The S. cerevisiae proteinprotein interaction network we investigate has 1870 proteins as nodes, connected by 2240 identified direct physical interactions, and is derived from combined, non-overlapping data, obtained mostly by systematic two hybrid analyses [8]. In figure 1 we show the degree distribution over the network. As you can see this protein-protein interaction network belongs to the family of scale free networks. In Figure 2 we depict the two VI curves corresponding respectively to the null model (in red) and to the real data (in blue) retrieved with our method using $infomap$ [15] as a community detection strategy. It is easy to see that these curves are well separated but we need a statistical validation of this evidence. In this case indeed $BATS$ testing methods yields a Bayes Factor $BF >> 10$ and GP testing method yelds a $p-value << 0.05$, both supporting the hypothesis that the originally chosen community detection algorithm is robust against the random. Moreover, using the same dataset, we compared the performance of $infomap$ to another published community detection algorithm $fastgreedy$ [4]. The corresponding VI curves plotted in Figure 3 show that the community structure found by the algorithm $infomap$ [15] is more robust than that found by $fastgreedy$, infact the variation of information corresponding to $infomap$ increases slower as a function of the perturbation level. Also in this case both $BATS$ and GP reveal statistical significant difference.

4 Conclusion

In this paper we have described a new algorithm to validate any network clustering, indeed it can be applied to any community structure with or without overlapping

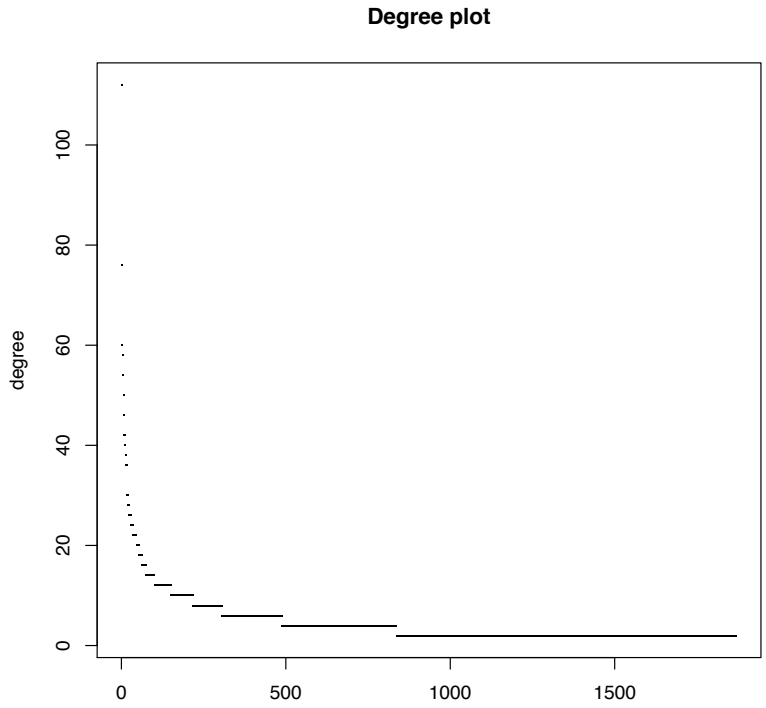


Figure 1: protein-protenin network degree. For each node in the plot the corresponding degree is plotted. The nodes are sorted in degree descending order.

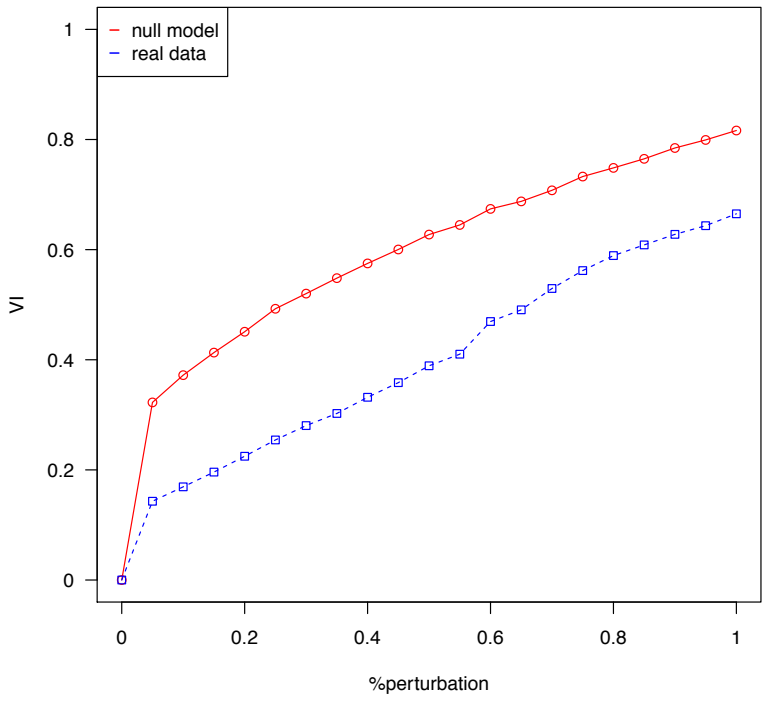


Figure 2: protein-protenin network VI curves

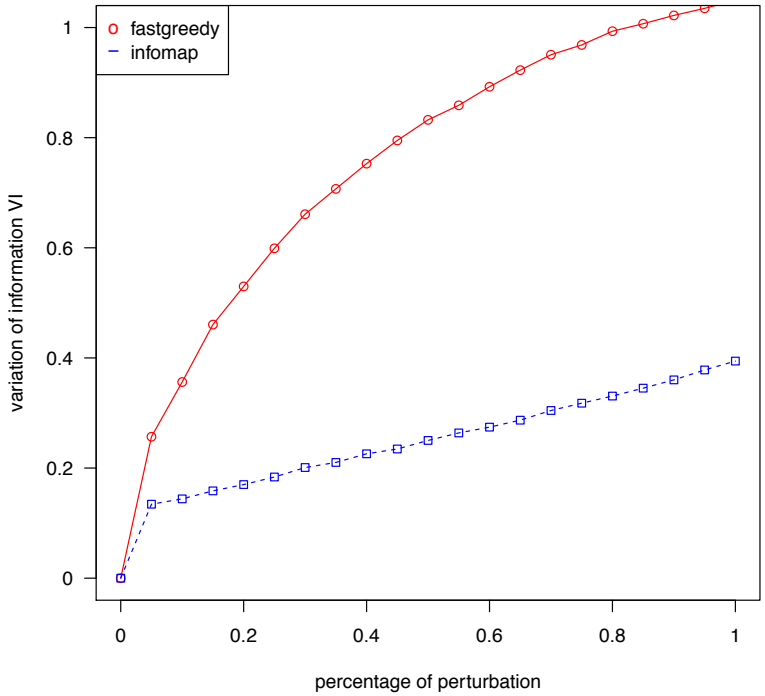


Figure 3: protein-protenin network VI curves

communities. This is a problem long studied in computer science, applied mathematics, and the social sciences, but it has lacked a satisfactory solution. We believe the method described here give such a solution. We address the problem of understanding when communities found in a network can be considered releable, and not the result of randomness in network structure. Our method assumes the Variation of Information as a measure of robustness under perturbations. The application to the real example network, described in the previous section, shows that our method clearly identifies strong community structures.

References

[1] ANGELINI C., DE CANDITIIS D., MUTARELLI M. AND PENSKY M. (2007). A Bayesian approach to estimation and testing in time-course microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **6:Article24**.

[2] ANGELINI C., CUTILLO L., DE CANDITIIS D., MUTARELLI M. AND PENSKY M. (2008). BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. *BMC Bioinformatics* **9:415**.

[3] COSCIA M., GIANNOTTI F. AND PEDRESCHI D. (2011). A Classification for Community Discovery Methods in Complex Networks. *Statistical Analy Data Mining* **4** 512–546.

[4] CLAUSET A., NEWMAN M.E.J. AND MOORE C. (2004). Finding community structure in very large networks. *Phys. Rev.* E **70**, 066111.

[5] FORTUNATO S. (2010). Community detection in graphs. *Phys. Rep.* 75–174 **486**.

[6] GOLDENBERG A., ZHENG A. X., SE FIENBERG S. E. AND AIROLDI E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning* **2** 129–233.

[7] HARENBERG S., BELLO G., GJELTEMA L., RANSHOUS S., HARLALKA J., SEAY R., PADMANABHAN K. AND SAMATOVA N. (2014). Community detection in large-scale networks: a survey and empirical evaluation. *WIREs Comput Stat* **6** 426–439.

[8] JEONG H., MASON S.P., BARABSI , A.-L.AND OLTVAI2 Z.N. Lethality and centrality in protein networks *Nature* **411, 41-42**

[9] CUTILLO L., CARISSIMO A. AND FIGINI S. (2012). Network Selection: A Method for Ranked Lists Selection. *PLoS ONE* **7(8):e43678**.

[10] KALAITZIS A. A. AND LAWRENCE N. D. (2011). A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression. *BMC Bioinformatics*

[11] KARRER B., LEVINA E. AND NEWMAN M. E. J. (2008). Robustness of community structure in networks. *Phys. Rev. E*

[12] KOLACYZK E. D. (2009). *Statistical Analysis of Network Models.* Springer, New York.

[13] MEILĂ M. (2007). Comparing clusterings – an information based distance. *J. Multivariate Anal.* **98** 873–895.

[14] PORTER M. A., ONNELA J. -P. AND MUCHA P. J. (2009). Communities in Networks. *Not. Amer. Math. Soc.* 1082–1097. **56**.

[15] ROSVALL M. AND BERGSTROM C. T. (2008).
Maps of random walks on complex networks reveal community structure. *PNAS* **105** (4) 1118-1123

[16] VIGER F. AND LATAPY M. (2005). Efficient and simple generation of random simple connected graphs with prescribed degree sequence. *COCOON*

Signal identification in ERP data by decorrelated Higher Criticism Thresholding

David Causeur[(1)], Émeline Perthame[(1)] and Ching-Fan Sheu[(2)]

(1) IRMAR UMR 6625 CNRS
Agrocampus Ouest, 65 rue de St-Brieuc CS 84215 35042 Rennes cedex, France,
david.causeur@agrocampus-ouest.fr

(2) National Cheng-Kung University
Institute of Education, 1 University Road, Tainan 701, Taiwan,
csheu@mail.ncku.edu.tw

Keywords: Correlated noise, Event-Related Potentials, High dimension, Higher Criticism, Signal identification.

Abstract. Event-related potentials (ERPs) are recordings of electrical activity along the scalp time-locked to perceptual, motor and cognitive events. Because significant association between ERPs and behavioral (or experimental) variables of interest are often rare, occurring only in brief moments during trials, and weak, relative to the huge between-subject variability, identification of ERP signals poses major challenges to statistical analysis. In this 'rare-and-weak' paradigm, the Higher Criticism method was shown in a number of recent papers to be optimal to determine signal detection threshold.

However, ERP time dependence exhibits a block pattern suggesting strong local and long-range autocorrelation components which violates the mild dependence assumption under which signal identification can be achieved efficiently. In high throughput settings, the detrimental effects of dependence on the accuracy of signal identification has indeed been widely known and a variety of decorrelation approaches have been developed to counter them. The presentation first highlights the impact of dependence in terms of instability of feature selection by Higher Criticism Thresholding. A second objective is to revisit the above issue using a flexible factor modeling for the covariance.

This framework introduces latent components of dependence, whose maximum-likelihood estimation enable decorrelation of the process of test statistics. In high-dimensional settings, the present method, and variants introducing a ℓ_1-penalized estimation of the inverse covariance of the process of test statistics, are compared to recent other decorrelation approaches either based on a shrinkage estimation of the inverse covariance or on its Cholesly decomposition.

1 Scientific Background

High-throughput instrumental data such as event-related potentials and functional magnetic resonance imaging (fMRI) data have increasingly become common in psychological research. The former provides high temporal resolution to chart the time course of mental processes, whereas the latter implicates spatial areas in the brain that might be responsible for experimental effects. With the routine collection of massive amounts of data from ERP or fMRI studies, researchers must face the challenge of signal identification: in shifting, simultaneously, through thousands or tens of thousands of tests for significant relationships with a response variable, a balance must be struck between keeping a low false positive error rate while maintaining sufficient power for correct signal identification. How to achieve this objective for ERP data exhibiting arbitrarily strong temporal dependence is the focus of the present paper.

Searching for time intervals of non-zero signals in ERP data can be viewed as a signal identification issue in the 'Rare and Weak' (RW) paradigm introduced by [5]. Large-scale significance analysis of ERP data is indeed based on a $m-$vector T of test statistics $T = (T_{t_1}, \ldots, T_{t_m})'$ where m is the number of time frames, for the collection of corresponding null hypotheses H_{0,t_i} of no association between the ERP measured at time t_i and the response variable. The RW setup is defined in [5] as the following sparse normal mixture model for T: for all t,

$$T_t \sim (1-\varepsilon)\mathcal{N}(0,1) + \varepsilon\mathcal{N}(\delta,1),$$

where the mixing parameter $0 \leq \varepsilon \leq 1$ is the proportion of non-null features and $\delta \geq 0$ is the signal amplitude. Note that the normality assumption introduced above holds for most ERP studies in which the tests for the association between the ERPs and the response variable is handled by t-tests for the significance of a single parameter. The alternative parameterization $\beta_\varepsilon = -\log(\varepsilon)/\log(m),\ r_\delta = (\delta^2/2)/log(m)$ is often preferred because it maps both the sparsity parameter β_ε and the amplitude parameter r_δ into $[0;1]$, if we observe that the expectation of the maximum test statistics under the null is bounded by $\sqrt{2\log m}$. Sparsity of the signal is characterized by $1/2 \leq \beta_\varepsilon \leq 1$ and weakness by $r_\delta < 1$ (see [10] for details).

The former RW assumption provides a simple yet insightful framework for the study of procedures, whose aim is to detect the presence of a nonzero signal. Indeed, closed-form theoretical detection bounds can be derived analytically and [5] demonstrates that HCT achieves the theoretically optimal decision limits. In the more challenging signal identification issue, aiming at the selection of non-null features for classification or prediction, [6] also demonstrates the superiority of HCT with respect to FDR-controlling multiple testing procedures.

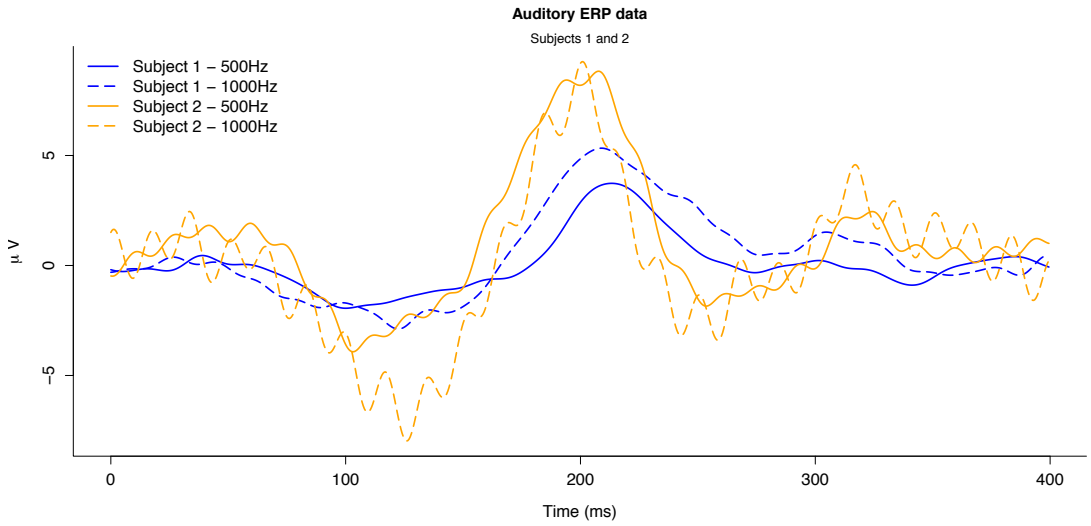
As reported in [3], the pronounced auto-correlation observed in ERP data can however induce a long-range regularity for the test statistics, resulting in spuriously low p-values outside of the support of the signal, which in turn can result in a misidentification of the non-null features. This instability of p-values'ranking due to dependence is also reported in many papers dealing with the impact of dependence on significance analysis of highly dimensional genomic data (see for example [7, 1, 11]). Equivalently, [9] reports that the theoretical detection bounds derived in the RW framework are markedly modified by a strong dependence among the test statistics. Therefore, [9] proposes to extend the RW framework as follows:

$$T = \delta + T_0,$$

where δ is a $m-$vector of signal amplitudes, in which a small proportion ε is non-zero and $T_0 \sim \mathcal{N}(0;\Sigma)$. If U is the inverse of the Cholesky factorization of Σ, namely $U\Sigma U' = I$, [9] introduces the so-called innovated HCT (iHCT) as the HCT procedure applied on the uncorrelated vector of innovations $UT = U\delta + UT_0$ and shows that iHCT restores the effectiveness of the HCT procedure in situations of strong dependence.

Correspondingly, in the feature selection issue for supervised classification in the Linear Discriminant Analysis (LDA) context, [1] shows that the HCT procedure is improved by replacing the z-scores T by correlation-adjusted z-scores $T^* = \Sigma^{-1/2}T$, where the inverse-square root of Σ is deduced from a James-Stein shrinkage estimator of Σ. As in [7, 11], we propose an alternative approach of innovated HCT based on a flexible factor model for Σ. As shown in [3], the complex dependence pattern observed in the correlation structure of test statistics derived from ERP data can be well approximated using the factor decomposition of Σ, often with a moderate number of factors. Moreover, it provides simple and efficient algebraic tools to derive $\Sigma^{-1/2}$. A Cyclic-Coordinate Descent (CCD) algorithm is also presented for a sparse ℓ_1- penalized estimation of $\Sigma^{-1/2}$.

Figure 1: ERP curves for subjects 1 (blue lines) and 2 (orange lines) of the auditory oddball experiment in conditions Hz500 (plain) and Hz1000 (dashed)



2 Materials and Methods

In ERP studies, perhaps the most commonly used experimental task is the oddball paradigm ([12]). In this paradigm, typically two classes of stimuli are presented, one occurring frequently (standard) and the other occurring infrequently (target). The subject is required to distinguish between the two stimuli and to respond to the stimuli that are designated as targets.

An auditory ERP study was performed at Kaohshung Medical University in Taiwan, providing an illustrative data set for the present investigation. The task uses two pure tones of 500 Hz and 1,000 Hz. The former is presented 120 out of 150 trials, whereas the latter (target) is presented only for 30 trials. The order of tone presentation is random and the subject is asked to (silently) count the number of targets. At each of 4 electrode locations (`FZ`, `C3`, `C4`, & `O1`), ERP waveform was obtained from each of the two tone conditions. For each of the $n = 15$ participants, the ERP curve begins at -100 milliseconds (ms) and terminates at 400 milliseconds (ms) with two records per 1 ms. The stimulus onset is at 0 ms. For subsequent analysis, only the ERPs from the electrode location FZ will be used.

[14] and many other studies have demonstrated that an ERP waveform across the parieto-central area of the skull is usually observed around 300 ms (the so-called P300 component) and is larger after the target event. The question to be addressed is whether it is possible to select time points at which ERP features can reliably detect which one of the two tones was presented to the subject and whether these time points are indeed around 300 ms as expected. This verification is of fundamental importance if the P300 component is to be considered as an electrophysiological marker for further assessment of psychiatric and neurological disorders.

It is conjectured that brain activations would be different over different time points depending on whether participants listen tones of 500Hz or 1000Hz. The data consist of $m = 800$ time points measured for $n = 15$ subjects and $J = 2$ conditions. Figure 1 shows ERP curves for subjects 1 (solid line) and 2 (dashed line) for condition Hz500 (grey) and Hz1000 (black) as an example. This figure illustrates the large variability among subjects.

For the ERP measurement Y_{jkt}, at time t, on subject j, $j = 1, \ldots, n$, in condition k, $k = 1$ for 500 Hz and $k = 2$ for 1000 Hz,

$$Y_{jkt} \quad = \quad \mu_t + \alpha_{jt} + \beta_{kt} + \varepsilon_{jkt}, \tag{1}$$

where $\alpha_t = (\alpha_{1t}, \ldots, \alpha_{nt})$ stands for the subject effect, with $\sum_j \alpha_{jt} = 0$ and $\beta_t = (\beta_{1t}, \beta_{2t})$ for the group effect. The condition 500hz being the reference, we set $\beta_{1t} = 0$, so that $t \mapsto \beta_{2t}$ is the difference curve. ε_{jkt} is the random error term, normally distributed with mean 0 and standard deviation σ_t.

In most similar situations, no dependence is assumed among the residual errors ε_{jkt}: the random vector $\varepsilon_{jk} = (\varepsilon_{jk,t_1}, \varepsilon_{jk,t_2}, \ldots, \varepsilon_{jk,t_m})'$ is assumed to be normally distributed with mean 0 and variance $D_\sigma = \text{diag}(\sigma^2_{t_1}, \sigma^2_{t_2}, \ldots, \sigma^2_{t_m})$, where diag(.) stands for the matrix operator which transforms a m-vector into the $m \times m$ diagonal matrix whose diagonal elements are given by the vector. For ERP data, the independence assumption is relaxed to account for time-dependence: $\text{Var}(\varepsilon) = \Sigma = D_\sigma^{1/2} R D_\sigma^{1/2}$, where R is a $m \times m$ residual correlation matrix.

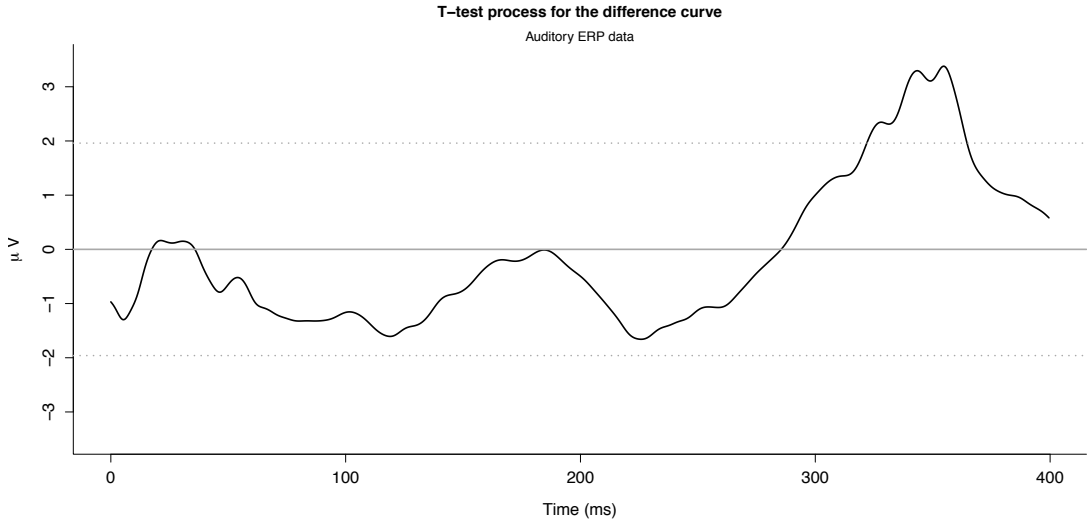


Figure 2: t-tests for the significance of β_{2t} at channel FZ. The dashed grey lines gives the 2.5th and 97.5th quantiles of the null distribution.

At each time t, the null hypothesis which is tested is $H_0^{(t)} : \beta_{2t} = 0$. The corresponding t-test process at channel FZ is displayed in Figure 2. The curve in Figure 2 shows a strong regularity which is not consistent with the expected profile of a process of independently distributed Student variables. This suggests a strong time-dependence among tests, which is known to affect the joint null distribution of test statistics. This strong auto-correlation is confirmed by Figure 3 in which the left panel is a histogram of the residual correlations of model 1 at electrode FZ and the right panel is an image plot of the residual correlation matrix. The histogram shows that a large proportion of correlations are strongly positive. The image plot shows a dependence pattern structured over time with an obvious auto-correlation component generating a large density of correlations close to 1 along the diagonal. The dependence pattern appears to be more complex than just an autoregressive structure of order one, with intervals of highly intercorrelated time points and an increasing lag-1 auto-correlation over time. The same kind of dependence is observed at the other electrodes.

The dependence structure of the $m-$vector of test statistics is directly inherited from the correlation among the residuals of model 1: $\text{Cov}(T_t, T_{t'}) = r_{tt'}$. Therefore, we propose the following RW framework for the $m-$vector of test statistics:

$$T \;=\; \delta + T_0,$$

where δ is a $m-$vector of signal amplitudes, in which a small proportion ε is non-zero and $T_0 \sim \mathcal{N}(0; \Sigma = \Psi + BB')$, where Ψ is a $m \times m$ diagonal matrix of specific

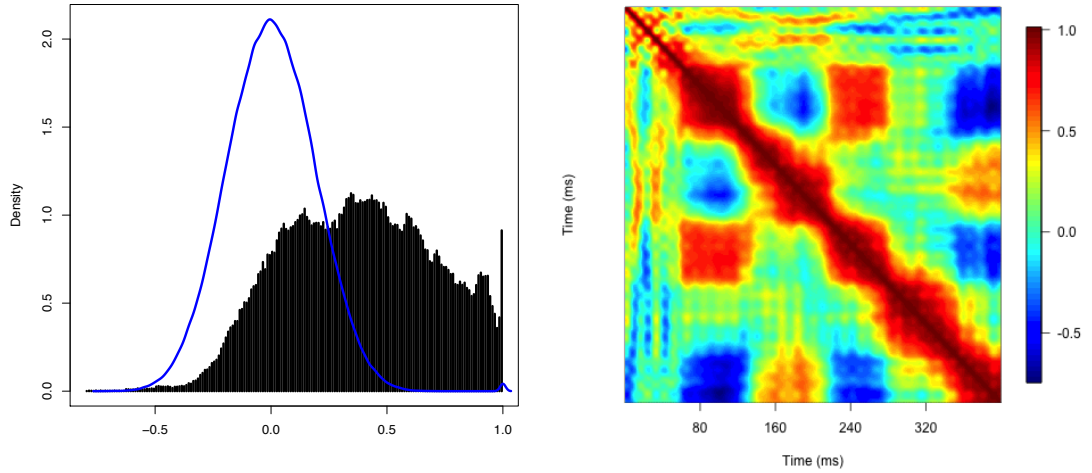


Figure 3: Left panel: histogram of correlations among residuals of model (1) at channel FZ over time. Right panel: image plot of the time correlations among residuals.

variances whose diagonal elements ψ_t^2 are in $[0;1]$ and B is a $m \times q$ matrix of factor loadings, with, for all t, $||b_t = (b_{t,1}, \ldots, b_{t,q})'||^2 = \sum_{l=1}^{q} b_{tl}^2 = 1 - \psi_t^2$. Note that, using the following new parameterization:

$$\begin{aligned} \varphi &= \Psi^{-\frac{1}{2}}, \\ \theta &= \Psi^{-\frac{1}{2}} B (I_q + B'\Psi^{-1}B)^{-\frac{1}{2}}. \end{aligned}$$

It is straightforward checked that $\Sigma^{-1} = \varphi(I_m - \theta\theta')\varphi$ has also a factor structure. Therefore, the maximum-likelihood (ML) estimation of Ψ and B, which can be derived in high-dimension using the EM algorithm presented in [7], also gives an estimator for the factor parameter of Σ^{-1} and in turn $\Sigma^{-1/2}$. Indeed,

$$\Sigma^{-1/2} = (I_m - U[(I_q + [I_q + D^2]^{1/2})^{-1} + I_q]^{-1}U')\Psi^{-1/2},$$

where U and D are deduced from the singular value decomposition of the standardized loadings $\Psi^{-1/2}B = UDV$. Hence, an innovated HCT procedure is defined as the HCT procedure designed by [5] under independence applied to the decorrelated test statistics $T^* = \hat{\Sigma}^{-1/2}T$.

Finally, noting that the strong time-dependence in Σ results in a sparse structure for Σ^{-1}, we also propose an alternative ℓ_1-penalized ML estimation of Σ^{-1} which leads to sparse estimate of θ.

3 Results and partial conclusion

Some variants of HCT procedures are compared hereafter, including the method proposed by [9] based on the Cholesky decomposition of Σ (iHCT for innovated HCT), the correlation-adjusted t-tests introduced by [1], based on a James-Stein Shrinkage estimator and the Factor-innovated HCT (F-iHCT) method presented above, taking advantage of a factor structure for Σ. The comparison is both based on the application of the HCT and iHCT procedures to the auditory oddball ERP dataset introduced above and on intensive simulations under various dependence patterns. We particularly focus on two criteria: the prediction performance based on the selected features and the number of selected features. Only partial simulation results are reported here, demonstrating that an innovated HCT procedure based on a factor decomposition of Σ^{-1} shows desirable properties in a simulation scenario which mimics the auditory oddball ERP data introduced above.

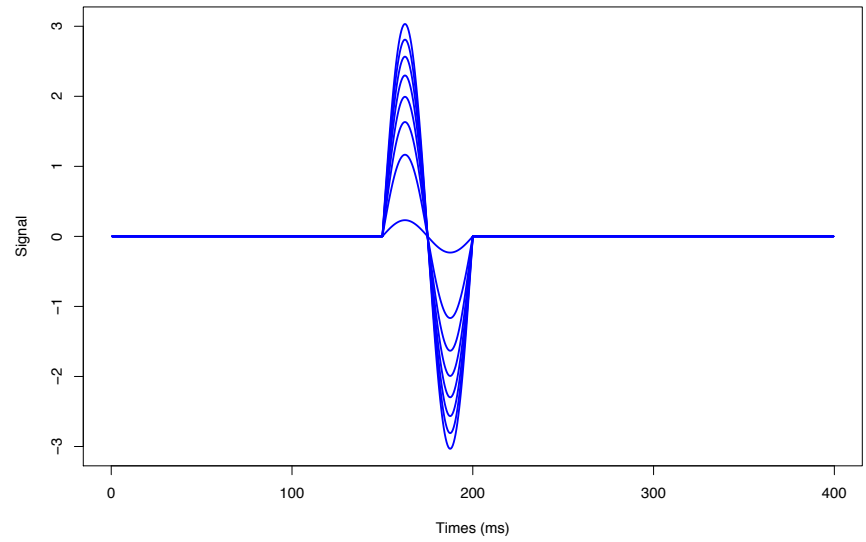


Figure 4: Simulation study - Signal amplitude curves along time

3.1 Simulation study

1,000 datasets with dimensions 30×800 are generated according to a multivariate normal distribution. Both the correlation structure and the within-condition variances are estimated from the auditory oddball ERP data introduced in section 2. This simulation plan therefore mimics the auditory ERP data by dimensions and covariance structure. Each dataset is split into two balanced groups. The normal distribution has expectation zero for the first 15 subjects (group 1) and the expectation for the 15 last subjects (group 2) is plotted on Figure 4. The difference curve is therefore a waveform with various amplitudes and the indices of non null features are in $[150ms, 200ms]$. $1,000$ training datasets are generated for each signal strength. Eight corresponding testing data of size 1000×800 with two balanced groups are also generated according to the same simulation plan for a prediction purpose. The RW model parameters for this simulation plan are $\varepsilon_T = 12\%$ and $A_T = \sqrt{2rlog(T)}$ with r taking 8 equally distributed values in $[0.004; 0.688]$. According to the RW setup, the present combination of r and β characterizes a not very sparse signal, with a weak to large strength.

As in [6], the variable selection step by different versions of HCT is followed by a supervised classification by Diagonal Discriminant Analysis on the subset of selected variables. Four methods are compared in this simulation study:

- Variable selection by standard HCT on raw p-values, classification by Naives Bayes (see [2]) denoted by *Standard HCT*;
- Variable selection by HCT on decorrelated test statistics using a shrinkage estimator of the whitening matrix (see [1]), classification by diagonal Shrinkage Discriminant Analysis (SDA, see [1]) denoted by *CAT-scores*;
- Variable selection by Factor-innovated HCT, classification by conditional Bayes classifier (proposed by [11]) denoted by *F-iHCT*;
- Variable selection by standard HCT performed on p-values adjusted for effects of latent factors as returned by the AFA ([13]) procedure using `erpfatest` function of `R` package `ERP` ([4]), classification by conditional Bayes classifier (see [11]) denoted by *AFA*.

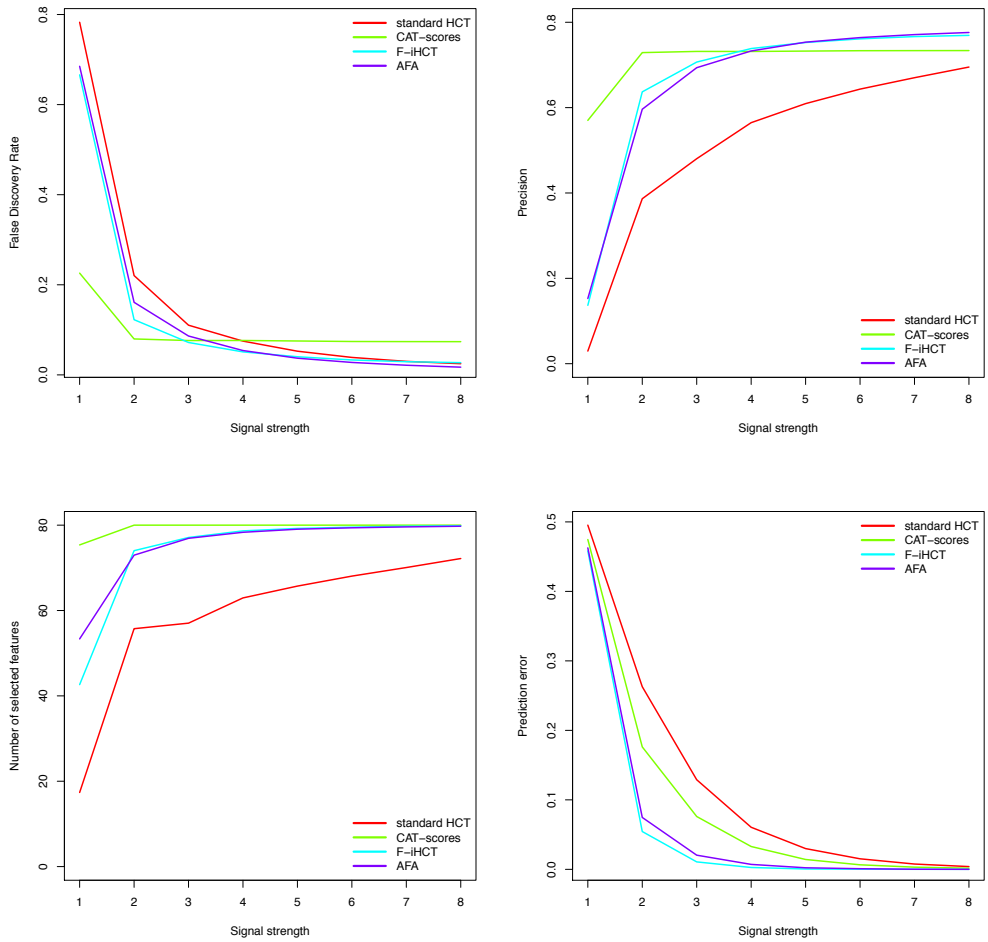


Figure 5: Results of the simulation study depending on signal strength: False Discovery Rate (top left), Precision (top right), Number of selected features (bottom left), Prediction error (bottom right).

For all the methods described above, the proportion of signal recovery, called precision, the false discovery rate (FDR), the number of selected features and the prediction error rate are computed. For all datasets, variable selection and estimation of classification rule are performed on training data (including the optimization of meta-parameters) and prediction error is computed on testing data.

Figure 5 shows that selection by CAT-scores appears to be the most efficient to catch weak signals, with both the smallest FDR and the largest precision for small amplitudes of signal. Even if CAT-scores does not achieve the best performance for large signal strengths, the FDR, precision and number of selected variables are remarkably stable. Standard HCT seems robust to dependence as the method performs well in term of FDR but its precision is small regarding methods based on decorrelation. Moreover, the number of selected variables is also small, which suggests that HCT is conservative under dependence. Lastly, classification by Naive Bayes fails as the error rates are the largest for weak to moderate strengths of signal. Variable selection and classification procedures based on the factor model assumption (AFA and F-iHCT) provide the best results both in terms of false positive, recovery of the signal and prediction error. FDR turns out to be small for moderate to high signal strengths and a correct power of signal identification is achieved.

3.2 ERP data analysis

The 4 methods compared in the simulation study are now applied on the auditory oddball ERP data presented in Section 2. For each method, the number of selected

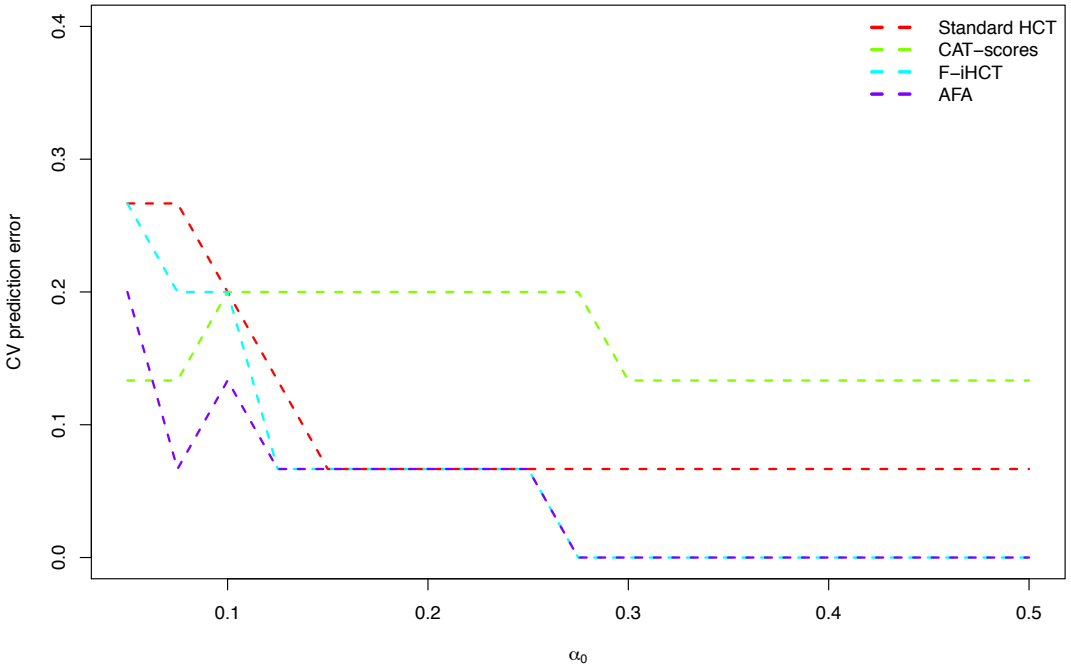


Figure 6: Real data study - Cross-validated prediction error of standard HCT, HCT performed on CAT-scores, factor innovated HCT and HCT performed on pvalues provided by AFA on auditory ERP experiment for several values of α_0

features and the prediction error are computed. As the number of observations if small, the classification error is computed by leave-one-out cross-validation (CV).

Note that the HCT method involves an hyper-parameter $0 \leq \alpha_0 \leq 1$, for which the recommendation varies depending on the authors. Small values of α_0 lead to more conservative selection procedures. Figure 6 presents the cross-validated error rates for several values of α_0. For values of α_0 larger than 0.15, standard HCT is stable and performs rather well. For more sparse models, standard HCT reaches larger error rates and is improved by decorrelation methods based on a factor model assumption (AFA and F-iHCT). The performance of the CAT-scores method varies slightly depending on α_0. The curves of F-iHCT and AFA are erratic for values of α_0 smaller than 0.125 but they stabilize when α_0 increases. For values of α_0 larger than 0.275, F-iHCT and AFA appear to be the most effective methods as they perfectly classify data. Nevertheless, one can notice that for equal error rates, the two methods do not select the same features as shown on the bottom of Figure 7.

Figure 7 shows the curve of the mean difference among the two groups and the time points selected by the 4 compared methods for $\alpha_0 = 0.125$ (top) and $\alpha_0 = 0.275$ (bottom). These values of α_0 are chosen because they provide two levels of sparsity but comparably small CV error. As expected, time points after 300ms are selected by all methods which is consistent with the literature but time points around 100 ms also appear to be significant.

To conclude, the AFA and F-IHCT methods which performs decorrelation by adjustment of covariates for the effect of latent factors seem to be the most suitable method in this example.

References

[1] M. Ahdesmäki and K. Strimmer. "Feature selection in omics prediction problems using cat scores and false non-discovery rate control". *Annals of Applied Statistics*, vol. 4, pp. 503-519, 2010.

[2] P.J. Bickel and E. Levina. "Some theory for Fisher's Linear Discriminant function, naive Bayes, and

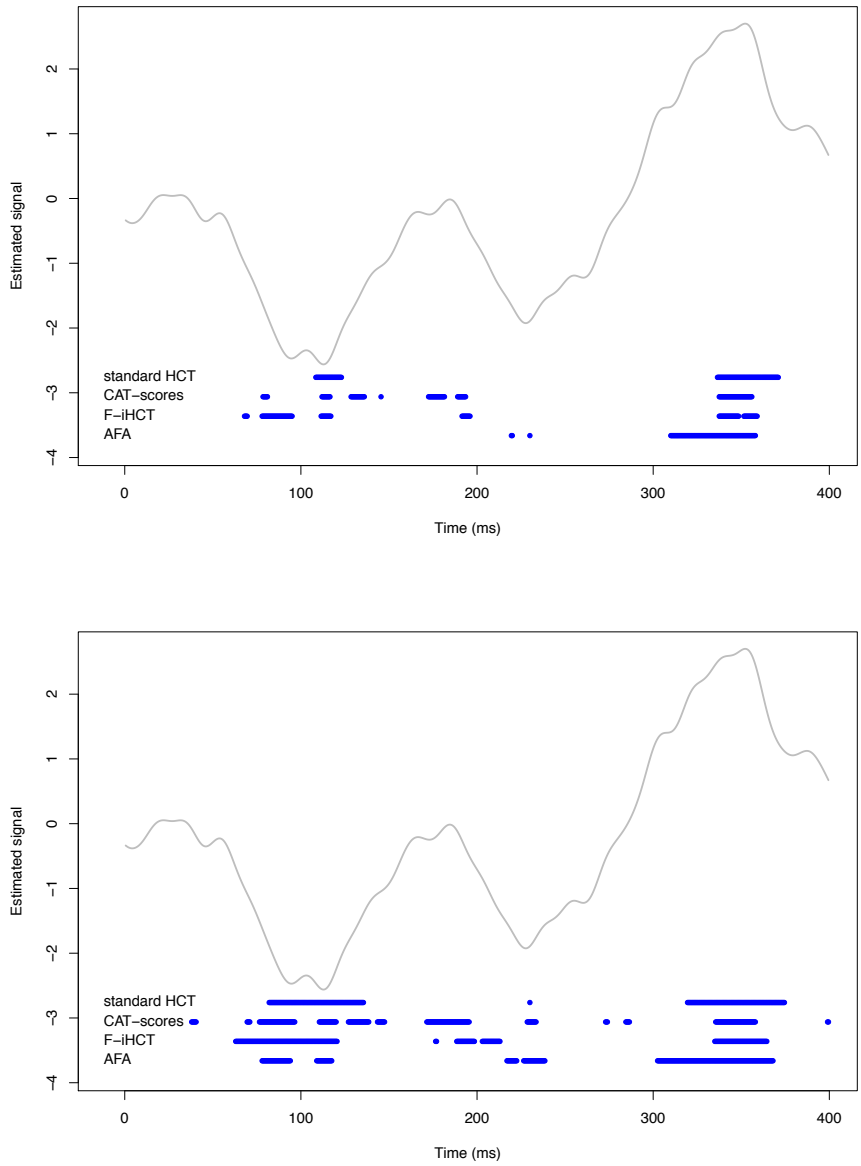


Figure 7: Real data study - Signal estimation (grey line) and significant time points (blue points) selected by standard HCT, HCT performed on CAT-scores, factor innovated HCT and HCT performed on p-values provided by AFA for $\alpha_0 = 0.125$ (top) and $\alpha_0 = 0.275$ (bottom) on auditory ERP experiment

some alternatives when there are many more variables than observations". *Bernoulli*, vol. 10, no. 6, pp. 989-1010, 2004

[3] D. Causeur and M.-C. Chu and S. Hsieh and C.-F. Sheu. "A factor-adjusted multiple testing procedure for ERP data analysis". *Behavior Research Methods*, vol. 44, pp. 635-643. 2012.

[4] D. Causeur, D. and C.F. Sheu. "ERP: Significance analysis of Event-Related Potentials data". R package version 1.0.1, http://CRAN.R-project.org/package=ERP, 2014.

[5] D. Donoho and J. Jin. "Higher criticism for detecting sparse heterogeneous mixtures". *The Annals of Statistics*, vol. 32, no. 3, pp. 962-994, 2004.

[6] D. Donoho and J. Jin. "Higher criticism thresholding: Optimal feature selection when useful features are rare and weak". *Proceedings of the National Academy of Sciences*, vol. 105, no. 39, pp. 14790-14795, 2008

[7] C. Friguet and M. Kloareg and D. Causeur. "A factor model approach to multiple testing under dependence". *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1406-1415, 2009

[8] P. Hall and J. Jin. "Properties of higher criticism under strong dependence". *The Annals of Statistics*, vol. 36, no. 1, pp. 381-402, 2008

[9] P. Hall and J. Jin. "Innovated higher criticism for detecting sparse signals in correlated noise". *The Annals of Statistics*, vol. 38, no. 3, pp. 1686-1732, 2010

[10] B. Klaus and K. Strimmer. "Signal identification for rare and weak features: higher criticism or false discovery rates?". *Biostatistics*, vol. 14, no. 1, pp. 129-143, 2013

[11] E. Perthame and C. Friguet and D. Causeur. "Stability of feature selection in classification issues for high-dimensional correlated data". *Statistics and computing*, to appear.

[12] T. W.Picton "The P300 wave of the human event-related potential". *Journal of Clinical Neurophysiology*, vol. 9, no. 4, pp. 456-479, 1992.

[13] C.F. Sheu and E. Perthame and D. Causeur and Y.S. Lee. "Accounting for time dependence in large-scale multiple testing of event-related potential data". Under revision. 2015.

[14] L.M. Williams and E. Simms and C.R. Clark and R.H. Paul "The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: neuromarker", *International Journal of Neuroscience*, vol. 115, pp. 1605-1630, 2005.

Validation Procedures for Predicted Gene Ontology Annotations

Davide Chicco[(1)(2)], Marco Masseroli[(2)]

(1) Princess Margaret Cancer Centre, University of Toronto, Ontario, Canada
(2) Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

Email: davide.chicco@gmail.com, marco.masseroli@polimi.it

Keywords: validation, Gene Ontology, biomolecular annotations, Receiver Operating Characteristic, ROC curves, Genomic and Proteomic Data Warehouse (GPDW)

Abstract In computational biology, controlled biomolecular annotations are very useful to describe the biological function features of genes and gene products through standard terminologies and ontologies. However, the available annotations contain errors, and the discovery and validation of new annotations are very time-consuming. Recently, scientists have taken advantage of different machine-learning algorithms to predict these *gene-function relationships*. While many of these methods have been easily adapted to the domain of bioinformatics, a difficult step is the validation of the predicted annotations. Here, we illustrate and compare three effective validation procedures that, together, are able to state the precision of any algorithm predictions with a reliable degree of accuracy. We show some validation results generated on Gene Ontology datasets of *Homo sapiens* gene annotations that prove the effectiveness of our validation techniques.

1 Scientific Background

In computational biology, a *controlled biomolecular annotation* is the association of a gene or gene product with a biological functional feature expressed through a controlled term, which can be part of a terminology or a controlled vocabulary structured within an ontology, such as the Gene Ontology (GO) [1]. Thus, the annotation states that the gene has the functional feature represented by the controlled term. For instance, the pair <*SLC1A6*, *L-glutamate transmembrane transporter activity*> represents the annotation of the *SLC1A6* gene to the *L-glutamate transmembrane transporter activity* molecular function. Despite their biological importance, there are some issues with available annotations, such as the presence of erroneous or missing ones. For this reason, computational algorithms and software tools able to produce ranked lists of reliably predicted annotations are a useful contribution.

In the past, we designed and developed several algorithms towards this goal. We started from a state-of-the-art algorithm based on truncated *Singular Value Decomposition* (tSVD) and developed some variants [2]. Then, in [3] article we designed an algorithm to choose the best truncation level for the tSVD, in [4] paper we designed and tested some *topic modeling* techniques, and in [5] manuscript we took advantage of a deep neural network approach.

All the prediction pipelines of these projects, as well as of any other similar project, share a common final pivotal step: the validation of results. Since biomolecular annotations are always incomplete (because our knowledge of biology is such incomplete), we do not have a *ground-truth* or *gold-standard* on which to rely; this makes us unable to take advantage of the common computational methods widely used for validation in applied machine-learning domains (such as computer vision or signal processing). To deal with this issue, we developed a method which assembles three different validation

procedures that, together, lead to a reliable determination of the predicted annotation accuracy. Here, we illustrate this method and its three techniques: the *analysis of the Receiver Operating Characteristic (ROC) curves*, the *comparison between available annotation versions*, and the *review of the scientific literature*.

After this Scientific Background, Section 2 illustrates our method and the included validation procedures. Section 3 shows some example results of the proposed validation method and Section 4 concludes.

2 Materials and Methods

In this section we describe the validation procedures that we assembled and implemented to test the effectiveness of the computational prediction methods: 2.1) ROC curve analysis, 2.2) comparison between different versions of available annotations, and 2.3) evaluation against the literature using available web tools.

2.1 Receiver Operating Characteristic (ROC) Curve Analysis

A ROC curve is a graphical plot which depicts the performance of a binary classifier system while its τ discrimination threshold varies [6]. Since usually in this field a reference *gold-standard* is not available, it is used to compare input and output annotations; for example, in [2], for all the possible values of the τ prediction likelihood, the algorithm computes the $TPrate = Sensitivity$ and $FPrate = 1 - Specificity$ with respect to the input annotation matrix. Thus, this ROC curve analysis is an efficient tool to understand the similarity between the input and output annotations of an annotation prediction method. A ROC curve showing a high *area under the curve* (AUC) corresponds to having many TPs (annotations present in the input and confirmed as present in the output) and many TNs (annotations absent in the input and confirmed as absent in the ouput). This means that the input matrix is very similar to the output matrix, and the output annotation profiles strongly reflect the input ones. On the contrary, a low AUC means a lot of differences between the input and output annotations. Given the comparison with the input annotations instead of with a gold standard, a good prediction should have a fairly high AUC. We consider a prediction insufficiently acceptable when its AUC is lower than $\omega = 2/3$. We chose this heuristic value to indicate that at least 66.67% of the output annotation matrix should be equivalent to the input matrix, since usually most of available annotations are correct although some errors and several missing annotations generally exist.

Despite the effectiveness of this ROC AUC analysis, our two other validation methods (annotation version comparison and literature review) are more useful and efficient.

2.2 Annotation Version Comparison

When an updated version of the controlled annotations previously used as input to a prediction method is available, the tally of the annotations predicted (AP) that are found confirmed in the updated version of the analyzed annotations provides an important validation. Note however that it can give only a lower estimate of the predicted annotation accuracy, since correctly predicted annotations could not be present in available updated annotations just because they have not been discovered yet, or simply because they have not yet been included in the available annotations.

The Genomic and Proteomic Data Warehouse (GPDW) [7] integrates numerous, multi-organism, gene and protein controlled annotation data from many different sources, including the Entrez Gene and GO databases. Relevant features of the GPDW are its periodical updates of the contained data and the storage in the GPDW of their outdated versions [8]. We leverage them by retrieving from the GPDW different, time distant versions of the available gene GO annotations, and using them as analyzed annotations and updated annotations for validation comparison, respectively (Figure 1).

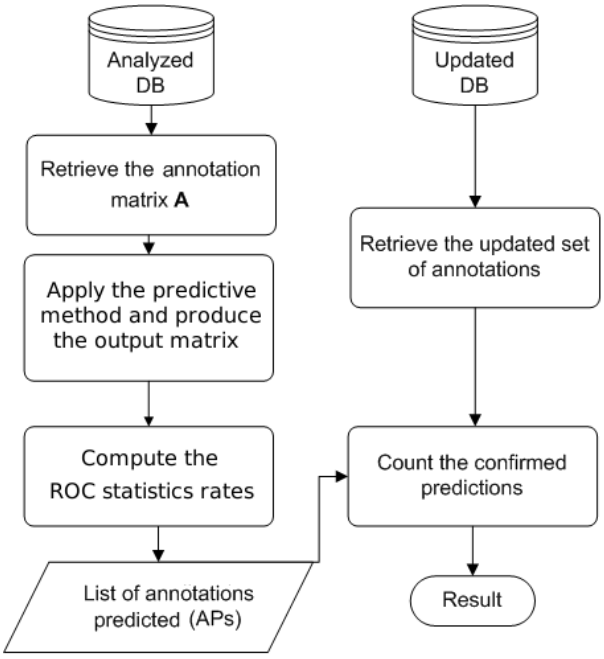


Figure 1: Flow chart of the validation procedure based on the comparison of database versions.

2.3 Literature Evaluation through Web Tools

The third and last step of our validation procedure is based upon searching updated literature resources for information supporting the predicted annotations. It is the only step not fully automated in our pipeline. The sources integrated in the GPDW mainly contain data from validated experiments, whose results are published in the literature. Yet, given the numerous research groups working independently all over the world and the many different journals in which results are published, some validated annotations published in the literature may have not yet been included in annotation databases. Thus, a literature review to search for confirmation of the annotations predicted by a computational method can provide effective validation results. For this last step of our validation procedure, we leverage the main online paper repository, PubMed [9], and the AmiGO [10] and GeneCards [11] web tools.

2.4 Evaluation

We applied all the described validation techniques to the gene GO annotations that we predicted with the methods described in [2]. Such methods are all based on the popular tSVD, also known as *principal component analysis*. We re-use the tests made by Khatri and colleagues [12], based on tSVD with a heuristic fixed truncation level (SVD-Khatri), and compared their results to those obtained with a tSVD variant that we developed (SVD-us), where the best truncation level is chosen through a ROC optimization algorithm [3].We also compared two other variants of the tSVD, named SIM1 and SIM2, both described in [2]. For the tests, we used as input the GO annotations of *Homo sapiens* genes available in the July 2009 version of the GPDW [7] (i.e. 14,341 annotations of 7,868 genes and 684 GO Cellular Components (CC), 15,467 annotations of 8,590 genes and 2,057 GO Molecular Functions (MF), and 21,048 annotations of 7,902 genes and 2,528 GO Biological Processes (BP)). For the result validation with the annotation version comparison techniques we used the corresponding gene GO annotations available in the March 2013 version of the GPDW (i.e. 31,135 annotations of 12,033 genes and 1,021 GO Cellular Components, 25,396 annotations of 10,460 genes and 2,603 GO Molecular Functions, and 64,212 annotations of 11,681 genes and 7,295 GO Biological Processes).

Table 1: ROC AUCs for the three *Homo sapiens* datasets and four prediction methods considered. The AUC area percentage is always greater than the minimum reliability threshold ω, which we heuristically fixed at 66.67%, except for the SVD-Khatri method applied to the GO CC dataset.

Method	CC	MF	BP	Method	CC	MF	BP
SVD-Khatri	58.98%	90.06%	77.24%	SIM1	80.94%	83.58%	70.20%
SVD-us	83.44%	85.40%	75.99%	SIM2	81.66%	83.32%	68.65%

3 Results

Using the three validation procedures defined, we compared the results and evaluated the performance of four different annotation prediction methods: (i) the tSVD as used by Khatri et al. [12] (with fixed truncation level $k = 500$ for all datasets), (ii) the tSVD with truncation level chosen by our automatic algorithm [3], (iii) the SIM1 with truncation level and cluster number chosen by our automatic algorithms [3], and (iv) the SIM2 with truncation level and cluster number chosen by our automatic algorithms [3] and using the Resnik similarity measure [2].

3.1 ROC Curve Analysis

We generated the ROC curves for the considered prediction methods and input datasets, and report their AUCs in Table 1. Almost all ROC AUCs are greater than $\omega = 66.67\%$, which is the minimum "reliability" threshold that we consider for the predictions. Only the ROC AUC generated by the SVD-Khatri method for the GO CC gene annotations did not reach that threshold; thus, we do not explore those predicted annotations further.

3.2 Annotation Version Comparison

In the first three cases in Table 2 we report the results obtained with a single GO sub-ontology dataset as input and output, while the results obtained with the complete GO dataset (CC+MF+BP) are in the last case in the Table: (a) Our tSVD method always outperforms the Khatri tSVD method with fixed truncation; the percentage of annotations found confirmed in the new GPDW version (last column) is greater for the MF and the BP datasets. (b) Our SIM1 method always outperforms the tSVD methods, except for the CC dataset, where it has the same performance as our tSVD method, and for the CC+MF+BP dataset, where the SVD-us outperforms all the other methods. The percentage of annotations found confirmed in the new GPDW version is greater for the MF and the BP datasets, and equal for CC dataset, for all the SIM1 tests. (c) Our SIM2 method always outperforms the the SIM1 and tSVD methods, except for the CC dataset, where they all have the same results; the percentage of annotations found confirmed in the new GPDW version (last column) is greater for the MF and BP datasets, and equal for the CC dataset. The complete GO dataset (CC+MF+BP) shows an increased number of validated predicted annotations, which are much more than the ones predicted in the single GO sub-ontology tests. In addition, the SVD-us method outperforms SIM1 and SIM2.

3.3 Literature Evaluation

Once we had the lists of the annotations predicted by our methods, we searched for confirmation of their existence in the literature, as described previously. Out of the annotations predicted with the tSVD method with our best truncation level for each single GO sub-ontology, we found in the literature the annotations which we then reported in Table 3. Out of the total 153 annotations predicted (CC: 8, MF: 81, BP: 64), 8 (5.30%) annotations (MF: 4, BP: 4) were found in published scientific papers, GeneCards or AmiGO. Out of the total 56 annotations predicted through the SIM1 method (CC: 8, MF: 13, BP: 35), 2 (3.57%) annotations (1 MF and 1 BP) were found in published sci-

Table 2: Comparison of the results of the tSVD with truncation level as in Khatri et al. [12] (*SVD-Khatri*), tSVD with our automatic truncation level (*SVD-us*), SIM1 and SIM2 methods. The τ threshold minimizes the sum $FPs + FNs$. C: number of clusters for SIM1 and SIM2. SIM2 uses Resnik's similarity. *APs*: number of annotations predicted; *anDB*: number of predicted annotations found in the November 2009 GPDW version; *upDB* (*upDB%*): number (percentage) of predicted annotations found in the March 2013 updated GPDW version (percentage over the predicted ones). The most important values are **bolded**: the percentages of APs found on the updated GPDW version. The values of the ROC AUC of these tests are in Table 1; the SVD-Khatri, SVD-us, SIM1 and SIM2 methods are described in [2].

Method	k	τ	C	APs	anDB	upDB	upDB%
Homo sapiens, GO Cellular Component - CC							
SVD-Khatri	500	0.45		0	0	0	**0.00**
SVD-us	378	0.49		8	0	4	**50.00**
SIM1	378	0.49	2	8	0	4	**50.00**
SIM2	378	0.49	2	8	0	4	**50.00**
Homo sapiens, GO Molecular Function - MF							
SVD-Khatri	500	0.48		108	0	4	**5.56**
SVD-us	607	0.48		81	2	5	**6.17**
SIM1	607	0.48	5	13	0	1	**7.69**
SIM2	607	0.48	5	30	0	3	**10.00**
Homo sapiens, GO Biological Process - BP							
SVD-Khatri	500	0.48		358	1	48	**13.51**
SVD-us	1,413	0.45		64	2	12	**18.75**
SIM1	1,413	0.45	2	35	1	10	**28.57**
SIM2	1,413	0.45	5	14	0	8	**57.14**
Homo sapiens, GO CC+MF+BP							
SVD-Khatri	500	0.45		794	196	234	**29.47**
SVD-us	1,905	0.43		112	3	51	**45.54**
SIM1	1,905	0.43	2	116	3	45	**38.79**
SIM2	1,905	0.43	2	111	3	49	**44.14**

Table 3: *Homo sapiens* GO annotations predicted by our tSVD method (SVD-us, Table 2) and confirmed in the literature search. If the annotation was added to the latest available Gene Ontology version, its evidence is reported. The single annotation not found in the version comparison analysis is in **bold**.

tSVD with truncation level chosen by our automatic algorithm				
Sub-Ontology	Gene Symbol	GO Term ID	GO Term Name	Evidence
MF	SLC1A6	GO:0005313	L-glutamate transmembrane transporter activity	IEA
MF	HDAC6	GO:0004407	Histone deacetylase activity	IEA
MF	POR	GO:0004128	Cytochrome-b5 reductase activity	IEA
MF	NT5M	GO:0008253	5'-Nucleotidase activity	EXP
BP	ITGA6	GO:0007155	Cell adhesion	IEA
BP	**ITGA6**	**GO:0007160**	**Cell-matrix adhesion**	**IEA**
BP	CPA2	GO:0006508	Proteolysis	IEA
BP	AHR	GO:0006805	Xenobiotic metabolic process	TAS

entific papers, GeneCards or AmiGO. Out of the total 52 annotations predicted through the SIM2 method (CC: 8, MF: 30, BP: 14), 4 (7.69%) annotations (3 MF and 1 BP) were found in published scientific papers, GeneCards or AmiGO.

Through the literature analysis, we found a predicted annotation that was not in the updated GPDW version, i.e. <*ITGA6*, *Cell-matrix adhesion*> (in **bold** in Table 3). Out of the total 153 annotations predicted by our tSVD method, 21 (13.73%) of them were validated in the updated GPDW version, and we found only 1 additional annotation in the literature. Given the time required to perform the literature evaluation, this result may seem very limited; this is why we consider to be more useful and reliable the first two validation procedures (*ROC curve analysis* and *annotation version comparison*),

particularly the latter one.

4 Conclusions

Validation of functional annotation predictions in biology is always a difficult task. Available annotations continuously increase while scientists discover new biology; furthermore, some of the available annotations may contain errors, which could be corrected in their subsequent versions. A *gold-standard* to use in validation is not available, so stating if a machine learning prediction algorithm is performing well is quite difficult. In this paper, we illustrated three validation procedures that we developed and used to validate the GO annotations of *Homo sapiens* genes predicted through some computational learning methods. These three techniques mutually compensate for each others' strengths and weaknesses and, even if not fully innovative, all together represent an useful tool to state the quality of biomolecular annotations predicted through any computational algorithm.

Despite our evaluation of validation procedures using only GO annotations, such procedures are not bound to the Gene Ontology or even to the biological domain, but can be used in any scientific validation in which a full *gold-standard* does not exist or is always changing. In the future, we plan to improve the use of our overall validation method by additionally automating the literature evaluation step, through the use of text mining techniques.

Acknowledgments

This work was partially supported by the Data-Driven Genomic Computing (GenData 2020) PRIN project (2013-2015), funded by Italy's Ministry of Education, Universities and Research (MIUR). Authors thank Coby Viner (University of Toronto) for his help in the English proof-reading of this article.

References

[1] The Gene Ontology Consortium, "Creating the Gene Ontology resource: Design and implementation". *Genome Res.*, 11(8): 1425-1433, 2001.

[2] D. Chicco, M. Tagliasacchi, and M. Masseroli, "Biomolecular annotation prediction through information integration". *CIBB*, 1-9, 2011.

[3] D. Chicco, and M. Masseroli, "A discrete optimization approach for SVD best truncation choice based on ROC curves". *IEEE BIBE*, 96, 1-4, 2013.

[4] P. Pinoli, D. Chicco, and M. Masseroli, "Enhanced Probabilistic Latent Semantic Analysis with Weighting Schemes to predict genomic annotations". *IEEE BIBE*, 92, 1-4, 2013.

[5] D. Chicco, P. Sadowski, and P. Baldi, "Deep autoencoder neural networks for Gene Ontology annotation predictions". *Proceedings of ACM BCB*, 533–540, 2014.

[6] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers". *Mach. Learn.* 31: 1-38, 2004.

[7] A. Canakoglu, M. Masseroli, S. Ceri, L. Tettamanti, G. Ghisalberti, and A. Campi, "Integrative warehousing of biomolecular information to support complex multi-topic queries for biomedical knowledge discovery". *IEEE BIBE*, 159, 1-4, 2013.

[8] Genomic and Proteomic Knowledge Base (GPKB), `http://www.bioinformatics.deib.polimi.it/GPKB/`

[9] NCBI PubMed, `http://www.ncbi.nlm.nih.gov/pubmed/`

[10] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, AmiGO Hub, and Web Presence Working Group, "AmiGO: online access to ontology and annotation data". *Bioinformatics*, 25(2): 288-289, 2009.

[11] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support". *Bioinformatics* 14(8): 656-664, 1998.

[12] P. Khatri, B. Done, A. Rao, A. Done, and S. Draghici, "A semantic analysis of the annotations of the human genome". *Bioinformatics*, 21(16): 3416-3421, 2005.

MANAGING NGS DIFFERENTIAL EXPRESSION UNCERTAINTY WITH FUZZY SETS

Arianna Consiglio[(1)], Corrado Mencar[(2)], Giorgio Grillo[(3)], Sabino Liuni[(4)]

(1) Institute for Biomedical Technologies
CNR, Via Amendola 122/D 70126 Bari Italy, and
Dept. of Informatics
University of Bari Aldo Moro, Via Orabona 4 70125 Bari Italy,
arianna.consiglio@ba.itb.cnr.it

(2) Dept. of Informatics
University of Bari Aldo Moro, Via Orabona 4 70125 Bari Italy,
corrado.mencar@uniba.it

(3) Institute for Biomedical Technologies
CNR, Via Amendola 122/D 70126 Bari Italy, giorgio.grillo@ba.itb.cnr.it

(4) Institute for Biomedical Technologies
CNR, Via Amendola 122/D 70126 Bari Italy, sabino.liuni@ba.itb.cnr.it

Keywords: RNA-Seq, differential expression, multireads, fuzzy sets, possibilistic modeling.

Abstract. The application of high-performance Next-Generation Sequencing (NGS) technologies is widely used to characterize case-control comparison studies for RNA transcripts, such as mRNAs and small non-coding RNAs. The first step in the analysis strategies is mapping NGS reads against a reference database, and a critical issue is choosing how to deal with multiread problem. In this paper we present a novel approach to represent and quantify read mapping ambiguities through the use of fuzzy sets and possibility theory. The aim of this work is to obtain a list of candidate differential expression events, ordered by significance, providing a description of the uncertainty of the results due to the multiread issue. A preliminary experiment on a case-control study of human endobronchial biopsies resulted in the identification of 9 genes with possible differential expression, four of them with an uncertain fold change. This result was confirmed by FDR adjusted Fisher's test, while the same data processed with DESeq2 did not provide significant differences between case and control.

1 Scientific Background

NGS technology is continuously improving and the produced reads are increasingly numerous. When working with alignment-based methods, a confounding factor is the presence of gene duplication, repetitive regions and overlapping genes. These events induce the problem of *multireads* in the NGS mapping procedure when a significant proportion of reads map to more than one location. This issue can lead to mistakes and imprecision in differential expression or alternative splicing analysis based on counts of reads mapping to some reference databases.

When multireads are sporadic, usually such reads are discarded from analysis, but this option leads to an underestimation of the read counts. In the last years, alternative strategies have been developed for the estimation of read counts in presence of multireads. The simplest choice is to randomly assign multireads to references (as in best-match mapping) or proportionally to the expression of uniquely mapped reads [1].

More complex techniques compute an estimation of the read counts using probabilistic models, based on some assumptions on the distribution of data [2, 3, 4].

The estimated expressions are given as input to the tools for the analysis of differential expression [5, 6]. Such tools scale the counts in order to make the expression values comparable, then they compute the fold change and a p-value with a statistical test, and eventually select a list of candidate differentially expressed genes. These results may contain many false positives and must be validated with further laboratory assays.

In this paper we propose a novel method, based on fuzzy sets and possibility theory, that deals with the inherent uncertainty of multiread mapping. The approach used is compliant with the work of Zadeh [7], who proposed possibility distributions as suitable interpretations of fuzzy sets. The possibility measure is used in this paper following the notation introduced by Pedrycz [8].

The aim of this work is to obtain a list of candidate differential expression events ordered by significance, while also providing a description of the uncertainty of the results due to the multireads issue, for an easier detection of false positives. The proposed approach is based on the idea of representing and quantifying read mapping ambiguities without heavy simplifications or stringent probabilistic assumptions.

2 Materials and Methods

2.1 *Fuzzy representation of gene expression*

The uncertainty of multireads is modeled through fuzzy sets describing the possibility that each gene has a given read count. When a read is mapped against a reference database, we have one of the following results: i) the read does not map to any reference sequence; ii) the read maps to only one reference sequence (unique mapping); iii) the read maps to more than one reference sequence with equal or different mapping quality (multiple mapping). As a consequence, for each gene we can quantify the number of reads according to four different cases:

- A = number of uniquely mapping reads;
- B = number of reads having the gene as unique best match (i.e. other genes may match, but with lower quality);
- C = number of reads having the gene as best match, although not unique (i.e. other genes may match with the same quality);
- D = number of reads having the gene as match, even if not best.

According to their definition, $A \leq B \leq C \leq D$. Tab. 1 shows some examples of multireads, mapped genes and quality values of mapping (scaled to 1 for each read).

These cases enable the definition of a possibility degree that a given count x is the real number of reads mapping to a particolar gene. More specifically, this possibility is null for $x < A$ and $x > D$, while it is maximal (=1) for $B \leq x \leq C$. A gradual, increasing possibility can be conceived for $A \leq x \leq B$, while a gradually decreasing possibility can be assumed for $C \leq x \leq D$. Linear increase/decrease can be assumed for simplicity.

The possibility distribution of the gene expression - based on read counts - is determined by A, B, C and D, and can be defined by a trapezoidal fuzzy set, as follows:

$$\mathrm{Tr}\left[A', B, C, D'\right](x) = \begin{cases} 0, & x \leq A' \vee x \geq D \\ \frac{x-A'}{B-A'}, & A' < x \leq B \\ 1 & B < x < C \\ \frac{x-D'}{C-D'} & C \leq x < D' \end{cases} \tag{1}$$

Table 1: Example of mapping results. For gene gene-1: $A = 1, B = 2, C = 3, D = 3$. For gene-2: $A = B = 0, C = 1, D = 2$.

read id	gene id	mapping quality
read-1	gene-1	1.0
read-2	gene-1	1.0
read-2	gene-2	0.8
read-3	gene-1	1.0
read-3	gene-2	1.0

where $A' = A - 1$ and $D' = D + 1$ to give non-null possibility to counts A and D respectively. The width of the fuzzy set (1) (defined as $D' - A'$) quantifies the uncertainty in the evaluation of the expression value, which in turn generates uncertainty in differential expression evaluation.

2.2 *Graphical comparison of differential expression*

For a qualitative evaluation of differential expression of genes in a case-control study, a graphical method can be proposed as a first approach.

Two trapezoids representing the expression of the same gene in different samples can be plotted on a 3-dimensional graph, which is useful to fully understand the use of fuzzy sets and related possibility distributions. The count values for the two experimental samples are drawn on the x axis and y axis respectively, while the possibility degrees are represented on the z axis. As shown in fig. 1, the Cartesian product of two trapezoidal fuzzy sets, representing the expression of the same gene in different samples, yields a truncated pyramid of possibilities (3D plot). The z-value of the pyramid is the possibility degree that the first sample has x reads *and* the second sample has y reads for the gene under consideration.

The projection on a 2D plot highlights two rectangles which bound the possibility: the innermost covers the area with highest possibility, while the outermost limits the area with non-null possibility. Larger rectangles represent wider uncertainty, small rectangles (possibly degenerating to a single point) represent more definite results. The position of the rectangle with respect to the bisector line describes the differential expression result in the case-control comparison.

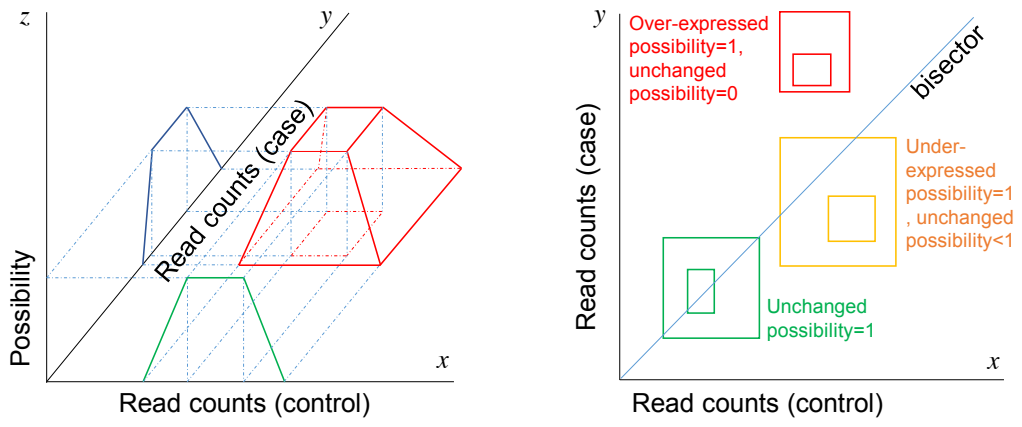


Figure 1: Graphical interpretation of the fuzzy sets and their comparison for differential expression evaluation.

2.3 *Fuzzy Fold Change Computation*

The proposed quantitative method for the evaluation of differential expression extends the fold change metric, usually adopted for differential expression, by integrating fuzzy sets representing uncertain read counts. In particular, given a control sample with fuzzy expression $\mathrm{Tr}\,[A'_1, B_1, C_1, D'_1]$ and a control case with fuzzy expression $\mathrm{Tr}\,[A'_2, B_2, C_2, D'_2]$, we extend the usual fold change metric to the following fuzzy fold change metric:

$$\mathrm{Tr}\left[\log_2 \frac{A'_1}{D'_2}, \log_2 \frac{B_1}{C_2}, \log_2 \frac{C_1}{B_2}, \log_2 \frac{D'_1}{A'_2}\right] \tag{2}$$

This trapezoidal fuzzy set follows from the application of the extension principle to the standard fold change metric, eventually simplified to a trapezoidal fuzzy set for ease of computation.

The fuzzy fold change is very useful to highlight potential false positives when the value of 0 (corresponding to null variation between case and control) belongs to fuzzy fold change with high possibility degree.

2.4 *Fuzzy representation of data and differential expression*

For a complete differential expression analysis, all the genes of both the case and the control samples must be taken into account. The last approach we propose ranks the genes in both samples in order of possibility that their expression in the case and control is significantly different. This approach combines the fuzzy fold change metric with a fuzzy representation of the dataset of genes.

In order to analyze the trend of the logarithmic fold change, we represent expression data in an MA-plot, as in fig. 2 (main plot)[1]. For simplicity, each gene is represented as a point and its expression value is the centroid of its trapezoid[2].

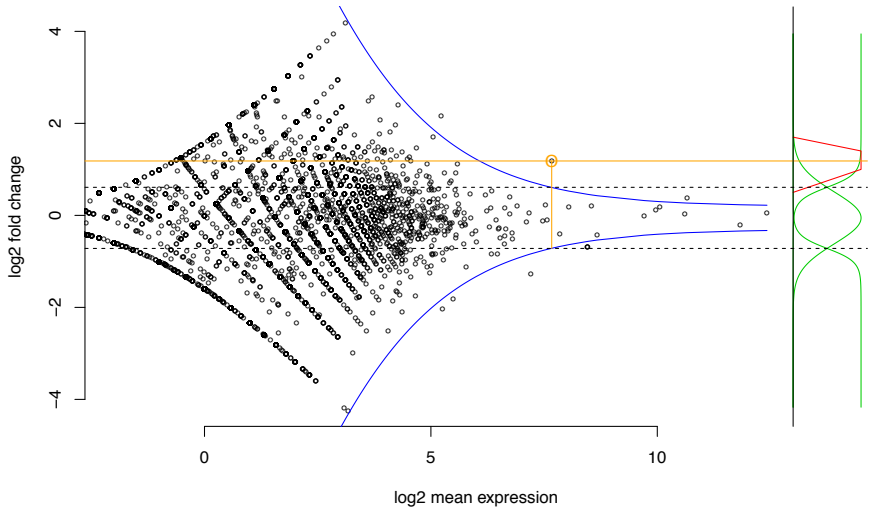


Figure 2: MA-plot of estimated data shows the variability of log2 fold change for low expression values (main plot). Computation of the differential expression possibility of a gene (projections on the right).

The plot clearly shows that the variability of fold change decreases as the mean expression value increases. The genes that are distant from the main rhomboidal figure are the best candidates for differential expression.

In order to rank genes according to their differential expression, we draw two curves enclosing the right part of the rhomboidal figure (we are not interested in the left part,

[1]Given two expressions e_1 and e_2 of a gene in two samples, the MA-plot places the gene on a plane (M,A) where $M = \log_2(e_1/e_2)$ (the fold change) and $A = (1/2)\log_2(e_1 e_2)$ (average intensity).

[2]If the centroid falls outside the interval $[B, C]$, it is limited to the closest extreme of this interval.

because it represents genes with too small expressions). The enclosing curves represent the limits for a varied expression to be considered as unrelated to the experimental conditions. Therefore, the genes lying on these frontiers can be associated to a possibility of being differentially expressed equal to 0.5; on the other hand, genes above the upper curve or below the lower curve have higher possibility of being differentially expressed.

Thereby, it is not difficult to compute a differential expression possibility value for each gene of the plot. Given a gene, its corresponding point is located in the MA-plot. If it is located in the left part of the rhomboid, it is excluded from further analysis because its expression is not significant. Otherwise, in correspondence of its abscissa, the ordinates y^+, y^- of the two enclosing curves are produced. Three fuzzy sets are then defined on the vertical axis: the first fuzzy set represents over-expression, the second fuzzy set represent insignificant variation, and the third fuzzy set represent under-expression (see fig. 2, projections on the right). The fuzzy sets representing under-expression and over-expression are defined in terms of a sigmoidal membership function, while the fuzzy set representing insignificant variation is defined as a Gaussian fuzzy set. The fuzzy set representing under-expression (resp. over-expression) intersects the fuzzy set representing insignificant variation at y^- (resp. y^+), with membership degree equal to 0.5.

The three fuzzy sets are used to evaluate the possibility of the gene to be differentially expressed. The fuzzy fold change of the gene is compared to the three fuzzy sets. More specifically, the possibility measure is computed between the fuzzy fold change and the under-expression fuzzy set, the insignificant variation fuzzy set, and the over-expression fuzzy set. (The possibility measure between two fuzzy sets F_1 and F_2 is defined as $\Pi(F_1, F_2) = \max_x \min\{F_1(x), F_2(x)\}$.) The possibility measure, between the fuzzy fold change and the fuzzy set representing over-expression (resp. under-expression), quantifies the possibility that the gene is over-expressed (resp. under-expressed) in the control sample. The possibility measure between the fuzzy fold change and the fuzzy set representing insignificant variation evaluates the possibility of false positiveness.

By repeating the procedure for all the genes, a ranked list is eventually produced with genes sorted according to their possibility of being differentially expressed, and accompanied with an additional information of possible false-positiveness.

3 Results

The proposed model was tested using two datasets downloaded from NCBI-SRA archive: DRP000527 and SRP014005. The datasets were mapped against Vega transcript database [9], while DESeq2 [5], Cuffdiff [6] and Fisher's Exact test p-value (adjusted with False Discovery Rate) were used to evaluate differential expression.

The first dataset contains two samples coming from the HeLa cells. In one sample the U2AF35 gene is suppressed, and no other gene should result differentially expressed in the analysis results. Illumina reads were mapped using Bowtie. Because of to the low number of mismatches allowed, all the mapping have the same quality and are considered as equivalent best matches. In this case in the trapezoids A coincides with B and C with D. The 17% of mapped reads are multireads. The mapping identified 19486 genes. The U2AF35 gene is the only varied gene, with a possibility = 1. This result was confirmed by DESeq2, Cuffdiff and Fisher's test.

The second dataset contains a case-control study of Asthma, through 454 Roche sequencing of human endobronchial biopsies. 454 reads were mapped using BLAST, with 97% of identity required. The 16% of mapped reads are multireads. The mapping identified 14802 genes, 9 genes have a possibility of being differentially expressed $>$ 0.5 and 4 of them have an uncertain fold change. Both centroid data and uniquely mapping read counts have been processed with DESeq2, but the tool only warns about the absence of replicates and outputs no significant differences in the two samples. The adjusted p-value of Fisher's test selects 13 differentially expressed genes: 9 of them are

the same highlighted with fuzzy possibility sets, while the other 4 show a possibility of being differentially expressed between 0.08 and 0.46 and one of them has also maximum uncertainty. Cuffdiff cannot be run on 454 data.

4 Conclusion

The described method exploits fuzzy sets to manage the uncertainty in multireads, in particular during the evaluation of differential expression analysis with NGS RNA-Seq data. The model has been tested on case-control transcriptomic data produced by Roche 454 and Illumina sequencers.

Gene expressions are represented with trapezoidal fuzzy sets, which represent the ambiguities resulting from read mapping. Genes are ranked through a possibility measure of differential expression, accompanied with information about the uncertainty that could be present in the results, caused by multireads.

The uncertainty representation can also be used just to add information to the results obtained with other differential expression tools, in order to highlight the risk of false positives in the results.

The model can also be applied to different types of data, like genomic and metagenomic reads, and it will be extended to cope with biological replicates and different types of sample comparison (e.g. with more than two conditions or time series data).

Acknowledgments

We thank Dr. Flavio Licciulli, Dr. Mariano Caratozzolo and Dr. Flaviana Marzano for their suggestions and help with NGS data elaboration. A.C. is supported by Progetto MICROMAP PON01_02589.

References

[1] G.J. Faulkner, A.R. Forrest, A.M. Chalk, K. Schroder, Y. Hayashizaki, P. Carninci, D.A. HUme, S.M. Grimmond. “A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE”. *Genomics*, 91(3), 281-288, 2008.

[2] H. Jiang, W.H. Wong. “Statistical inferences for isoform expression in RNA-Seq”. *Bioinformatics*, 25(8), 1026-1032, 2009.

[3] B. Li, V. Ruotti, R.M. Stewart, J.A. Thomson, C.N. Dewey. “RNA-Seq gene expression estimation with read mapping uncertainty”. *Bioinformatics*, 26(4), 493-500, 2010.

[4] B. Li, C.N. Dewey. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. *BMC Bioinformatics*, 12(1), 323, 2011.

[5] M.I. Love, W. Huber, S. Anders. “Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2”. *Genome Biology*, 15(12), 550, 2014.

[6] C. Trapnell, D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn, L. Pachter. “Differential analysis of gene regulation at transcript resolution with RNA-seq”. *Nature biotechnology*, 31(1), 46-53, 2013.

[7] C. Negoita, L.A. Zadeh, H.J. Zimmermann. “Fuzzy sets as a basis for a theory of possibility”. *Fuzzy sets and systems*, 1, 3-28, 1978.

[8] W. Pedrycz, F. Gomide. “An introduction to fuzzy sets: analysis and design”. *Mit Press*, 1998.

[9] L.G. Wilming, J.G.R. Gilbert, K. Howe, S. Trevanion, T. Hubbard, J.L. Harrow. “The vertebrate genome annotation (Vega) database”. *Nucleic Acids Research*, 36(suppl 1), D753-D760, 2008.

SUPERVISED TERM WEIGHTS FOR BIOMEDICAL TEXT CLASSIFICATION

Mounia Haddoud[1,2], Aïcha Mokhtari[2], Thierry Lecroq[1], Saïd Abdeddaïm[1]

(1)Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes (LITIS)
Université de Rouen, 76821 Mont-Saint-Aignan Cedex, France.
{mounia.haddoud1,thierry.lecroq,said.abdeddaim}@univ-rouen.fr

(2) Recherche en Informatique Intelligente, Mathématiques et Applications (RIIMA)
USTHB, BP 32, El-Alia, Bab-Ezzouar, 16111 Algiers, Algeria

Keywords: Bio-medical text mining, text classification, supervised term weighting, support vector machines, nearest centroid.

Abstract. Maintaining accessibility of biomedical literature databases has led to development of text classification systems to assist human indexers by recommending thematic categories to biomedical articles. These systems rely on using machine learning methods to learn the association between the document terms and predefined categories. The accuracy of a text classification method depends on the metric used in order to assign a weight to each term. Weighting metrics can be classified as supervised or unsupervised according to whether they use prior information on the number of documents belonging to each category. In this paper, we propose two supervised weighting metrics and an extended term representation which both improve the quality of biomedical document classification.

1 Scientific Background

Medical Subject Headings (MeSH), a controlled set of keywords, are used to index all the article abstracts contained in the MEDLINE database to facilitate search and retrieval. The increasing size of the MEDLINE needs efficient text classification tools to assist indexers in labeling document texts with the predefined thematic categories of MeSH [1, 2, 3]. In the two last decades a huge number of machine learning techniques were proposed to automatically classify text documents. In text classifier systems, documents are preprocessed in order to be suitable as training data for a learning algorithm. Traditionally, each text document is converted into a vector where each dimension represents a term which value is the weight that will be used in the learning process. As the weight reflects the importance of the term in the document, an appropriate choice of the metric function used for weighting terms is crucial for correct classification. Common term weighting metrics for text classification were unsupervised and generally borrowed from information retrieval (IR) field. The simplest IR metric is the binary representation BIN which assigns a weight of 1 if the term appears in the document and 0 otherwise. The term can be assigned a weight TF that reflect its frequency in the document. TFIDF is the most commonly used weighting metric in text classification. TFIDF is the product of TF and IDF, the inverse document frequency which favors rare terms in the corpus over frequent ones. However, there are some drawbacks on using unsupervised weighting functions, as the category information is omitted.

Previous studies proposed different supervised weighting metrics where the document frequency factor IDF of TFIDF is replaced by a factor that use prior information on the number of documents belonging to each category. Several classical metrics were tested in the literature, for instance, chi-square (χ^2), information gain (IG), gain ratio (GR) and odds ratio (OR) [4]. These early studies get an improvement with TF.χ^2,

TF.IG, TF.GR and TF.OR term weights trained with SVM. Accurate SVM text classification was obtained using Bi-Normal Separation (BNS) metric for supervised term weighting [5]. More recently, other specific metrics were proposed for the supervised term weighting problem. Liu et al. [6] uses a probability based (PB) term weight in order to tackle the problem of imbalanced distribution of documents among categories. Lan et al. [7] utilizes a term weight TF.RF based of the Relevance Frequency (RF) metric. Altinçay and Erenel [8] combined RF metric with mutual information and the difference of term occurrence probabilities in the collection of the documents belonging to the category and in its complementary set.

The rest of the paper is organized as follows. Section 2 describes the supervised metrics we propose for weighting and the corpus used as benchmark. The experimental comparison of our metrics with those proposed in the literature is presented in Section 3.

2 Materials and Methods

We propose two metrics (One-way Klosgen and Loevinger) and compare them with 10 metrics that have been used for term weighting problem in the literature. We also propose to represent a text document as a vector where each dimension can be either term frequencies or term positions.

2.1 Extended term representation

In this classical representation, terms are viewed as the dimensions of the learning space. A term may be a single word or a phrase (n-gram). In this work, we propose to represent each dimension by a term together with its minimal frequency in the document. Let us consider for example, a particular term t such that 25% of the documents where t appears are in category c. If 45% of the documents where t appears at least 3 times are in category c, then the term t is probably more correlated with the category c when its frequency exceeds 2. Hence, we propose features of the form (t, n) in documents containing t with a term frequency at least n. If a document d contains ten times a term t, we must generate ten features (t, i) ($i = 1, 2, \ldots, 10$), meaning that t occurs at least once, twice,. . . , ten times. This could unnecessarily grow the number of features so we consider only n powers of 2. Then, if t occurs ten times, we will generate the features $(t, 1)$, $(t, 2)$, $(t, 4)$ and $(t, 8)$. The number of frequency features associated to a term t which appears n times in a document d will only be $\log_2 p$ in the worst case.

Most of the terms that are related to the main topics of a document occur at its beginning. In order to validate this assumption we propose features of the form (t, p), meaning that the first position of t in the document is lower or equal to p. The position being defined as the number of words preceding the term occurrence. As for term frequency features, we generate only features (t, p) with p powers of 2. For example, if a term t first appears at position 5 in a document of size 100 words, we generate the features $(t, 8)$, $(t, 16)$, $(t, 32)$ et $(t, 64)$, meaning that the first position of t is lower or equal than 8, 16, 32 and 64. The number of position features associated to a term t which appears in a document d at first position p will be $\log_2 |d|$ in the worst case, where $|d|$ is the size of d in number of words.

2.2 Weighting metrics

We consider a corpus D of N documents and d a particular document of D. Let x denotes a nominal feature of d representing either, 1) t a term that occurs in d, 2) (t, n) a term that occurs at least n times in d, or 3) (t, p) a term which first position is lower or equal to p in the document d.

Each document can belong to one or many categories (labels or classes) $c_1, c_2, \ldots, c_k$. We denote by y a particular category c_i. We denote by $\bar{x}$ the fact that the feature x is not present in d and by $\bar{y}$ the fact that d does not belong to the category y. The number

Table 1: Two-way contingency table for nominal feature x (term) and category y (document label). $f(uv)$ denotes the number of documents containing u and belonging to v. $*$ represents any term or category.

	y	$\bar{y}$	$*$
x	$f(xy) = a$	$f(x\bar{y}) = b$	$f(x*)$
$\bar{x}$	$f(\bar{x}y) = c$	$f(\bar{x}\bar{y}) = d$	$f(\bar{x}*)$
$*$	$f(*y)$	$f(*\bar{y})$	$f(**) = N$

Table 2: Expected contingency table for nominal feature x and category y. $\hat{f}(uv)$ denotes the expected number of documents containing u and belonging to v under the null hypothesis of independence H_0.

	y	$\bar{y}$	$*$
x	$\hat{f}(xy) = \frac{f(x*)f(*y)}{N}$	$\hat{f}(x\bar{y}) = \frac{f(x*)(N-f(*y))}{N}$	$\hat{f}(x*)$
$\bar{x}$	$\hat{f}(\bar{x}y) = \frac{(N-f(x*))f(*y)}{N}$	$\hat{f}(\bar{x}\bar{y}) = \frac{(N-f(x*))(N-f(*y))}{N}$	$\hat{f}(\bar{x}*)$
$*$	$\hat{f}(*y)$	$\hat{f}(*\bar{y})$	N

of documents containing the feature x and belonging to the category y is denoted by $f(xy)$ and represents the document frequency. In general, $f(uv)$ denotes the number of documents containing u and belonging to v, u being x, $\bar{x}$ or $*$ (documents containing any term) and v being y, $\bar{y}$ or $*$ (documents belonging to any category). These frequencies are represented in the contingency Table (Table 1) in which the number of documents is denoted by N, $f(xy)$ by a and f_{11}, $f(x\bar{y})$ by b and f_{12}, and so on.

Many metrics are based on the estimation of the probability $P(uv)$ the probability that a document containing u belongs to the category v, u being x, $\bar{x}$ or $*$ and v being y, $\bar{y}$ or $*$. Under the maximum-likelihood hypothesis this probability is estimated by: $p(uv) = \frac{f(uv)}{N}$. Some metrics are based on the difference between the observed and the expected frequencies. The expected contingency frequencies under the null hypothesis of independence H_0 are given in the table 2.

Giving a weight to a feature x associated to a term in a document labeled with y depends on the correlation between x and y in the training corpus. This correlation can be estimated by different metrics, all the metrics used in this paper depends only on four values: N the number of training documents, $f(xy)$ the joint frequency, $f(x*)$ and $f(*y)$ the marginal frequencies. Given these values one can compute the contingency table and than compute any of the metrics described in Table 3. The first 10 metrics of Table 3 are those already been used for the problem of term weighting in the literature [4, 9, 5, 6, 7, 8]. The last 2 metrics Loevinger and One-way Klosgen are proposed by the authors of this paper. These metrics are collected from papers dealing with association rules and classification rules [10] and were not used for supervised term weighting in the litterature.

2.3 Benchmark

In order to compare experimentally the metrics, we use the Ohsumed corpus. Ohsumed is a test collection that includes 13,929 medical abstracts (6,286 for training and 7,643 for testing) from MEDLINE indexed by 23 cardiovascular diseases MeSH categories. Ohsumed is small when compared to the entire MEDLINE corpus that contains over 21 million references indexed by 27,149 descriptors in 2014 MeSH. However it was necessary in the first instance to use a small dataset for all the experiments we have done, namely 120 learn/prediction tasks with 12 metrics, 5 different weighting schemes and two machine learning methods (see table 5 and 6 next section).

We have done a summary preprocessing on these data and did not use feature selection in order to compare the weighting metrics independently from other methods of selecting the terms. Each document was stemmed (Porter stemming) and reduced to a vector of features representing 1-grams or 2-grams terms. Traditionally the performance of a classifier on a corpus is estimated by learning the classification on the training data

Table 3: Metrics used for supervised feature weighting

Metric	Mathematical form
IDF	$\log(\frac{N}{f(x*)})$
Pearson's χ^2 test	$\sum_{i,j} \frac{(f_{ij}-\hat{f_{ij}})}{\hat{f_{ij}}}$
Information gain	$\sum_{u\in\{x,\bar{x}\}} \sum_{v\in\{y,\bar{y}\}} p(uv) \log \frac{p(uv)}{p(u*)p(*v)}$
Odds ratio	$\frac{ad}{bc}$
Log odds ratio	$\log \frac{ad}{bc}$
Bi-normal separation (BNS)	$\lvert F^{-1}(p(x\lvert y)) - F^{-1}(p(x\lvert\bar{y}))\rvert$ (*)
Probability based term weight	$\log(1+\frac{a^2}{bc})$
Pointwise mutual information	$\log \frac{p(xy)}{p(x*)p(*y)}$
Relevance frequency	$\log_2(2+\frac{a}{\max(b,1)})$
Relevance frequency$_{OR}$	$\log_2(2+\frac{a}{\max(b,1)})(1-(p(x\lvert y)-p(x\lvert\bar{y})))$
Relevance frequency$_{\chi^2}$	$\log_2(2+\frac{a}{\max(b,1)})\lvert p(x\lvert\bar{y})-p(x\lvert y)\rvert$
One-way Klosgen	$\sqrt{p(xy)}(p(y\lvert x)-p(*y))$
Loevinger	$1-\frac{p(x*)p(*\bar{y})}{p(x\bar{y})}$

(*) F^{-1} is the inverse Normal cumulative distribution function.

Table 4: Experimented term frequency weights as a function of the frequency *tf*(x,d) of a feature x in a document d

Term frequency weight	Value	Description
BIN(x,d)	1 if *tf*$(x,d)>0$, 0 otherwise	binary weight
RTF(x,d)	*tf*(x,d)	raw term frequency
LTF(x,d)	$\log(1+$*tf*$(x,d))$	term frequency logarithm
ITF(x,d)	$1-\frac{1}{1-$*tf*$(x,d)}$	inverse term frequency

and evaluating the accuracy of the prediction obtained on the evaluation data. The evaluation metrics used are the *precision* which is the proportion of documents placed in the category that are really in the category, *recall* which is the proportion of documents in the category that are actually placed in the category, and the F_1-Score is defined as: $F_1\text{-Score} = \frac{2\,.\,precision\,.\,recall}{precision\,+\,recall}$. The microaveraged F_1-Score is computed globally for all the categories, while the macroaveraged F_1-Score is the average of the F_1-Scores computed for each category. This later measures the ability of a classifier to perform well when the distribution of the categories is unbalanced, while the microaveraged F_1-Score gives a global view of the document classification performance.

3 Results

For each category y, every document d is transformed to a vector W_d where each feature x is weighted by : $w(x,y,d) = w_{TF}(x,d) \times w_{DF}(x,y)$. Where term frequency weight w_{TF} (see Table 4) depends on the frequency of x in the document d and document frequency weight w_{DF} is one of the metrics described in Table 3. The feature x represents either a term feature t, a term frequency feature (t,n) or a term position feature (t,p) as defined in section 2.1. For the classical term representation, we have experimented three possible term frequency weights (see Table 4). For our model, we use only binary term weights ($w_{TF}(x,d) = \text{BIN}(x,d)$), because the frequency of the term is already considered in the extended term representation $x=(t,n)$.

In order to estimate the performance of both our model and the 2 metrics we propose, we have compared the F_1-Score of SVM and nearest centroid classification on Ohsumed documents with classical and extended term representations using different weighting schemes. For each document frequency weight metric w_{DF} we have experimented 5 weighting schemes:

- raw term frequency weight ($w_{TF} = \text{RTF}$) for term features t
- term frequency logarithm weight ($w_{TF} = \text{LTF}$) for term features t

Table 5: F_1-Scores of nearest centroid classifier with different term representations and weighting metrics

Term representation Term frequency weight	t RTF	t LTF	t ITF	(t,n) BIN	(t,n) &(t,p) BIN
Microaveraged F_1-Score					
Loevinger	0.585	0.603	0.610	0.620	0.628
IDF	0.490	0.564	0.584	0.604	0.616
Pointwise mutual information	0.506	0.548	0.563	0.585	0.593
Log odds ratio	0.516	0.547	0.559	0.577	0.590
Odds ratio	0.515	0.527	0.530	0.547	0.563
Relevance frequency	0.394	0.465	0.489	0.513	0.532
Bi-normal separation	0.461	0.491	0.500	0.518	0.531
Relevance frequency$_{OR}$	0.388	0.456	0.479	0.503	0.518
One-way Klosgen	0.468	0.481	0.486	0.495	0.502
Pearson's χ^2	0.444	0.452	0.456	0.463	0.468
Information gain	0.345	0.346	0.347	0.357	0.373
Relevance frequency$_{\chi^2}$	0.319	0.334	0.335	0.325	0.330
Macroaveraged F_1-Score					
Loevinger	0.568	0.584	0.591	0.604	0.611
IDF	0.441	0.536	0.560	0.590	0.603
Log odds ratio	0.492	0.535	0.548	0.570	0.579
Pointwise mutual information	0.476	0.528	0.545	0.570	0.577
Odds ratio	0.517	0.524	0.526	0.543	0.564
Relevance frequency	0.384	0.463	0.491	0.522	0.542
Relevance frequency$_{OR}$	0.390	0.458	0.481	0.509	0.526
Bi-normal separation	0.439	0.476	0.484	0.505	0.513
One-way Klosgen	0.457	0.468	0.473	0.485	0.491
Pearson's χ^2	0.440	0.444	0.446	0.454	0.462
Information gain	0.366	0.375	0.384	0.400	0.401
Relevance frequency$_{\chi^2}$	0.337	0.352	0.368	0.360	0.365

Table 6: F_1-Scores of SVM classifier with different term representations and weighting metrics

Term representation Term frequency weight	t RTF	t LTF	t ITF	(t,n) BIN	(t,n) &(t,p) BIN
Microaveraged F_1-Score					
One-way Klosgen	0.587	0.604	0.609	0.631	0.639
Pearson's χ^2	0.593	0.598	0.600	0.618	0.629
Odds ratio	0.563	0.582	0.590	0.617	0.629
Loevinger	0.563	0.579	0.586	0.614	0.626
Bi-normal separation	0.553	0.586	0.593	0.614	0.623
Information gain	0.570	0.583	0.586	0.603	0.615
Relevance frequency$_{\chi^2}$	0.548	0.568	0.571	0.590	0.602
Log odds ratio	0.497	0.545	0.556	0.587	0.600
Relevance frequency$_{OR}$	0.475	0.531	0.541	0.571	0.588
Relevance frequency	0.460	0.521	0.535	0.564	0.583
Pointwise mutual information	0.459	0.520	0.533	0.566	0.582
IDF	0.296	0.363	0.380	0.417	0.444
Macroaveraged F_1-Score					
One-way Klosgen	0.538	0.569	0.575	0.595	0.602
Pearson's χ^2	0.553	0.562	0.568	0.587	0.598
Odds ratio	0.520	0.545	0.553	0.576	0.594
Loevinger	0.518	0.541	0.550	0.580	0.590
Information gain	0.529	0.547	0.550	0.560	0.578
Relevance frequency$_{\chi^2}$	0.501	0.522	0.524	0.540	0.565
Bi-normal separation	0.468	0.513	0.521	0.552	0.564
Log odds ratio	0.401	0.461	0.476	0.510	0.534
Relevance frequency$_{OR}$	0.384	0.450	0.462	0.504	0.523
Relevance frequency	0.365	0.435	0.453	0.486	0.515
Pointwise mutual information	0.353	0.421	0.439	0.480	0.501
IDF	0.185	0.237	0.255	0.289	0.319

- inverse term frequency weight (w_{TF} = ITF) for term features t
- binary term frequency weight (w_{TF} = BIN) for term frequency features (t, n)
- binary term frequency weight (w_{TF} = BIN) for term frequency features (t, n) and term position features (t, p)

The Table 5 reports the microaveraged and macroaveraged F_1-Score obtained with nearest centroid classifier considering different term representations and weighting metrics (the five table columns represent the five weighting schemes). After calculation of the F_1-Score for each classifier, the metrics are ranked in descending order of the best weighting scheme score. It is clearly observed from these results that the proposed representation model (t, n) & (t, p) performs significantly better than the classical representation (the three first columns) and achieves the best performances in all experiments in terms of microaveraged F1-scores for all the metrics. We can also observe that the proposed Loevinger metric yields to better microaveraged and macroaveraged F1-scores for all weighting schemes. Table 6 provides the F_1-Scores with SVM classification. It can be seen that by using the One-way Klosgen metric we obtain the best classification scores on Ohsumed data.

4 Conclusion

In this paper, we have proposed two term weighting metrics that have not been previousely used for term weighting in the literature. We have also proposed an extended term representation where the term frequency and the term position in the document are adequately integrated to the document frequency. We showed that both the metrics and the term representation can improve significantly the classification of Ohsumed biomedical documents. After this study we intend to assess our approach with large-scale experiments on all MEDLINE corpus with all MeSH categories.

References

[1] Manuel Wahle, Dominic Widdows, Jorge R Herskovic, Elmer V Bernstam, and Trevor Cohen. Deterministic binary vectors for efficient automated indexing of medline/pubmed abstracts. In *AMIA annual symposium proceedings*, volume 2012, page 940. American Medical Informatics Association, 2012.

[2] Minlie Huang, Aurélie Névéol, and Zhiyong Lu. Recommending mesh terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–667, 2011.

[3] Vidya Vasuki and Trevor Cohen. Reflective random indexing for semi-automatic indexing of the biomedical literature. *Journal of biomedical informatics*, 43(5):694–700, 2010.

[4] Franca Debole and Fabrizio Sebastiani. Supervised term weighting for automated text categorization. In *Proceedings of the 2003 ACM Symposium on Applied Computing (SAC), March 9-12, 2003, Melbourne, FL, USA*, pages 784–788. ACM, 2003.

[5] George Forman. BNS feature scaling: an improved representation over tf-idf for svm text classification. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 263–270. ACM, 2008.

[6] Ying Liu, Han Tong Loh, and Aixin Sun. Imbalanced text classification: A term weighting approach. *Expert Systems with Applications*, 36(1):690–701, 2009.

[7] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):721–735, 2009.

[8] Hakan Altinçay and Zafer Erenel. Using the absolute difference of term occurrence probabilities in binary text categorization. *Applied Intelligence*, 36(1):148–160, 2012.

[9] Zhi-Hong Deng, Shiwei Tang, Dongqing Yang, Ming Zhang, Liyu Li, and Kunqing Xie. A comparative study on feature weight in text categorization. In Jeffrey Xu Yu, Xuemin Lin, Hongjun Lu, and Yanchun Zhang, editors, *Advanced Web Technologies and Applications, 6th Asia-Pacific Web Conference, APWeb 2004, Hangzhou, China, April 14-17, 2004, Proceedings*, volume 3007 of *Lecture Notes in Computer Science*, pages 588–597. Springer, 2004.

[10] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3), 2006.

Improving genome assemblies using multi-platform sequence data

Pınar Kavak [1,2,*], Bekir Ergüner [1], Duran Üstek [3], Bayram Yüksel [4],
Mahmut Şamil Sağıroğlu [1], Tunga Güngör [2], and Can Alkan [5,*]

(1) Advanced Genomics and Bioinformatics Research Group (İGBAM)
BİLGEM, The Scientific and Technological Research Council of Turkey (TÜBİTAK),
41470 Gebze, Kocaeli, Turkey, pinar.kavak@tubitak.gov.tr

(2) Department of Computer Engineering
Boğaziçi University, 34342 Bebek, İstanbul, Turkey

(3) Department of Medical Genetics
İstanbul Medipol University, 34810 Beykoz, İstanbul, Turkey

(4) TÜBİTAK - MAM - GMBE (The Scientific and Technological Research Council of Turkey, Genetic Engineering and Biotechnology Institute), 41470 Gebze, Kocaeli, Turkey

(5) Department of Computer Engineering
Bilkent University, 06800 Bilkent, Ankara, Turkey, calkan@cs.bilkent.edu.tr

Keywords: *de novo* assembly, assembly improvement, next generation multi-platform sequencing.

Abstract. *De novo* assembly using short reads generated by next generation sequencing technologies is still an open problem. Although there are several assembly algorithms developed for data generated with different sequencing technologies, and some that can make use of hybrid data, the assemblies are still far from being perfect. There is still a need for computational approaches to improve draft assemblies. Here we propose a new method to correct assembly mistakes when there are multiple types of data obtained using different sequencing technologies that have different strengths and biases. We apply our method to Illumina, 454, and Ion Torrent data, and also compare our results with existing hybrid assemblers, Celera and Masurca.

1 Scientific Background

Since the introduction of high throughput next generation sequencing (NGS) technologies, traditional Sanger sequencing is being abandoned especially for large-scale sequencing projects. Although cost effective for data production, NGS also imposes increased cost for data processing and computational burden. In addition, the data quality is in fact lower, with greater error rates, and short read lengths for most platforms. One of the main algorithmic problems to analyze NGS data is the *de novo* assembly: i.e. "stitching" billions of short DNA strings into a collection of larger sequences, ideally the size of chromosomes. However, "perfect" assemblies with no gaps and no errors are still lacking due to many factors, including the short read and fragment (paired-end) lengths, sequencing errors in basepair level, and the complex and repetitive nature of most genomes. Some of these problems in *de novo* assembly can be ameliorated through using data generated by different sequencing platforms, where each technology has "strengths" that may be used to fix biases introduced by others.

*to whom correspondence should be addressed

Overlap-layout-consensus (OLC) graph based assemblers [1, 2] work well on the long read assembly. Assemblers that are based on de Bruijn graphs [3, 4, 5] are designed primarily for short reads. Several assemblers use multiple read libraries [6, 7, 8] for better assembly construction. Additionally, strategies to merge different assemblies using different data sources into a single coherent assembly are described in literature (e.g. [9]). Our method differs from that of [9], in both pre- and post-processing steps.

In this work, we propose a method to improve draft assemblies (i.e. produced using a single data source, and/or single algorithm) by incorporating data generated by different NGS technologies, and applying novel correction methods. To achieve better improvements, we exploit the advantages of both short but low-error-rate reads and long but erroneous reads. We show that correcting the contigs built by assembling long reads through mapping short and high quality read contigs produce the best results, compared to the assemblies generated by algorithms that use hybrid data.

2 Materials and Methods

A part of human chromosome 13 was cloned into a bacterial artificial chromosome (BAC) in a previous study. We sequenced the BAC clone separately using Illumina, Roche/454, and Ion-Torrent platforms (see Table 1). A "gold standard" reference assembly was also obtained using template-based assembly with Mira [8] using Roche/454 data, which is then corrected using the Illumina reads. Since Roche/454 and Ion Torrent platforms have similar sequencing biases (i.e. problematic homopolymers), we separated this study into two different groups: Illumina & 454 and Illumina & Ion-Torrent, which gives us the opportunity to compare Roche/454 and Ion-Torrent data.

Table 1: Properties of the data

Technology	Length range	Mean length	Mean base qual (phred s.)	Paired
Illumina	101bp (all reads have equal length)	101bp	38	paired
Roche/454	40bp-1027bp	650bp	28	single-end
Ion-Torrent	5bp-201bp	127bp	24	single-end

Technology: The name of the sequencing technology used to produce the reads. **Length range**: Minimum and maximum lengths of the generated reads. **Mean length**: The mean length among all reads. **Mean base qual**: The average phred score sequence quality of all reads. Calculated by summing up all phred scores of the bases in a read and dividing it to sequence length over all reads. **Paired:** Represents whether the sequencing is performed as paired-end or single-end.

Pre-processing: First, the reads that has low average quality value (phred score 17, i.e. $\geq$2% error rate) were discarded. Then, the reads with high N-density (with $>$10% of the read consisting of Ns) were removed. Third, groups of bases that seem to be non-uniform according to sequence base content were trimmed. Finally, each assembler's pre-processing operations were also inevitably applied.

Assembly: Several assembly tools were used: Velvet [3], a de Bruijn graph based assembler to assemble the short reads; and two different overlap-layout-consensus (OLC) assemblers: Celera [1], and SGA [2] to assemble the long read data sets (Roche/454 and Ion Torrent) separately. In addition, a de Bruijn graph based assembler, SPAdes [4] was also used on the long read data. Then all draft assemblies were mapped to the E. coli reference sequence using BLAST [10] to identify and discard E. coli contamination due

to the cloning process. At the end, one short read, and three long read assemblies were obtained.

Correction: We used BLAST [10] to map the contigs obtained with the short reads onto the contigs generated by assembling the long reads. Since BLAST may report multiple mapping locations due to repeats, only the "best" map locations were accepted. Reasoning from the fact that the short reads show less sequencing errors, the sequence reported by the short read based contigs were preferred over the long read contigs (LRC) when there are disagreements between the pair. By doing this, the "less fragmented" long read assemblies were patched. Figure 1 shows a visual representation of the strategy on correcting the LRCs. The strategy we applied is as follows: if there is a mapping between a short read contig (SRC) and a LRC, and if the mapping does not start at the beginning of the LRC, add the unmapped prefix of the LRC. Also, if the mapping does not start at the beginning of the SRC (very rare situation), add the unmapped prefix of the SRC with lowercase letters. On the mapping part between SRC and LRC, pick the SRC values. If the mapping does not end at the end of the SRC (rare), add the unmapped suffix of the SRC, again with lowercase letters. One may argue that it might disturb the continuity of the resulting contig, however, we observe such mapping properties very rarely. The reason for using lowercase letters is to keep track of the information that there is a disagreement between the SRC and LRC on these sections, so the basepair quality will be lower than other sections of the assembly. Finally, add the unmapped suffix of the LRC and obtain the corrected contig.

Evaluation: We mapped each of the final corrected assemblies to the "gold standard" reference assembly we constructed (described above), and calculated various statistics based on the comparisons, and estimated assembly qualities (Table 2). We also used two hybrid assemblers, Celera-CABOG [6] and Masurca [7], with the same data to compare our correction methodology with those of hybrid assembly algorithms.

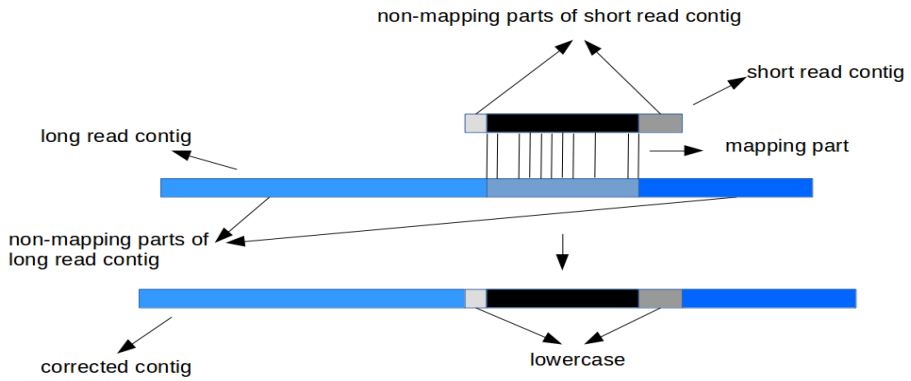


Figure 1: Correction method: Correct the long read contig according to the mapping information of the short read contig.

3 Results

A summary of the results is presented in Table 2. Briefly, the Velvet assembly using only the Illumina reads showed better coverage (99%) and high average identity (97.5%) rates compared to Celera assembly using long reads. Correcting the Celera assembly with our method improves both coverage and average identity rates, which are then further improved by reiterative application of our method.

The coverage of 454 assembly increases up to 99.7% and the average identity rate increases up to 94.4% on the first correction cycle. The repetitive correction cycles increase the coverage and average identity rates. The cycles are stopped if there is no improvement (≥ 0.001) or decrease on the average identity because of the increase on the coverage. One can see that correcting the long read assembly with the SRCs works

well with all kind of assemblers. However, corrected SGA assembly has the highest coverage rate among all.

Assembling short and long reads separately with de Bruijn and OLC graph based assemblers and correcting them give better results than assembling short and long reads together with a hybrid assembler such as Masurca or Celera. Masurca seems to have the best average identity rate on Illumina-Ion Torrent data, but the coverage for this run is just 1%. Celera-CABOG performs very well on Illumina-454 data, but no better than corrected SGA or corrected Celera with Illumina and 454. Celera-CABOG does not have any contigs which successfully map onto the reference sequence, with Illumina-Ion-Torrent data, because all 487 resulting contigs were eliminated on the E.coli contamination filtering phase.

Table 2: Results of assembly correction method on BAC data.

Name	Length	# of Contigs	# of Mapped Contigs	# of Covered bases	Coverage	Avg. Identity	# of Gaps	Size of Gaps
Reference	*176.843*							
Velvet								
Ill. Velvet	197,040	455	437	175,172	0.99055	0.97523	39	1,671
Celera								
454 Celera	908,008	735	735	172,563	0.97580	0.92599	18	4,280
Ion Celera	39,347	27	27	47,638	0.26938	0.96932	47	129,205
Corrected Celera								
Ill-454 Celera	4,945,785	895	270	176,368	0.99731	0.94370	5	475
Ill-454 Celera^{2*}	5,078,059	890	265	176,640	0.998852	0.944527	4	203
Ill-Ion Celera	93,909	30	28	81,819	0.46267	0.96327	36	95,024
Ill-Ion Celera^2	145,262	30	28	91,962	0.52002	0.97412	33	84,881
Ill-Ion Celera^3	216,167	30	28	99,645	0.56347	0.98066	34	77,198
SGA								
454 SGA	62,909,254	108,095	101,514	176,546	0.99832	0.97439	1	297
Ion SGA	842,997	6,417	6,122	153,092	0.86569	0.99124	197	23.751
Corrected SGA								
Ill-454 SGA	295,009	335	335	176,757	0.99951	0.96823	5	86
Ill-454 SGA^2	279,034	305	305	176,757	0.99951	0.96769	5	86
Ill-Ion SGA	197,509	291	291	175,052	0.98987	0.97501	45	1,791
Ill-Ion SGA^2	203,064	291	291	175,676	0.99340	0.97413	34	1,167
SPADES								
454 SPADES	12,307,761	49,824	49,691	176,843	1.0	0.98053	0	0
Ion SPADES	176,561	110	107	167,890	0.94937	0.92909	9	8,953
Corrected SPADES								
Ill-454 SPADES	290,702	298	298	176,454	0.99780	0.96538	5	389
Ill-Ion SPADES	198,665	52	52	171,977	0.97248	0.94215	4	4,866
Ill-Ion SPADES^2	200,307	52	52	172,101	0.97319	0.94230	2	4,742
Masurca								
Ill-454 Masurca	380	1	0	0	0	0	0	0
Ill-Ion Masurca	2,640	8	8	1,952	0.01104	0.98223	9	174,891
Celera-CABOG								
Ill-454 Celera	1,101,716	891	891	174,330	0.98579	0.92452	12	2,513
Ill-Ion Celera	0	0	0	0	0.0	0.0	0	0.0

Name: the name of the data group that constitute the assembly; # of contigs: the number of contigs that belong to the resulting assembly; # of Mapped Contigs: the number of contigs that successfully mapped onto the reference sequence; # of Covered bases: the number of bases on the reference sequence that are covered by the assembly; Coverage: percentage of covered reference; Avg. identity: percentage of the correctly predicted reference bases; # of Gaps: The number of gaps that cannot be covered on the reference genome; Size of Gaps: total number of bases on the gaps.

* "2" represents the results of the second cycle of correction, "3" represents the third cycle.

4 Conclusion

Assembly correction by using advantages of different technologies improves the resulting assembly. Here we presented a new method to improve draft assemblies by correcting high contiguity assemblies using the contigs obtained with high quality reads.

Our results show that our method is useful and it gives better results than using all data for once with a hybrid assembler compared to the results of two hybrid assemblers. However, the need to develop new methods that exploit different data properties of different NGS technologies, such as short/long reads or high/low quality of reads, remains. In this manner, as future work, our correction algorithm can be improved by exploiting the paired end information of the short, high quality reads after the correction phase, to fill in the gaps between corrected contigs.

Funding

The project is supported by the Republic of Turkey Ministry of Development Infrastructure Grant (no: 2011K120020), BİLGEM - TÜBİTAK (The Scientific and Technological Research Council of Turkey) grant (no: T439000), and a TÜBİTAK grant to C.A.(112E135).

References

[1] E.W.Myers *et al* (2000) A Whole-Genome Assembly of Drosophila, *Science*, 287(no:5461):2196-2204, doi:10.1126/science.287.5461.2196.

[2] J.Simpson *et al* (2012) Efficient *de novo* Assembly of Large Genomes Using Compressed Data Structures, *Genome Research*, 22:549-556, doi:10.1101/gr.126953.111.

[3] D.Zerbino, E.Birney (2000) Velvet: Algorithms for *de novo* Short Read Assembly Using de Bruijn Graphs, *Genome Research*, 18(5):821-829, doi: 10.1101/gr.074492.107.

[4] A.Bankevich *et al* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, *Journal of Computational Biology*, 19(5):455-477, doi:10.1089/cmb.2012.0021.

[5] J.Butler *et al* (2008) ALLPATHS: *De novo* Assembly of Whole-Genome Shotgun Microreads, *Genome Research*, 18(5):810-820, doi:10.1101/gr.7337908.

[6] J.R.Miller *et al* (2008) Aggressive Assembly of Pyrosequencing Reads with Mates, *Bioinformatics*, 24(24):2818-2824, doi:10.1093/bioinformatics/btn548.

[7] A.Zimin *et al* (2013) The MaSuRCA Genome Assembler, *Bioinformatics*, 29(21):2669-2677, doi:10.1093/bioinformatics/btt476.

[8] B.Chevreux *et al* (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information, *Computer Science and Biology:Proceedings of the German Conference on Bioinformatics (GCB)*, 99:45-56.

[9] Y.Wang *et al* (2012) Optimizing Hybrid Assembly of Next-Generation Sequence Data from Enterococcus Faecium: a Microbe with Highly Divergent Genome, *BMC Systems Biology*, 6(Suppl 3):S21, doi:10.1186/1752-0509-6-S3-S21.

[10] S.Altschul *et al* (1990) Basic Local Alignment Search Tool, *Journal of Molecular Biology*, 215(3):403-410.

A simulation study of the relationship between replication stress detection pathway and the cell cycle

Monika Kurpas(1), Krzysztof Puszynski(1)

(1) Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology
Akademicka 16, 44-100 Gliwice, Poland, monika.kurpas@polsl.pl

Keywords: cell cycle, ATR, deterministic model, DNA, damage detection.

Abstract. Ataxia telangiectasia mutated and Rad3-related (ATR) detects single-stranded DNA areas (ssDNA) caused by stalled replication forks. ATR-p53 pathway induces cell cycle arrest, necessary to repair damage. Using mathematical modeling we examined how detection of ssDNA influences the cell cycle. Our model confirms that the cell cycle phase, during which ssDNA are detected, also has the impact on genetic material susceptibility to damage. Our results indicate that during cell cycle progression, with increasing cell size, cellular DNA becomes more prone to damage than in early stages of cycle. Reaction speed rates, which decrease over the cycle, and degree of DNA condensation have an impact on strength of DNA damage response. This result may explain why cells from heterogeneous population exhibit different responses to radiation, what is commonly observed during biological studies performed on the cell culture.

1 Scientific Background

The cell cycle is the sequence of events during which the genetic material is duplicated and equally segregated to two daughter cells. The cell enters the mitotic cycle with G1 phase, during which the cell component grows and the volume of the cell increases. G1 phase is then followed by the S phase of DNA replication and G2 phase of preparation of cell division machinery. The mitotic M phase of cell division starts when the previous steps were finished without errors. Additionally G0 phase can be distinguished, as quiescence or senescence phase, containing nonproliferative cells. Specialized mechanisms checking the integrity of the DNA, called cell cycle checkpoints, are present among others between G1 and S as well as G2 and M phases. G1/S checkpoint verifies if DNA is undamaged and suitable for replication, while G2/M replication checkpoint verifies whether all genetic material was replicated and the lesions repaired before the entry into mitotic division.

Single-stranded DNA areas are often formed after the blockade of the replication forks progression due to the presence of modifications of the DNA chain, such as 6-4 photoproduct or pirymidine dimer. These forms are often observed after irradiating cells with ultraviolet radiation (UV). Even without treating cells with any exogenous factor, we can observe spotaneous lesions formation due to replication errors. Areas of ssDNA arise also as an effect of double-strand breaks repair that require the resection of DNA free ends. Detection of such forms is performed by ATR module [1].

P53, the main effector of ATR pathway known also as the ”guardian of the genome”, is involved in many cellular processes, like DNA repair, cell cycle arrest or apoptosis. The major inhibitor of p53 is mouse double minute 2 homolog (Mdm2), the E3 ubiquitin ligase, which marks p53 to the degradation in proteasome. P53 is phosphorylated by ATR, checkpoint kinase 1 (Chk1) and 2 (Chk2), what additionally ampificates a detection signal. Chk1 and Chk2 also interact with proteins and complexes regulating the cell cycle (mainly with components of cyclin-CDK complexes) [2]. P53 transcriptionally upregulates p21 protein, the potent inhibitor of cyclin-dependent kinases (CDK) [5].

Very important pair of cyclin-CDK cell cycle regulatory complexes are cyclin B and Cdk1 protein kinase. Active complexes drive transition between S, G2 and metaphase of M phase and are responsible for DNA replication, chromosome condensation and mitotic spindle assembly. They are inactivated by the ubiquitin ligase anaphase-promoting complex/cyclosome (APC/C), which destroys the cyclins responsible for activation of Cdk1. When the cyclins are degraded, Cdk gets inactivated and APC dominate through G1 phase. In mammalian cells the APC complex consists of core with many polypeptide subunits, and two activators: Cdh1 and Cdc20. responsible for recognizing specific target proteins. At the end of G1 phase cyclin synthesis is induced and their degradation is inhibited, what causes rising cyclin-CDK activity [3].

1.1 *Existing models*

There are many models of cell cycle (reviewed in [4]), but very few cover the whole cycle for non-embryonic eukaryotic cells. The good example of cell cycle model for eukaryotes is Tyson's work [3]. It describes cyclin/CDK and APC interactions, making the core of cell cycle machinery. To our knowledge, except Iwamoto *et al.* work [5], there is lack of models of the cell cycle that contain additionally DNA damage detection module and describe the interactions between proteins and complexes belonging to both parts. According to our knowledge the proposed model is the first which takes into account the impact of cell cycle phase on ssDNA fragments formation after UV irradiation and detection of these lesions.

2 **Materials and Methods**

A detection system is activated by UV radiation resulting in formation of ssDNA fragments. Recognition of ssDNA is initiated by coating of single DNA strand by replication protein A (RPA) complexes, what induces independent movement of Rad17-RFC, Rad9-Rad1-Hus1 (9-1-1) and ATR-ATRIP complexes to the site of the damage. ATR interacting protein (ATRIP) allows to bind ATR-ATRIP complex to RPA-coated DNA strand, causing ATR autophosphorylation on Ser1989 [1]. ATR protein is then capable to phosphorylate 9-1-1 complex that after activation recruits topoisomerase 2-binding protein 1 (TopBP1) important for full ATR activation (through phosphorylation on other serine residue). Full activation of ATR enables to recognize its phosphorylation targets, among others Rad17, claspin, RPA, ATRIP, Chk1, Chk2, p53 and H2AX histone.

Presented model is a simplification of processes occuring in the cell during the detection of ssDNA. Some of the described above interactions are simplified to reduce the model complexity and the time necessary for completing the calculation. We distinguished two cell compartments: nucleus and cytoplasm. Our model takes into account cell volume, which varies between cell cycle phases and has impact on the reactions speed.

Our model was built with ordinary differential equations (ODE) with the use of basic laws known from biochemistry (the law of mass action, Michaelis-Menten kinetics). We used equations from Tyson's work [3], which serve as cell cycle core of our model. We combined them with our previously created model of ATR-p53 pathway [6]. Equations from ATR-p53 part of the model were multiplied by cell mass from cell cycle model in order to show dependence of cellular signaling kinetics on cell mass. We rescaled also Tyson's model to time-scale and number of molecules observed in average cell line. We assumed that cell cycle duration is about 24 hours (average time of division in typical proliferating human cell), but the model can also be rescaled to other cycle lengths. We found in literature, that maximum amount of cyclin B in cells is 1600000 molecules [8]. We rescaled the model also for this value (fig. 2B). Moreover, we selected mammalian homologs of proteins used in Tyson's work. We used cyclin B-Cdk1 complexes, as well as APC/C activators Cdh1 and Cdc20, and polo-like kinase 1 (Plk1). Plk1 is the im-

portant cell cycle component necessary among others to create a time lag (as observed) between the rise in cyclin B-Cdk1 activity and the activation of Cdc20.

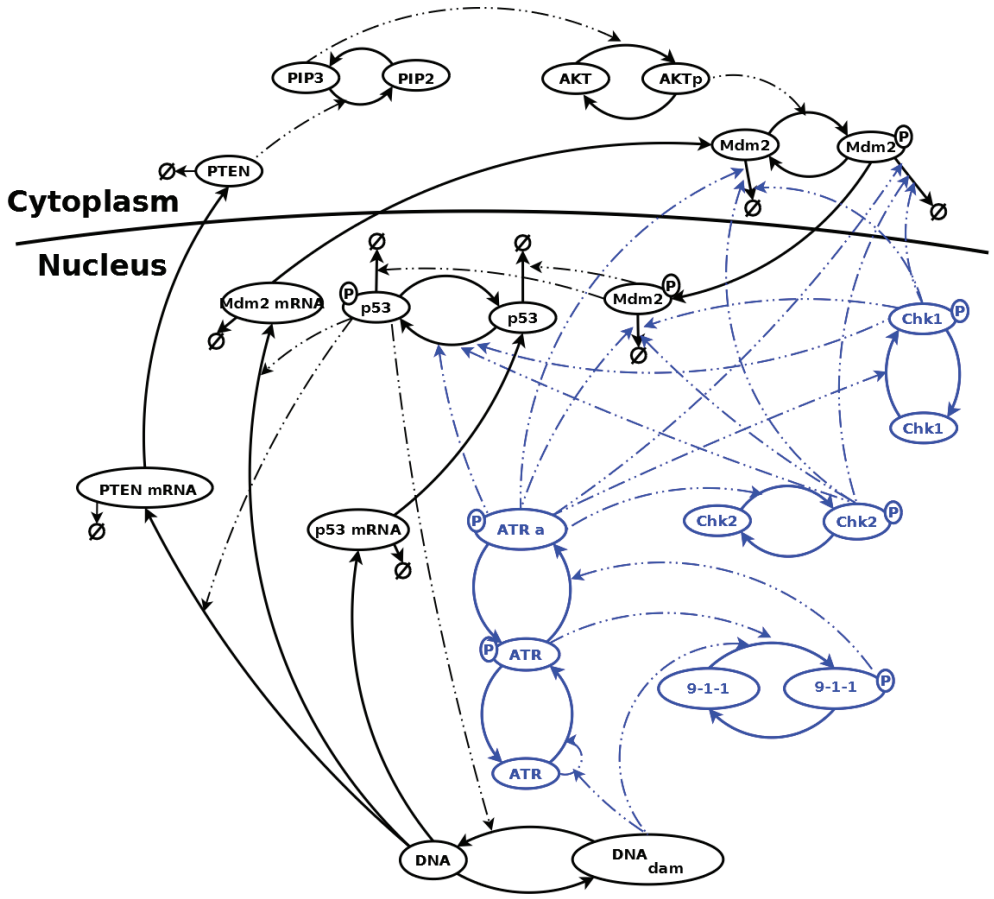


Figure 1: **Schematic representation of the model of ssDNA detection.** Solid lines represent changes in the form of proteins. Dashed lines represent interactions between the model elements. The ATR detector module is marked with blue color [6].

Equation for cyclin B from Tyson's work [3] after modification:

$$\begin{aligned} \frac{d}{dt}CYCB(t) &= k1 - (k2' + \frac{k2''}{resc}CDH1(t) + cd1 \cdot P53p(t) \\ &+ cd2 \cdot CHK1(t) + cd3 \cdot CHK2(t))CYCB(t) \end{aligned} \tag{1}$$

Equation for Cdh1 from Tyson's work [3] after modification:

$$\begin{aligned} \frac{d}{dt}CDH1(t) &= \frac{(k3' \cdot resc + k3'' \cdot CDC20A(t)) \cdot (resc - CDH1(t))}{J3 + 1 - \frac{CDH1(t)}{resc}} \\ &- \frac{k4 \cdot CDH1(t) \cdot resc \cdot k4chk1 \cdot CHK1(t) + \frac{M(t)}{resc} \cdot \frac{CYCB(t)}{resc}}{J4 + \frac{CDH1(t)}{resc}} \end{aligned} \tag{2}$$

New parameters:

- cd1 – active p53-dependent inactivation of cyclin B-Cdk1 complex (through transcriptional regulation of p21) [5]
- cd2 – active Chk1-dependent inactivation of cyclin B-Cdk1 complex (through phosphorylation and inactivation of Cdc25) [2]
- cd3 – active Chk2-dependent inactivation of cyclin B-Cdk1 complex (through phosphorylation and inactivation of Cdc25) [2]

- k4chk1 − active Chk1-mediated degradation of Cdh1 [9]
- resc − scaling factor for cell cycle

The output of the model is the levels of total concentration and active forms of p53, Chk1, Chk2 what determines the cell fate. According to Kracikova *et al.* report [7], cell fate depends on the level of active p53. Two threshold values for p53 are distinguished: lower, which causes cell cycle arrest (naturally present in the model, as an effect of lesions occurrence) and higher responsible for activation of apoptotic pathway. If p53 level is above this threshold, p53 induces apoptosis through cooperation with Bax protein. We determined this threshold as equal $2.1 \cdot 10^5$ by simulation analysis of model behaviour. If p53 level is continously elevated above this value for more than 6 hours, we consider this state as apoptotic for more than half of population. We do not take into account further levels of p53 or level of damage, because due to the degradation of cell components both levels might be increased.

The developed model is a general attempt to mapping the interaction between the paths of SSB detection and cell cycle and it is not designed for any particular cell line.

3 **Results**

Panel 2B shows normal functioning cell cycle. Cell division occurs when level of cyclin B drops under the treshold of 255800 molecules. Cell cycle is controlled by detection module through active p53-dependent [5], Chk1-dependent and Chk2-dependent cyclin B deactivation. A very important role is played also by Chk1-dependent Cdh1 inactivation [9], which gives the cell time to repair its damage.

3.1 *Spotaneous damage formation*

In our model, we implemented replication stress. It is the basal damage level which continously stimulates ATR-p53 pathway (fig. 2A). Most of ssDNA fragments are detected during S phase when genetic material is replicated. It results in elevated p53 level during this period and increased susceptibility to the damage (fig. 2E and 2F).

3.2 *Correlation between cell cycle phase and susceptibility to the damage*

We performed simulation analysis, to examine how the cell cycle phase influences DNA damage detection and repair. For this purpose, we simulated cell treatment with dose of 18 J/m^2, which we considered as apoptosis threshold in our previous work [6]. We observed that number of lesions and also repair time strictly depends on cell cycle phase.

The least damages occur in G1 phase (fig. 2C and 2D) when cell is condensed, has only one copy of the coiled genome and reaction rates are fast because of small cell volume (there is the biggest probability of molecules meeting). The cell cycle length was only changed slightly.

In S phase, when genetic material is replicated, probability of lesions arising increases significantly (fig. 2E). Unwinded, prepared for replication DNA is more lesion prone. Also repair time is extended. As mentioned above, during S phase of cell cycle occurs the additional replication stress which is normal phenomenon occuring in the course of replication. Cell cycle progression is being arrested, until all genetic material is repaired. According to Fingar *et al.* [10], growth of the cell mass is independent of cell division − even when cell cycle is arrested, the mass of the cell is still increasing (fig. 2F).

During G2 phase of cell cycle, when genetic material is replicated, cell is preparing for mitosis and size of the cell is almost doubled. It implicates decreased reactions speeds and biggest likelihood of the damage. However, DNA lesions are still repaired because with an increase of cell size, number of repair complexes is also grow-

ing (fig. 2G). In this case, long-term cell cycle arrest prevents cell before entry to the mitotic division (fig. 2H).

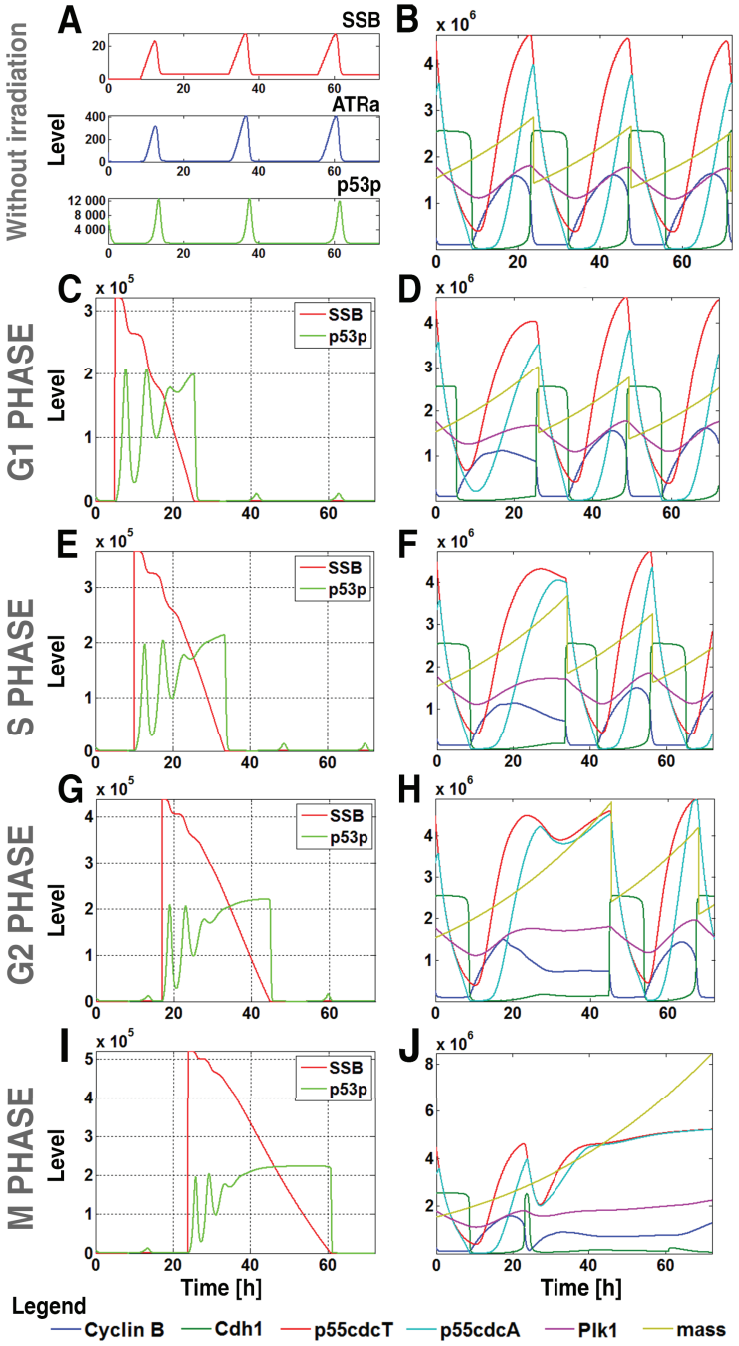


Figure 2: **Impact of irradiation on cells in various stages of cell cycle.** A-B cell cycle without irradiation, with considered spontaneous damage during S phase: level of damage, active ATR and active p53 protein (A), core of the cell cycle (B); C-J irradiation with dose of 18 J/m^2 in various phases of cell cycle: level of damage and active p53 protein (left panel), core of the cell cycle (right panel). The mass of the cell was rescaled to the order of magnitude of proteins involved in cell cycle regulation and does not reflect the real mass of the cell.

Irradiation during mitotic division destabilizes functioning of all cellular components. Lesions are repaired very slowly and active p53 is elevated continously during the long period of time (fig. 2I). DNA damage is too extensive to be repaired. Cell do not pass the correct division (fig. 2J). As mentioned before, when the p53 level in the cell is elevated above given thereshold for at least 6h, the cell is considered as apoptotic so the future mass growth in the fig. 2J is irrelevant.

4 Conclusion

Cell cycle regulation is a complicated mechanism still not clearly understood. Mathematical models like the model presented in this work might serve to broaden our knowledge about cellular interactions without costly and long lasting biological experiments.

Our results obtained using the deterministic model indicate that ATR module DNA damage response is strictly correlated with cell cycle phase. Moreover interactions between cell cycle and detection module may serve as potential anticancer therapeutic target.

In the future we plan to investigate the response of heterogeneous population of cells to irradiation. We will measure fractions of apoptotic cells to recreate conditions in the in vitro culture. We also plan to perform biological experiments which will give us necessary models parameters and verification data for our models.

Acknowledgments

This project was funded by the Polish National Center for Science granted by decision number DEC-2012/05/D/ST7/02072 (K.P.) and by BKM-514/RAU1/2015 : 18 (M.K.).

References

[1] S. Liu, B. Shiotani, M. Lahiri, A. Marechal, A. Tse, *et. al.* "ATR Autophosphorylation as a Molecular Switch for Checkpoint Activation". *Molecular Cell*, 43:192-202, 2011.

[2] H.C. Reinhardt and M.B. Yaffe. "Kinases that Control the Cell Cycle in Response to DNA Damage: Chk1, Chk2, and MK2". *Current Opinion in Cell Biology*, 21(2):245–255, 2009.

[3] J.J. Tyson and B. Novak. "Regulation of eukaryotic cell cycle: molecular antagonism, hysteresis and irreversible transitions". *Journal of Theoretical Biology*, 210:249–263, 2001.

[4] J.E. Ferrel Jr., T.Y-Ch. Tsai, Q. Yang. "Modeling of cell cycle: why do certan circuits oscillate?". *Cell*, 144:874–885, 2011.

[5] K. Iwamoto, H. Hamada, Y. Eguchi, M. Okamoto. "Mathematical modeling of cell cycle regulation in response to DNA damage: exploring mechsnisms of cell-fate determination". *BioSystems*, 103:384–391, 2011.

[6] M. Kurpas, K. Jonak, K. Puszynski. "Simulation Analysis of the ATR Module as a Detector of UV-Induced DNA Damage". *Information Technologies in Biomedicine*, 3:317–326, 2014.

[7] M. Kracikova, G. Akiri, A. George, R. Sachidanandam, S.A. Aaronson. "A threshold mechanism mediates p53 cell fate decision between growth arrest and apoptosis". *Cell Death and Differentiation*, 20:576–588, 2013.

[8] P.S. Frisa, J.W. Jacobberger. "Cell Cycle-Related Cyclin B1 Quantification". *PloS ONE*, 4(9).

[9] M. Yamada, K. Watanabe, M. Mistrik, E. Vesela, I. Protivankova, N. Mailand, M. Lee, H. Masai, J. Lukas, J. Bartek. "ATRChk1APC/CCdh1-dependent stabilization of Cdc7ASK (Dbf4) kinase is required for DNA lesion bypass under replication stress". *Genes and Development* 27:2459-2472, 2013.

[10] D.C. Fingar, S. Salama, C. Tsou, E. Harlow, J. Blenis. "Mammalian cell size is controlled by mTOR and its downstream targets S6K1 and 4EBP1/eIF4E". *Genes and Development*, 16:1472-1487, 2002.

IDENTIFYING MODULES IN BIOLOGICAL NETWORK WITH WG-CLUSTER

Paola Lecca[(1)], Angela Re[(1)]

(1) Centre for Integrative Biology - University of Trento
via Sommarive 9, 38123 Povo (Trento), Italy, {lecca, re}@unitn.it

Keywords: biological networks, modules detection, graph clustering, network entropy.

Abstract. Network modular structure is a pressing challenge to gain great power in biological discovery from omics data. Most clustering algorithms seek to achieve this objective through node classification by means of node-related topological or quantitative properties. However, so doing, they disregard the additional richness which is now possible to introduce by edge weights whose information content is augmenting due to fast-evolving technical advances of omics profilings. Here, we present WG-Cluster (Weighted Graph CLUSTERing), a novel technique for network modular structure reconstruction which, compared to other techniques, exploits network edge weights to improve network clustering efficiency and biological interpretability. Additional distinctive features include the assessment of statistical significance for detected modules, and the identification of emerging topological properties in reconstructed modules. WG-Cluster can be applied to direct and indirect networks and scales up to large omics data sets.

1 Scientific Background

Cellular interconnectivity is daunting with 86,771 protein isoforms, 1,000 metabolites and and an incrementally number of newly discovered functional non-coding RNAs. Consequently, a network formalization of cellular processes and subsequent reconstruction of its modular organization is a pressing necessity. Network-based approaches provide insight from the molecular interrogation of many types of interactions, including but not limiting to protein-protein, protein-DNA or protein-RNA interactions.

A rich collection of algorithms has been introduced for module detection in weighted networks, which vary in the choices of weights assignments to nodes and/or edges, final objectives of module detection and consequently in procedural steps [1]. Major limitations of most clustering methods include the high computational cost and the inefficiency in exploiting the knowledge on edge weight, which in fact can encode biologically relevant information.

Here, we present a new algorithm for module reconstruction, implemented by the WG-Cluster (Weighted Graph CLUSTERing) tool, which leverages information on both node and edge weights to efficiently provide statistically significant modules. To ensure this result, WG-Cluster integrates an edge-based unsupervised k-means network clustering upfront a fast-greedy algorithm to identify modules. The initial detection of sub-graphs of similar edge weights lends a gain in efficiency, and computing an entropy score on the modules extracted from each sub-graph allows for a network-based estimate of modules statistical significance. Finally, a convolution analysis of sub-graph mean weight and module entropy returns a general overview of module reconstruction results which is meant to inform downstream analysis. Figures 1 and 2 present the WG-Cluster workflow and GUI screenshots.

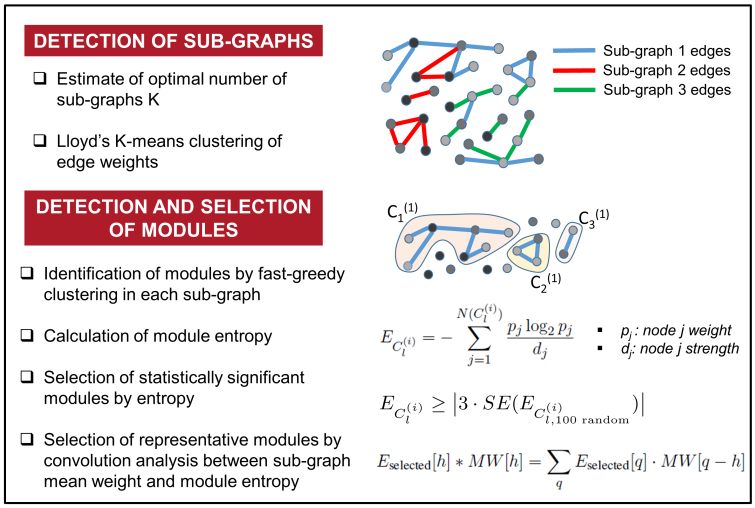


Figure 1: **WG-Cluster workflow.** The complexity of modules for estimating the optimal number of sub-graphs and for running the Lloyd's K-means is linear in the number of edges NE and number of iterations; the complexity for detecting connected components is $\mathcal{O}(V(\log V)^2)$, where V is the number of vertices.

2 Materials and Methods

The WG-Cluster algorithm is implemented in R (http://wwww.r-project.org) and takes as input network edges in Simple Interaction File (SIF Cytoscape) format and node weights in tabular format (node, weight). The algorithm sequentially executes four computational modules. First it estimates the optimal number of clusters (sub-graphs) that split up the graph. Then it executes a Lloyd's K-means clustering ([2]) of the edge weights to detect sub-graphs with edges of similar weights. Fast-greedy modularity optimization procedure finds (if any) the connected components in each sub-graph. Finally, an entropy score is computed for each connected component and it is used as a measure of the statistical significance of the component.

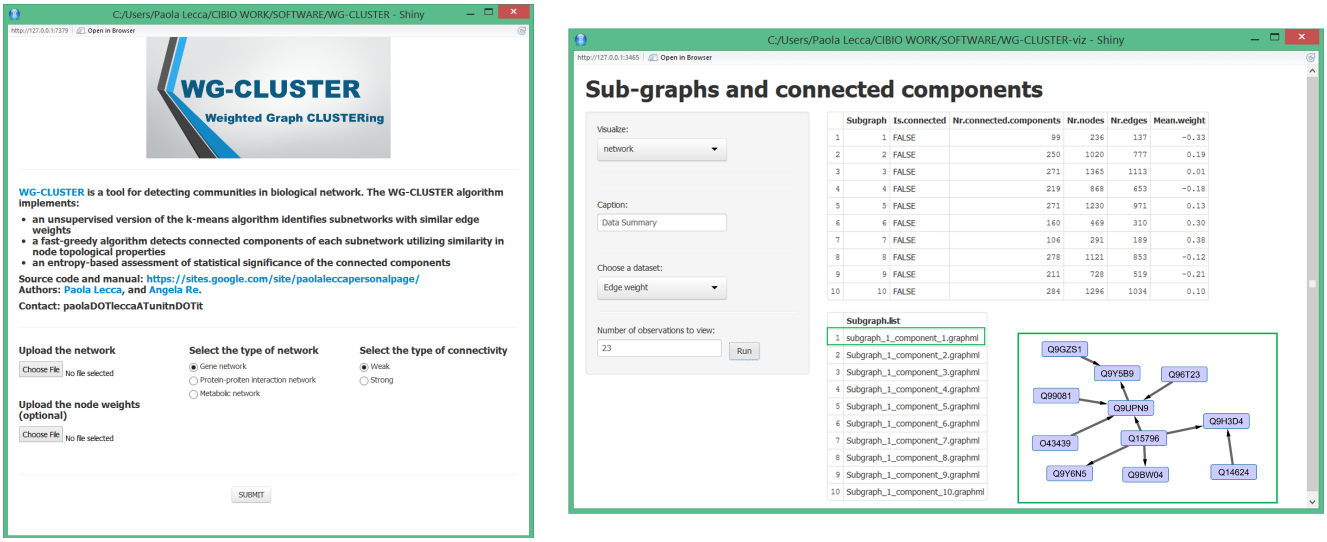


Figure 2: **Front-end and screenshot of WG-Cluster.** On the left, a user-friendly front-end asks the user to (i) upload the input file of network edges in SIF format and optionally the input file of network node weights in a text table; (ii) select the type of biological network to analyze (gene, protein and metabolic networks), and (iii) the type of connected components to detect (weak or strong). On the right, screenshot summarizing counts and properties on the sub-graphs and connected components detected by the algorithm. A GRAPHML file is created for each component to enable visualization and follow-up analysis.

Computational modules of WG-Cluster are now described in detail. Hereafter, we will denote with V the number of vertices and with NE the number of edges in the input graph.

2.1 *Detection of sub-graphs*

The number of optimal sub-graphs which partition the input graph is estimated by minimizing the total within-clusters sum of squares (WCSS) obtained with a K-means

procedure. For a set of edge weights $\mathbf{w} = (w_1, w_2, \ldots, w_E)$, K-means clustering tries to find a set of K cluster centers $S = (S_1, S_2, \ldots, S_K)$ that is a solution to the minimization problem:

$$WCSS = \sum_{i=1}^{K} \sum_{\mathbf{w} \in S_i} ||\mathbf{w} - \mu_i||^2$$

where μ_i is the mean of the edge weights $\mathbf{w}$ in sub-graph S_i.

An elbow in the curve interpolating the points ($n_{\text{sub-graphs}}$, WCSS) suggests the appropriate number of sub-graphs n_{optimal}. n_{optimal} is estimated as the minimum value of n_{clusters} at which the first derivative of WCSS w.r.t. $n_{\text{sub-graphs}}$ is null within a tolerance $0 < \epsilon \ll 1$, i.e. $\left| \frac{d\, WCSS}{dn_{\text{sub-graphs}}} \right| \leq \epsilon$. The first derivative of the curve ($n_{\text{sub-graphs}}$, WCSS) is calculated by the Stineman algorithm [3]. The problem of WCSS minimization is known to be NP-hard. Furthermore, if the input data do not have a strong clustering structure, the procedure may not converge. For this reason, WG-Cluster adopts the Lloyd's algorithm whose complexity is linear in the number of edges and number of sub-graphs, and is recommended in case of data poorly clustered ([2]).

2.2 *Detection and selection of connected components*

Each sub-graph S_i ($i = 1, \ldots, K$) returned by the K-means clustering is decomposed into connected components $C_l^{(i)}$ (with $l = 1, 2, \ldots, L_i$, where L_i is the number of connected components in sub-graph S_i) via a fast-greedy optimization procedure ([4]). The entropy of each connected components is calculated as follows:

$$E_{C_l^{(i)}} = -\sum_{j=1}^{N(C_l^{(i)})} \frac{p_j \log_2 p_j}{d_j} \tag{1}$$

where $N(C_l^{(i)})$ is the number of nodes in the connected component $C_l^{(i)}$, p_j is the fold change of the expression level (from normal to tumor condition) of gene j (normalized between 0 and 1) and d_j is the sum of the weights of the edges adjacent to the node representing gene j (known as *node strength*). Denoting with $D^{(j)}$ the number of nodes directly connected to node j, d_j is thus defined as $d_j = \sum_{h=1}^{D^{(j)}} w_{jh}$, where w_{jh} is the edge weight between the node j and its directly connected node h.

The entropy is used as a measure of significance of the connected components. In order to establish a threshold on the significance, for each connected component $C_l^{(i)}$, an ensemble of 100 random connected components with the same degree distribution of the reference connected component $C_l^{(i)}$ is generated. A connected component is considered significant, and retained, if its entropy value is more than three standard deviations from the mean entropy of the corresponding ensemble of random connected components. Let denote with $\{C_{l'}^{(i')}\}$, where $l' \in \{1, 2, \ldots, L_i'\}$ with $L_i' \leq L_i$, and $i' \in \{1, 2, \ldots, K'\}$ with $K' \leq K$.

Since sub-graph mean weight and connected component entropy are complementarily informative, both of them are employed to classify statistically significant connected components. The convolution of the entropy of selected connected components (E_{selected}) with the mean edge weight MW of the sub-graph to which they belong is performed, as follows:

$$C_{\otimes} = E_{\text{selected}}[h] * MW[h] = \sum_{q} E_{\text{selected}}[q] \cdot MW[q - h] \tag{2}$$

where $E_{\text{selected}} = \{E_{C_{l'}^{(i')}}\}$ is the vector of the entropies of the significant connected components, and $MW = \{(1/NE^{(i')}) \sum_{l=1}^{NE^{(i')}} w_l\}$ is the mean edge weight of the sub-

graph to which they belong. The convolution in Eq. (2) calculates the area overlap (thus similarity) between the probability distributions of the entropy and of the mean edge weight as a function of the amount by which one of the distribution is translated. The frequency spectrum of the convolution offers a flexible way to classify connected components. Indeed, a summarizing view of the network organization can be provided by the connected components that have the most frequent value $C_{\otimes}$.

2.3 *Clustering validation and performances assessment*

To evaluate WG-Cluster solutions, we used standard validation indices, the silhouette index and the maximal R Davies-Bouldin index [7]. Those indices do not rely on benchmarking graphs but compare intra-cluster similarity and inter-cluster similarity based on the rationale that a well performing clustering algorithm should return compact and distinct clusters. Silhouette index close to 1 and maximal R Davies-Bouldin index close to 0 (Figure 3) support the high quality of the clustering. Furthermore, we evaluated WG-Cluster performances to process Erdos-Renyi random graphs of 500 nodes and an increasing number of edges in terms of user CPU and system CPU running times. Compared to state-of-the-art approaches, WG-Cluster resulted efficient by requiring short running time on huge and complex networks (Figure 4).

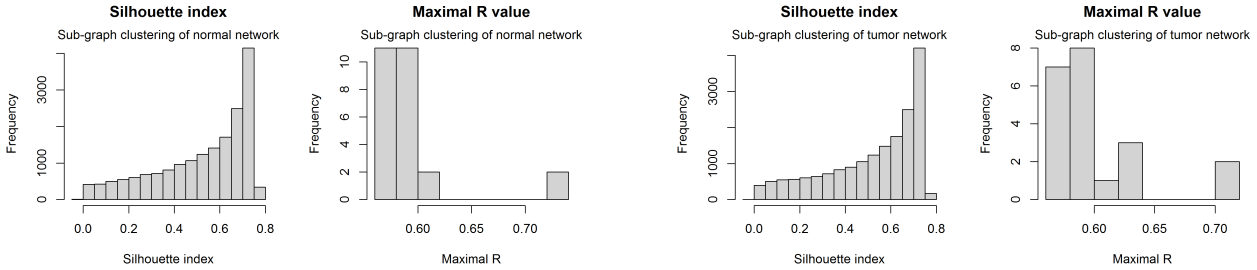
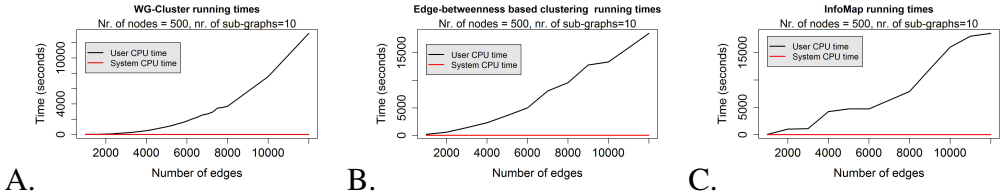


Figure 3: **Clustering validation indices.** Distribution of silhouette and maximal R Davies-Bouldin index [7] of sub-graphs detected by WG-Cluster for normal and tumor network. Silhouette index close to one and Davies-Bouldin index close to zero indicate the good clustering performances of WG-Cluster.



A. B. C.

Figure 4: **Running times to cluster random weighted graphs with increasing number of edges.** WG-Cluster running time on a random weighted graph of 500 nodes and an increasing number of edges in (**A**) is compared with that achieved by the edge betweenness graph clustering algorithm ([5]) in (**B**) and that of InfoMap ([6]) in (**C**). Each algorithm was utilized in its R implementation on a desktop Windows 8.1 PC with a 3.1 GHz CPU.

3 Results

As a proof-of-principle, we applied WG-Cluster to compare the modular structure of a protein-protein interaction (PPI) network in tumour/normal conditions. PPIs were extracted from the open-access IntAct database (http://www.ebi.ac.uk/intact/) whose PPI scoring system fits into community standards. We contextualized the PPI network in cancer and normal conditions by crossing the IntAct network with The Cancer Genome Atlas (http://cancergenome.nih.gov/) colorectal cancer transcriptome dataset, which provides processed $\log_2$ mRNA expression values for 11,589 genes in 19 samples extracted from normal tissues and in 155 samples from tissues affected by cancer. By this operation we constructed two networks (of 18,078 edges) corresponding to the normal and tumour conditions. In each network, node and edge weights were computed as follows. A co-expression score for each gene pair was defined by the Pearsons correlation coefficient of the mRNA levels, either across the normal samples or across the

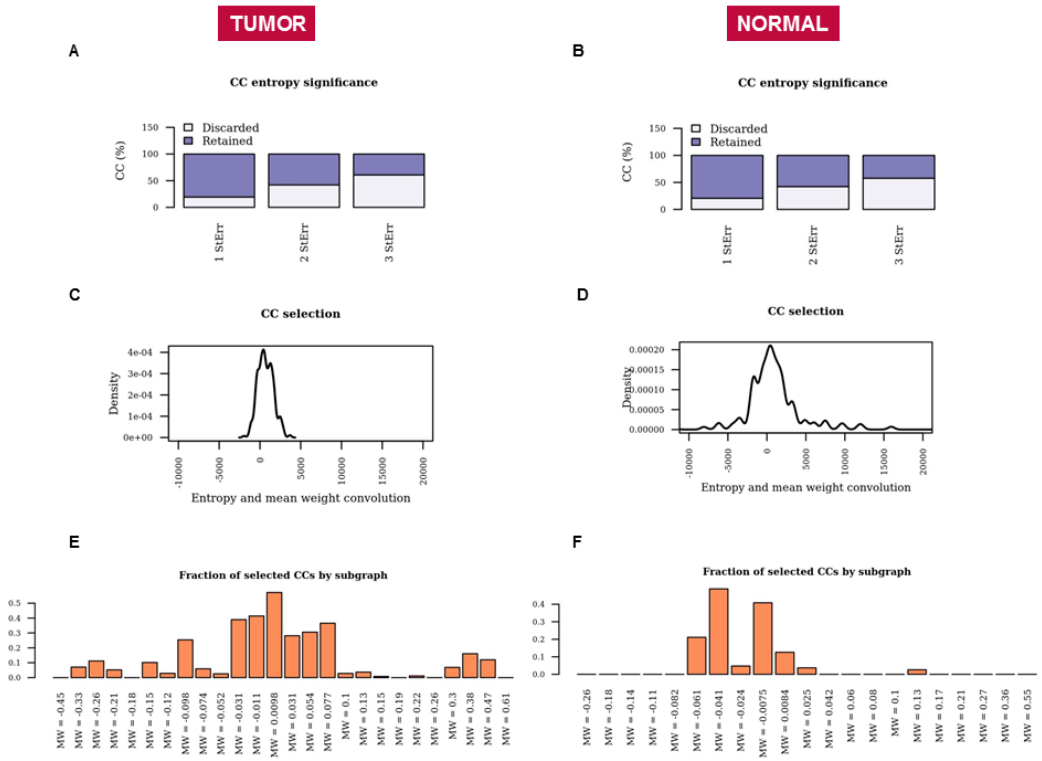


Figure 5: **Summary of connected components reconstructed in the tumor/normal cases.** (A-B) Barplots displaying the fractions of retained components by the number of standard errors from the mean entropy of the ensemble of random components. (C-D) Convolution density plots. (E-F) Barplots displaying, in each sub-graph, the fraction of components associated with convolution values within the central peaks of the convolution density plots. Sub-graphs are labeled by their mean edge weights.

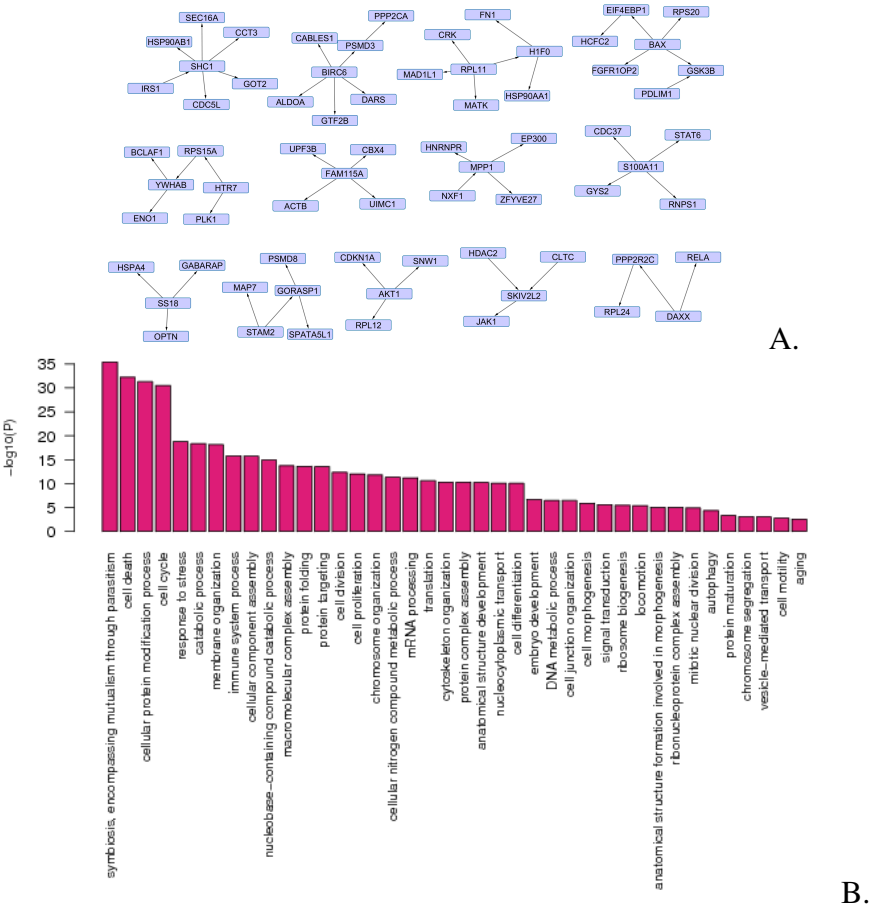


Figure 6: (A) Network of interactions which were extracted from sub-graphs of extreme mean edge weights and which resulted more intense in the tumor versus normal case. (B) Enrichment analysis in Gene Ontology categories. Y-axis reports the negative log10 algorithm of FDR (Fisher's exact test).

tumour samples. The product between the IntAct score of a pair of interacting proteins and the co-expression score of the corresponding mRNAs defines the final weight of an edge in a network. The weight of a node in a network was defined by the average mRNA expression, either across the normal samples or across the tumour samples. WG-cluster was then run in parallel to the weighted networks in each condition. The fraction of selected connected components as a function of increasing stringency in the entropy-based assessment of statistical significance was similar in the two conditions (Figure 5A-B). Interestingly, the clustering solutions in the tumor/normal cases pointed at an overall substantially different network (Figure 5C-D). Indeed, the density plot of the convolution between sub-graph mean weight and connected component entropy was neatly peaked in the tumor case, whereas it was spread over a wider range of values in the normal case. To delve into this observation, highly frequent convolution values, belonging to the central peaks of the density plots, were selected and interpreted in terms of corresponding mean weight and entropy. We observed connected components in sub-graphs yielding extreme mean weights (MW $\leq$ -0.26 or MW $\geq$ 0.13) in the tumor case but not in the normal case. We therefore focused on the tumor-related sub-graphs yielding extreme mean edge weights and extracted edges from corresponding connected components. We then asked whether those 551 interactions belonged to connected components with highly frequent convolution values also in the normal network. Interestingly, 18% of the interactions resulted to get more intense in the tumor condition with respect to the normal one. Functional enrichment analysis of the genes involved in those interactions (Figure 6A) permitted to prioritize relevant Gene Ontology [8] biological processes (Figure 6B).

4 Conclusion

Network analysis is among the most common tools in systems biology. In this study we presented WG-Cluster which leverages information encoded through network node weights and edge weights to efficiently provide statistically significant modules. Among the main motivations for WG-Cluster development there is the pressing need to ground network clustering solutions on both node and edge weights, considering the fast improvements in the annotation not only of the entities of a system but also of their interactions. Furthermore, since technological advances now permit to address the dynamics of a biological system, for instance along time or across conditions, the development of differential network analysis is crucial. Due to its flexibility, WG-Cluster can be applied also in those complex scenarios.

References

[1] J. Wang, M. Li, Y. Deng, Y., Recent advances in clustering methods for protein interaction networks, *BMC Genomics* 1, 11 Spuppl. 3:S10, 2010.

[2] Q. Du, M. Emelianenko, L. Ju, Convergence of the Lloyd algorithm for computing centroidal voronoi tesellation, *SIAM J. NUMER. ANAL.*, 44, 1, 102–119, 2006.

[3] T. Johannesson, H. Bjornsson, Stineman, a consistently well behaved method of interpolation, `http://rpackages.ianhowson.com/cran/stinepack/`, accessed: 2015-01-07, 2012.

[4] A. Clauset, M. E. J. Newman, C. Moore, Finding community structure in very large networks, *Phys. Rev. E*, 70, 066111, 2004.

[5] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, *PNAS*, 99, 12, 78217826, 2001.

[6] M. Rosvall, C. T. Bergstrom, Maps of information flow reveal community structure in complex networks, *PNAS*, 105, 4, 1118–1123, 2008.

[7] M. Halkidi, Y. Batistakis, M,Vazirgiannis, Clustering validity checking methods, *SIGMOD Rec.*, 31, 3, 19–27, 2002.

[8] M. Ashburner, C. A. Ball, J. A. Blake et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet.*, 25, 1, 25–9, 2000.

A DEEP LEARNING APPROACH TO DNA SEQUENCE CLASSIFICATION: FIRST RESULTS

Riccardo Rizzo[(1)], Antonino Fiannaca[(1)], Massimo La Rosa[(1)], Alfonso Urso[(1)]

(1) ICAR-CNR
National Research Council of Italy, viale delle Scienze Ed.11, 90128 Palermo, Italy,
{ricrizzo,fiannaca,larosa,urso}@pa.icar.cnr.it

Keywords: Convolutional Network, Deep Learning, Artificial Neural Network, Spectral Sequence Representation, K-mers representation.

Abstract. Deep learning neural networks are capable to extract significant features from raw data, and to use these features for classification tasks. In this work we present a deep learning neural network for DNA sequence classification based on spectral sequence representation. The framework is tested on a dataset of 3000 16S genes and compared to the GRNN that we tested outperform the Support Vector Machine classification algorithm.

1 Scientific Background

Classification tasks are strongly based on the features that represent the objects to classify. In order to build a good representation it is necessary to recognize and measure meaningful detail of the object, but in some cases it is quite difficult to understand which features use, and this affects the performances of the classification model.

Recently neural deep learning architectures or deep learning models, were proved to be able to extract useful features from input patterns. These architecture are mainly used in image processing and are capable to identify objects on natural images.

The term "deep" refers intuitively to the number of layers that are used in these networks, and, more precisely, is related to the path from an input node to the output node in the network (considering the network as a directed graph) [3].

Among the deep learning architecture it is usually comprised the LeNet-5 network, or convolutional network, a neural network that is inspired by the visual system's structure [1]. This network was used for character recognition in the original paper, and for image processing [2] and speech detection [4].

The main drawback of the deep learning methods is that it is still impossible to reuse the knowledge acquired by the network; deep networks are still black boxes and it is complicated to correct wrong answers or understand the reasons of a good one. As proved in [11] it is possible to build artificial images with no recognizable objects in it, but classified with high confidence in specific categories, as "chair" or "lion", by a deep neural network.

The application of these techniques to gene classification requires a fixed dimension representation of the sequences like the spectral representation based on k–mers occurrences. This representation was used for sequence classification in many works [7, 8]. In particular in our work [8] is noticed that some k–mers are much more important than the other for sequence representation, this means that in the representing vectors there are details that should be taken into account. This observation resembles the problem of feature extraction from image and this idea is at the core of the present work.

In this work we want to understand if the convolutional network is capable to identify and to use these features for sequence classification, outperforming the classifier proposed in the past.

2 Materials and Methods

2.1 *Convolutional Neural Network*

The Convolutional Neural Networks (CNN) are made by a very large number of connections and layers. The one used in this work is a modified version of the LeNet-5 network introduced by LeCun et al. in [1] and is implemented using the python Theano package for deep learning [5, 6].

The LeNet-5 is a network made by two lower layers of convolutional and max-pooling processing elements, followed by two "traditional" fully connected Multi Layer Perceptron (MLP) processing layers, so that there are 6 processing layers.

The convolutional layers calculate L 1-D convolutions between the kernel vectors w^l, of size *2n+1*, and the input signal x:

$$q^l(i) = \sum_{u=-n}^{n} w^l(u)x(i-u) \tag{1}$$

In eq. 1 $q^l(i)$ is the component i of the l-th output vector and $w^l(u)$ is the component u of the l-th kernel vector. After a bias term b^l is added and a non–linear function is applied:

$$h^l(i) = \tanh(q^l(i) + b^l) \tag{2}$$

The vector h^l is the output of the convolutional layer. The max-pooling is a non-linear down-sampling layer. In these processing layers the input vector is partitioned into a set of non-overlapping regions (of 2 elements in this implementation) and, for each sub-region, the maximum value is considered as output. This processing layer reduces the complexity for the higher layers and operates a sort of translational invariance. Convolution and max-pooling are usually considered together and are represented in Fig.1 as two highly connected blocks.

In the proposed architecture the first convolutional layer has $L = 10$ filters of 5 elements ($n = 2$), followed by a max-pooling layer of dimension 2, while the second layer has $L = 20$ filters of the same dimension, and the same max-pooling layer.

The two upper level layers corresponds to a traditional fully-connected MLP: the first layer of the MLP operates on the total number of output from the lower level (the output is flattened to a 1-D vector) and the total number hidden units is 500. The output layer has one unit for each class.

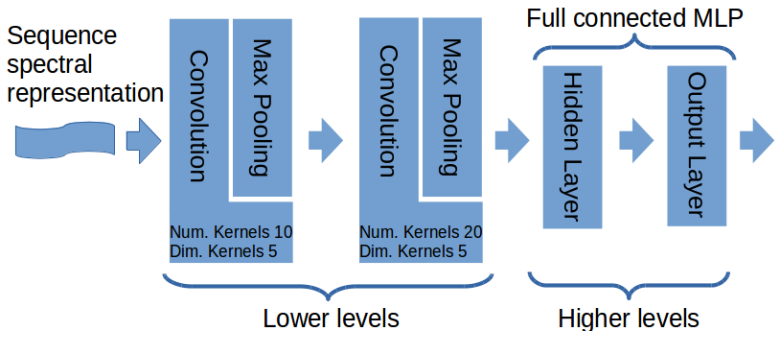


Figure 1: The architecture of the network used.

2.2 *Spectral Representation*

The spectral representation has been successfully used in many bioinformatics works [7, 8], in facts, as demonstrated by [9], each biological species has a proper modal spectrum, that can distinguish it from the others. Given a fixed value k, a spectral representation is a vector of size 4^k. Its components are computed by counting the

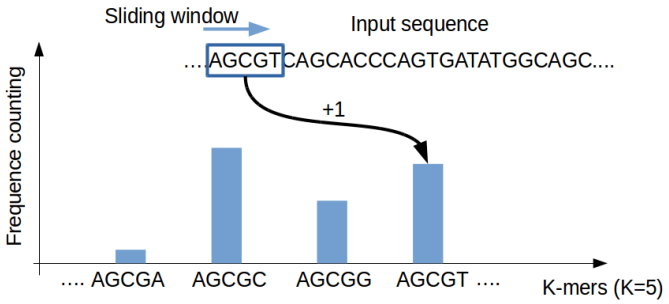


Figure 2: k–mers spectral representation.

occurrences of small DNA snippets of length k, called k–mers, which are extracted from the genomic sequences by means of a sliding window, with step = 1 and length = k. In case of k–mers containing one or more undefined nucleotides, for example the "N" character, they are discarded. Since the CNN is able to discover and exploit those k-mers representing distinctive features, we adopt the so called "bag–of–words" model, which does not take into account the position of kmers in the original sequence. This procedure is summarized in Figure 2. The main computational advantages of using this representation are: (1) to obtain a fixed-size vector representation of genomic sequences and (2) to take into account only distinctive k-mers.

2.3 *Dataset of 16S sequences*

The 16S rRNA sequences have been downloaded from the RDP Ribosomal Database Project II repository [10], release 10.27. We randomly selected 1000 sequences from each of the three most common bacteria phyla, Actinobacteria, Firmicutes, Proteobacteria, collecting in total 3000 sequences. All the sequences have length greater than 1200 bp, are classified as type strain, i.e. they are the best representative of their own species, and are certified as "of good quality" by the RDP database.

3 **Results**

Experimental tests have been carried out using the algorithm and the dataset described in Section 2. Two kinds of experiments have been made. In the first case, using a ten fold cross validation scheme, the prediction performances of the CNN have been tested at each taxonomic rank (from phylum to genus) and considering full length sequences. In the second case, the ten-fold cross validation scheme was repeated considering as test set the sequence fragments of shorter size, 500 bp long, obtained randomly extracting 500 consecutive nucleotides from the original full length sequences. This way, we wanted to asses if the network is able to correctly predict the taxonomic rank of the test sequences even if they also contain only a small part (500 bp) of the original information content. In our experiments, we set the k-mers size to k = 5, as done in other works adopting the spectral representation [7, 8]. The CNN has been run considering two different kernels sizes: kernel_0 = kernel_1 = 5 in the first run; kernel_0 = 25, kernel_1 = 15 in the second run. From here on, the first kernels configuration will be named *kern_1*, whereas the second one will be named *kern_2*. In both configurations the training phase has been run for 200 epochs.

Classification scores, in terms of accuracy, precision and recall, have been compared with another classifier, based on the General Regression Neural Network (GRNN) algorithm, presented in our previous work [8]. The GRNN is a one-pass training neural network, usually adopted for regression purposes, that we adapted for the classification of barcode sequences of animal species, taking into account the COI gene. Moreover we developed three different versions of the GRNN, each one implementing a different distance model: euclidean distance, city-block (Manhattan) distance, Jaccard distance. We compared the CNN classifier with the classification approach based on GRNN algorithm

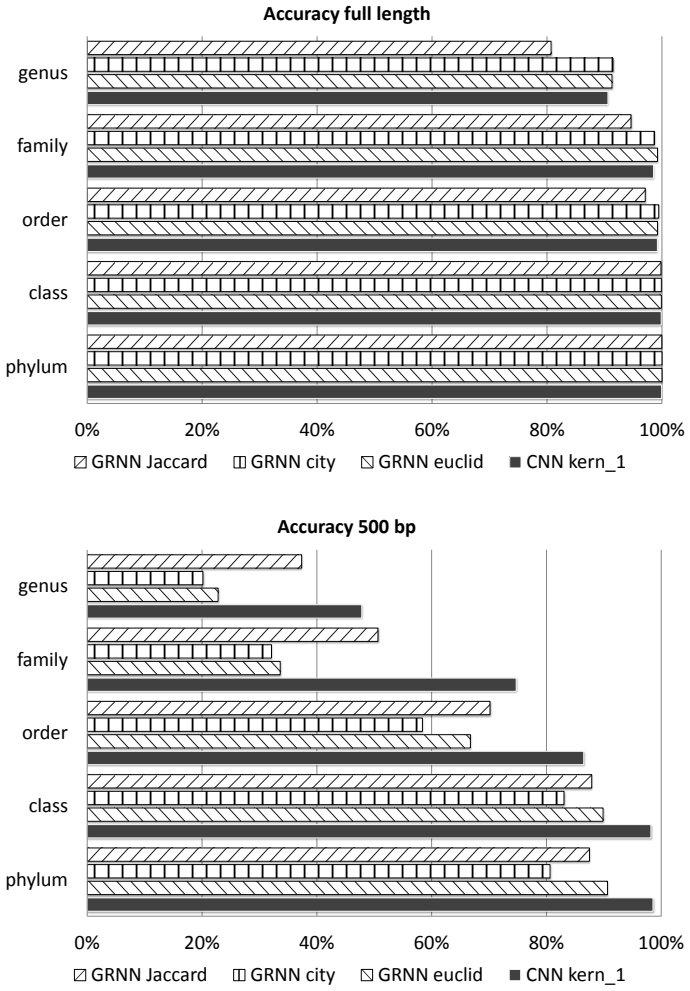


Figure 3: Accuracy scores for full length sequences (upper chart) and 500 bp sequences (lower chart).

because in our previous work [8] we demonstrated that the GRNN outperformed one of the most used classification algorithm, that is the Support Vector Machine (SVM), for the classification of barcode and mini–barcode sequences.

In our experiments, the CNN network with *kern_1* configuration always provided better results with respect to the CNN with kern_2 configuration. For this reason, in the following we will only discuss the results obtained with *kern_1* configuration.

All the classification scores have been summarized in the charts of Figures 3, 4, 5. Considering the full length sequences, it is evident that our approach based on the CNN network, with *kern_1* configuration, reaches almost identical scores, with variance lesser than 1%, with regards with the GRNN classifiers based on the euclidean and the city block distance models. Otherwise the GRNN with Jaccard distance model produced lower results.

Classification scores considering 500 bp sequences showed very interesting results. Our CNN approach, with *kern_1* configuration, clearly outperforms all the other classifiers in terms of accuracy at all taxonomic levels. Only at genus level, accuracy score does not reach the 50%: this behaviour can be explained considering the great number of different genera (393) of the dataset. With regards to the precision chart (Figure 4), the CNN with *kern_1* configuration outperforms the other classifiers at phylum and class level; while at order, family and genus level the GRNN with city block distance model reached better results. This behaviour is, however, balanced if we look at the recall scores (Figure 5). There once again the CNN with *kern_1* configuration always reaches the highest scores, demonstrating that our approach has a better true positive rate, that is the percentage of retrieving correctly classified samples.

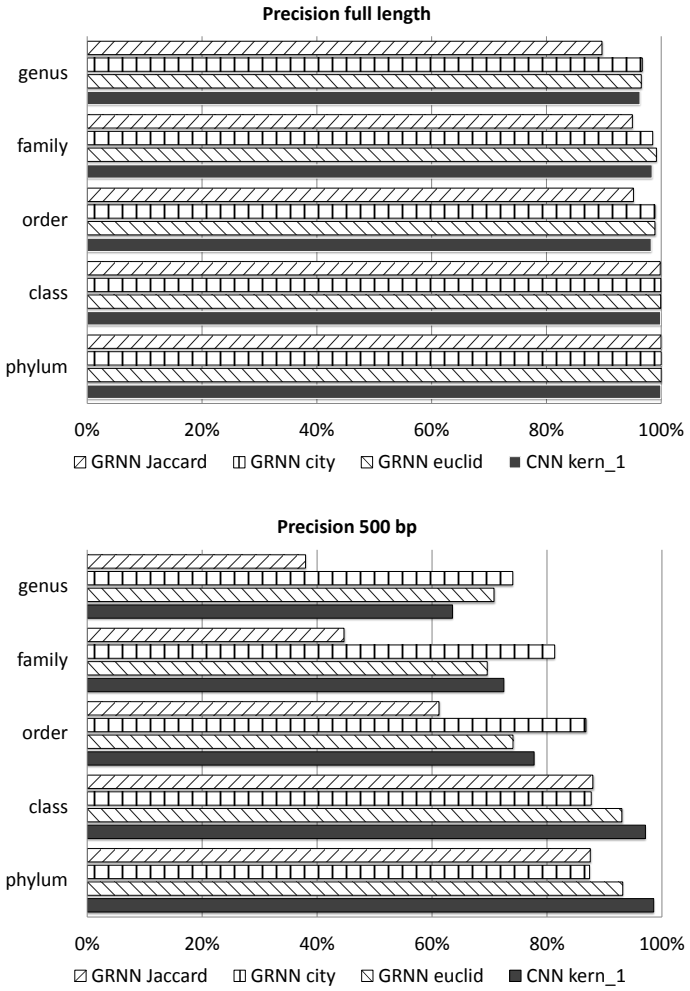


Figure 4: Precision scores for full length sequences (upper chart) and 500 bp sequences (lower chart).

4 Conclusion

These first experiments confirms that the approach is worth of attention and future work. There are at least two things that need much more investigations: the surprisingly not so good recall results with full length sequences, compared with the 500 bp, that carry much less information and more noise, and the precision results.

It is also strange that a network with a larger kernels is not able to give better results, and we plan to investigate also this problem is future works.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* vol.86, n.11, pp. 2278-2324, 1998.

[2] C. Farabet, and C. Couprie, L. Najman,Y. LeCun, "Learning hierarchical features for scene labeling". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, n.8, pp. 1915–1929

[3] Y. Bengio, "Learning deep architectures for AI." *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1-127, 2009.

[4] S. Somsak, A. C. Surendran, J. C. Platt, and C. J.C. Burges. "Convolutional networks for speech detection." *Interspeech*. 2004.

[5] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley and Y. Bengio. "Theano: new features and speed improvements". *NIPS 2012 deep learning workshop*.

[6] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley and Y. Bengio. "Theano: A CPU and GPU Math Expression Compiler". *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 30 - July 3, Austin, TX, 2010.

[7] P. Kuksa, V. Pavlovic. "Efficient alignment-free DNA barcode analytics". *BMC Bioinformatics*, vol.10, Suppl.14, pp.S9, 2009.

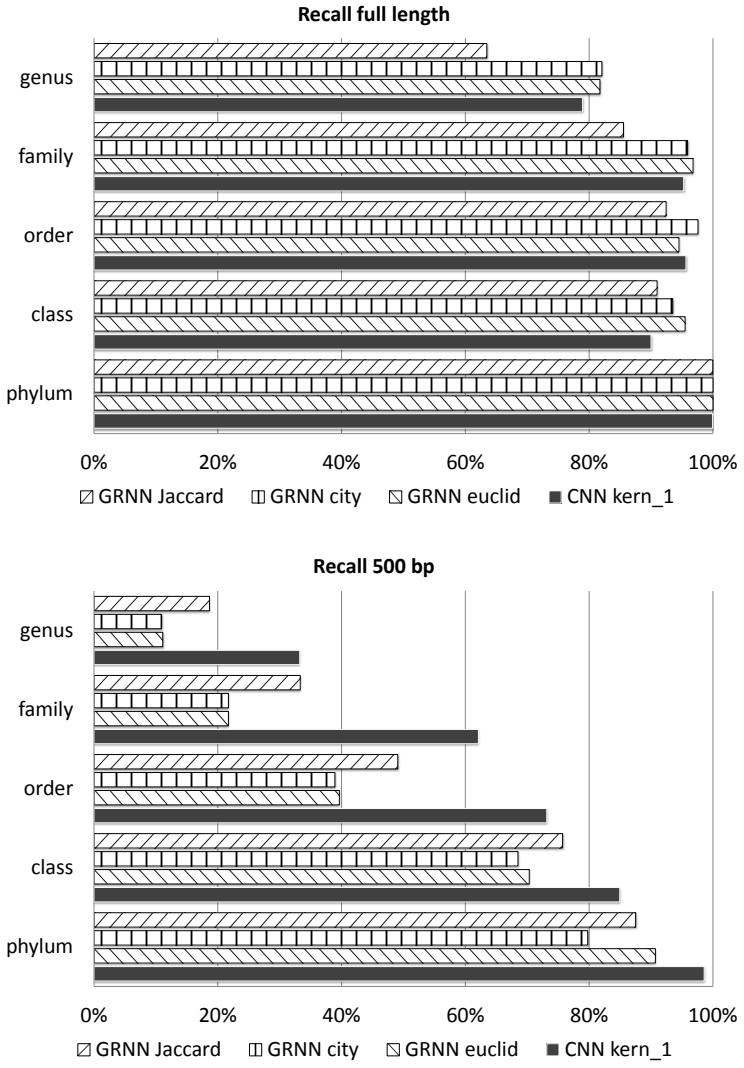


Figure 5: Recall scores for full length sequences (upper chart) and 500 bp sequences (lower chart).

[8] R. Rizzo, A. Fiannaca, M. La Rosa, A.Urso. "The General Regression Neural Network to Classify Barcode and mini-barcode DNA". *Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB)*. Lecture Notes in Computer Science, in press.

[9] B. Chor, D. Horn, N. Goldman, Y. Levy, T. Massingham. "Genomic DNA k-mer spectra: models and modalities". *Genome Biology*, vol.10:R108, 2009.

[10] J.R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R.J. Farris, A.S. Kulam-Syed-Mohideen, D.M., McGarrell, T. Marsh, G.M. Garrity, J.M. Tiedje. "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis". *Nucleic acids research*, vol.37(Database Issue), pp.D141–145

[11] A. Nguyen, J. Yosinski, J. Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", *Computer Vision and Pattern Recognition (CVPR'15), 2015*

A walking tour in Reproducible Research and Big Data Management with RNASeqGUI and R

Francesco Russo, Dario Righelli, Claudia Angelini

National Research Council (CNR)
Institute for applied mathematics "Mauro Picone" (IAC)
Via Pietro Castellino 111, 80131, Naples, Italy
email: f.russo@na.iac.cnr.it
website: http://bioinfo.na.iac.cnr.it/BioinfoLab/

Keywords: Reproducible Research, Big-Data, GUI, RNA-Seq, Differentially Expressed Genes.

Abstract In this paper, we discuss the concept of Reproducible Research and its importance to produce transparent and high quality scientific papers. In particular, we illustrate the advantages that both paper authors and readers can receive from the adoption of Reproducible Research and we discuss a strategy to develop computational tools supporting such a feature. We present a novel version of RNASeqGUI, a user friendly computational tool capable to handle and analyse RNA-Seq data. This tool exploits Reproducible Research feature to produce RNA-Seq analyses easy to read, inspect, understand, study, reproduce and modify. Overall, this paper is a proof of concept on how it is possible to develop complex and interactive tools in the spirit of Reproducible Research.

1 Introduction

In order to asses and verify the results of scientific publications, it is fundamental to inspect and reproduce the entire analysis carried out by the authors. Unfortunately, the way by which the analysis is described does not facilitate the reproducibility of results. In the recent years, the issue of reproducibility [1, 2, 3, 4, 5, 6] has been emerging as a crucial aspect also in Medicine and Genetics. The implementation of this feature requires a lot of attention and the utmost effort from the researchers that have to keep track of all steps performed. However, if a paper is published in reproducible spirit, a reader, starting from the same data and following all the steps described as a "computational protocol", will be able to obtain the same results presented in that paper. Therefore, Reproducible Research is a fundamental feature that gives the possibility to have deep insight into an analysis, to verify the authenticity of results (eventually to find out bugs and mistakes) and to improve knowledge transfer. Overall, it contributes to publish high quality Science work.

Nevertheless, the analyses of Next Generation Sequencing data are usually quite complex and require the use of several different tools and pipelines. Moreover, several preliminary steps (e.g. filtering, normalization, rounding) are often performed before processing the data to obtain meaningful results. Usually, those types of analyses are long and time-consuming and often require to carefully choose the appropriate methods and parameters to handle a specific case. Therefore, it is very difficult to keep track of all these details without using automatic procedures that exploit all the advantages of literate statistical programming and version control. Moreover, Next Generation Sequencing analyses are computationally expensive and have to handle very large data sets (e.g. files of sequences or alignments that are usually many gigabytes large). Thus, the implementation of tools and pipelines have to take into account both reproducibility and time/memory consuming issues.

In the recent years, many different tools (e.g. filehash, ReportingTools, cacher) have been built to help developers to incorporate the Reproducible Research feature inside their software for different programming languages, including R. However, the use of such instruments as a routinely way for the implementation of software dedicated to NGS data analysis is still vary little unexplored.
For this reason, we present a novel version of our software, called RNASeqGUI [8], that combines Reproducible Research and management of big data sets. RNASeqGUI is a open-source software dedicated to the analysis of RNA-Seq data.
The rest of the paper is organized as follows: In Section 2, we discuss how the Reproducible Research feature has been incorporated in RNASeqGUI. In Section 3, we discuss about the general features of this software, its structure and functionalities. In Section 4, we discuss the conclusions.

2 Methods

Graphical user interfaces are usually very easy to use and do not require a specific knowledge of a programming language. By means of "point and click" approach, a non-expert user can run complex and personalised analyses on the dataset of interest. However, the price to be paid for this flexibility is the difficulty of keeping track of all actions performed while using these kind of tools.
In order to solve this issue, we incorporated the Reproducible Research feature inside RNASeqGUI. As a result, all actions and steps are automatically recorded by our software and can be automatically re-executed when needed.
The strategy we adopted is the following one. During the usage of RNASeqGUI, all lines of code that RNASeqGUI executes after clicking on a particular button are written in a R markdown file that is automatically saved. The file can be compiled, re-executed, and showed to the user in the form of an HTML report whenever he wants. Therefore, not only does RNASeqGUI provide an open source code, but also all those lines, that have been actually executed during an analysis process, are clearly reported in a self-contained way, called *chunks of code*. These constitute complete and independent units of code that can be run independently in an R console without the need to install RNASeqGUI. A reader just needs to install R and Bioconductor (not RNASeqGUI) and by copying & pasting the code chunks of interest contained in the report he can reproduced the same analysis carried out by a RNASeqGUI user.
In this way, some reader does not need to read the entire code of some execution but just the specific chunks of code that have been used during a particular step of the analysis. Therefore, the report contains all the information about the lines of code that have been used by RNASeqGUI plus all the initialization parameters and the input and output data. Such a report can be considered as a full detailed log file, written in human readable and friendly format usable as a supplementary material, containing executable code along with all initializations and printed results (plots, tables, arrays etc.). Moreover, each chunk of code inside the report can be run independently to obtain the results shown in that report.
Using this simple implementation, each time a the report is generated all chunks of code are re-executed. However, full re-execution might be very time consuming both for the authors and for readers since the amount of data involved in a RNA-Seq study can be very large. Moreover, a potential reader might not have the resources to run all the analyses. This problem is solve by a method, called *caching* [7]. Caching is a strategy that allows to store data into several objects in order to retrieve these data in a faster and secure way. This prevents the re-computation of time consuming chunks of code during the generation of the report by saving some intermediate results into several objects that are called during the re-creation of the report file. The user does not need to perform any step to store the code chunks and intermediate results. Both data storage and chunk

generation are automatically executed.
Therefore, caching makes available all the intermediate results that can be checked separately and can be used as starting points for different analyses. As a consequence, the implementation of caching - that works automatically inside RNASeqGUI - allows the user to run in a more efficient way different types of analyses on the same dataset and to easily modify an analysis while still preserving reproducibility. By combining Reproducible Research tools and caching in RNASeqGUI, we allow a better management of the entire data analysis and an automatic way to keep track of the computational protocol used for analysing a specific dataset.

3 RNASeqGUI

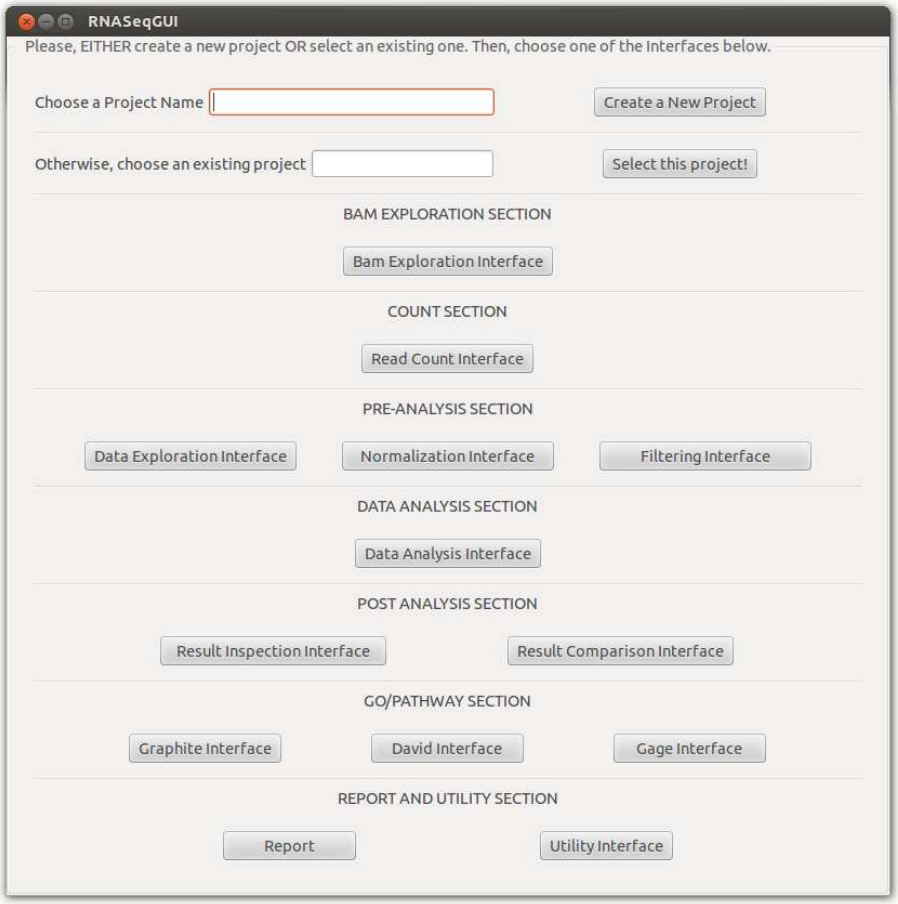


Figure 1: Sections of RNASeqGUI main interface. A user must create a project by choosing a project name and by clicking on `Create a New Project` button.

RNASeqGUI is implemented in R and requires the RGTK2 graphical library [9] to run, it is completely open source and freely downloadable from
`http://bioinfo.na.iac.cnr.it/RNASeqGUI/`.
This GUI is built to identify differentially expressed genes from RNA-Seq experiments and to support the interpretation of the results. It includes several well known RNA-Seq tools, available as command line in Bioconductor.
This software, thanks to the R markdown language, is capable to automatically generate a dynamic report describing all the analysis carried out on a given project in a fully detailed way. The report includes all R code chunks used during the analysis, the figures and the summary of the results. These chunks can be executed and results are updated automatically whether some changes occur. It also includes all versions of the R packages used (session info), all steps, input/output parameters, file names, etc., and it can be exported as HTML file.

The current release of RNASeqGUI [8] is divided into seven main sections (see Figure 1). Each section is dedicated to a particular step of the data analysis process. Since the first version of the software, we have deeply increased both the number of sections and the number (and variety) of functionalities within each section. In the current release, the first section covers the exploration of the bam files. The second concerns the counting process of the mapped reads against a genes annotation file. The third focuses on the exploration of count-data, on the normalization procedures and on the filtering process. The fourth is about the identification of the differentially expressed genes that can be performed by several methods. Such a section now includes also the possibility to handle more complex designs. Moreover, all results of the methods for identifying differentially expressed genes can be explored via web thanks to `ReportingTools` R package [10]. The fifth section regards the inspection of the results produced by these methods and the quantitative comparison among them. The novel six section regards the Gene-Ontology, Gene-Set enrichment and Pathway Analysis. The seventh section focuses on report generation that contains all the chunks of codes that have been executed along with the plots generated during the analysis, and contains the Utility Interface that provides a series of general purpose functions.

4 Conclusions

Reproducible Research is receiving a lot of attention in the recent years. There are several approaches to Reproducible Research. One of these [11] can be achieved through a virtual machine (VM), by giving to a third part user the possibility to download and install it on his own computer in order for him to replicate the incorporated analysis. This method is useful to replicate an entire experiment. Unfortunately, if a user wants to reproduce more than one analysis, probably he needs to configure more than one VM to reproduce all the analyses. On the other hand, RNASeqGUI is interactive. Therefore, it does not have of the problem of setting up a VM machine that can execute one pipeline per time.

In RNASeqGUI, we choose to combine the peculiar features of a graphical user interface with the tools available in *Bioconductor* for Reproducible Research. Therefore, for each analysis flux (project), RNASeqGUI automatically generates a report that keeps track of all actions executed by the user, plus provides a set of cached objects saved in a database that stores some intermediate results of the analysis.

In this manner, the results (figures, tables, etc.) can be used in a publication and the report can be attached as supporting information data. Moreover, database of cached objects can be shared via the web to allow the possibility to perform the same analysis using the same data.

To conclude, thanks to this report and to the available cached objects database, not only does the user promote the transparency of his own work, but also he allows other readers to execute alternate analysis starting from intermediate results of the original analysis carried out.

Finally, this paper is a proof of concept an how to build computational tools able to handle complex and high dimensional data analysis in the spirit of Reproducible Research.

Acknowledgements

We want to thank M. Franzese, V. Costa and R. Esposito for suggestions and discussions, D. Granata for technical support.

This work was supported by the Italian Flagship **InterOmics** Project (PB.P05) and by BMBS **COST Action** BM1006.

References

[1] Gentleman R (2004) Reproducible Research: A Bioinformatics Case Study. Bioconductor Project Working Papers. Working Paper 3.

[2] Peng R D (2011) Reproducible Research in Computational Science. Science, 334(6060), 1226-1227.

[3] Peng R D (2009) Reproducible research and Biostatistics. Biostatistics, 10 (3): 405-408.

[4] Nekrutenko A and Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nature Reviews Genetics, 13 (9):667-672.

[5] Ince D C, Hatton L, Graham-Cumming J (2012) The case for open computer programs. Nature Perspective, (482), 485-488.

[6] Editorial (2013) Enhancing reproducibility. Nature Methods, (10), 367.

[7] Peng R D (2008) Caching and Distributing Statistical Analyses in R. Journal of Statistical Software, 26 7.

[8] Russo F and Angelini C (2014.) RNASeqGUI: A GUI for analysing RNA-seq data. Bioinformatics. Bioinformatics, 30 (17): 2514-2516.

[9] Lawrence M and Duncan TL (2010) RGtk2: A Graphical User Interface Toolkit for R. Journal of Statistical Software, 37(8).

[10] Huntley M A, Larson J L, Chaivorapol C, Becker G, Lawrence M, Hackney J A, Kaminker J S (2013) ReportingTools: an automated result processing and presentation toolkit for high throughput genomic analyses, Bioinformatics. 29 (24):3220.

[11] Hillman-Jackson J, Clements D, Blankenberg D, Taylor J, Nekrutenko A; Galaxy Team. (2012) Using Galaxy to perform large-scale interactive data analyses, Curr Protoc Bioinformatics. Chapter 10:Unit10.5.

CIBB 2015 Short contributed talks

BIOSTATISTICS TECHNICAL CHAIR

Paola MV Rancoita, University Vita-Salute San Raffaele, Italy

BIOINFORMATICS TECHNICAL CHAIR

Stefano Rovetta, University of Genova, Italy

Well-supported phylogenies using largest subsets of core-genes by discrete particle swarm optimization

Reem Alsrraj[1,2], Bassam AlKindy[1,3], Christophe Guyeux[1], Laurent Philippe[1], and Jean-François Couchot[1]
[1] FEMTO-ST Institute, UMR 6174 CNRS, University of Franche-Comté, France
[2] Iraqi Commission for Computers and Informatics, Baghdad, Iraq
[3] Department of Computer Science, University of Mustansiriyah, Baghdad, Iraq
email:christophe.guyeux@univ-fcomte.fr

Keywords: biomolecular phylogeny, particle swarm optimization, rosale order.

Abstract. The number of complete chloroplastic genomes increases day after day, making it possible to rethink plants phylogeny at the biomolecular era. Given a set of close plants sharing in the order of one hundred of core chloroplastic genes, this article focuses on how to extract the largest subset of sequences in order to obtain the most supported species tree. Due to computational complexity, a discrete and distributed Particle Swarm Optimization (DPSO) is proposed. It is finally applied to the core genes of *Rosales* order.

1 Introduction

Given a set of biomolecular sequences or characters, various well-established approaches have been developed in recent years to deduce their phylogenetic relationship, encompassing distance-based matrices, Bayesian inference, or maximum likelihood [1]. Robustness aspects of the produced trees can be evaluated too, for instance through bootstrap analyses. However the relationship between this robustness, the real accuracy of the phylogenetic tree, and the amount of data used for this reconstruction is not yet completely understood. More precisely, if we consider a set of species reduced to lists of gene sequences, we have an obvious dependence between the chosen subset of sequences and the obtained tree (its topology or robustness). This dependence is usually regarded by the mean of gene trees merged into a phylogenetic network.

This article investigates the converse approach: it starts by the union of whole core genes, and tries to remove the ones that blur the phylogenetic signals. More precisely, the objective is to find the largest part of the genomes that produces a phylogenetic tree as supported as possible, reflecting by doing so the relationship of the largest part of the sequences under consideration. Due to overwhelming number of combinations to investigate, a brute force approach is a nonsense, which explains why heuristics have been considered. The proposal of this research work is thus the application of a Discrete Particle Swarm Optimization (DPSO) that aims at finding the largest subset of core genes producing a phylogenetic tree as supported as possible. A new algorithm has been proposed and applied, in a distributive manner, to investigate the phylogeny of *Rosales* order.

The remainder of this article is constituted as follows. The DPSO metaheuristic is recalled in the next section. The way to apply it for resolving problematic supports in biomolecular based phylogenies is detailed in Section 3. The proposed methodology is then applied to the particular case of *Rosales* in Section 4. This article ends by a conclusion section, in which the article is summarized and intended future work is outlined.

2 Discrete Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a stochastic optimization technique developed by Eberhart and Kennedy in 1995 [2]. PSOs have been successfully applied in function optimization, artificial neural network training, and fuzzy system control. In this metaheuristic, particles follow a very simple behavior that is to learn from the success of neighboring individuals. An emergent behavior enables individual swarm members to take benefit from the discoveries or from previous experiences of the other members that have obtained more accurate solutions. In the case of the standard binary PSO model [3], the particle position is a vector of N parameters that can be set as "yes" or "no", "true" or "false", "include" or "not include", *etc.* (binary values). A function associates to such kind of vector a real number score according to the optimization problem. The objective is then to define a way to move the particles in the N dimensional binary search space so that they produce the optimal binary vector w.r.t. the scoring function.

In details, each particle i is thus represented by a binary vector X_i (its position). Its length N corresponds to the dimension of the search space, that is, the number of binary parameters to investigate. An 1 in coordinate j in this vector means that the associated j-th parameter is selected. A swarm of n particles is then a list of n vectors of positions $(X_1, X_2, \ldots, X_n)$ together with their associated velocities $V = (V_1, V_2, ..., V_n)$, which are N-dimensional vectors of real numbers between 0 and 1. These latter are initially set randomly. At each iteration, the new velocity is computed as follows:

$$V_i(t+1) = w.V_i(t) + \phi_1(P_i^{best} - X_i) + \phi_2(P_g^{best} - X_i) \tag{1}$$

where w, ϕ_1, and ϕ_2 are weighted parameters setting the level of each 3 trends for the particle, which are respectively to continue in its adventurous direction, to move in the direction of its own best position P_i^{best}, or to follow the gregarious instinct to the global best known solution P_g^{best}. Both P_i^{best} and P_g^{best} are computed according to the scoring function.

The new position of the particle is then obtained using the equation below:

$$X_{ij}(t+1) = \begin{cases} 1, & \text{if } r_{ij} \leq Sig(V_{ij}(t+1)), \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where r_{ij} is a chosen threshold that depends on both the particle i and the parameter j, while the Sig function operating as selection criterion is the sigmoid one [3], that is:

$$Sig(V_{ij}(t)) = \frac{1}{1 + e^{-V_{ij}(t)}}. \tag{3}$$

3 PSO applied to phylogeny

Let us consider, for illustration purpose, a set of chloroplast genomes of *Rosales*, which has already been analyzed in [4] using an hybrid genetic algorithm and Lasso test approach. We sampled 9 ingroup species and 1 outgroup (*Mollissima*), see Table 1 for details, which have been annotated using DOGMA [5]. We can then compute the core genome (genes present everywhere), whose size is equal to 82 genes, by using for instance the method described in [6, 7]. After having aligned them using MUSCLE, we can infer a phylogenetic tree with RAxML [1] (for a general presentation on phylogenetic tree construction see, *e.g.*, [8]). If all bootstrap values are larger than 95, then we can reasonably consider that the *Rosales* phylogeny is resolved, as the largest possible number of genes has led to a very well supported tree.

Species	Accession	Seq.length	Family	Genus
Chiloensis	NC_019601	155603 bp	*Rosaceae*	*Fragaria*
Bracteata	NC_018766	129788 bp	*Rosaceae*	*Fragaria*
Vesca	NC_015206	155691 bp	*Rosaceae*	*Fragaria*
Virginiana	NC_019602	155621 bp	*Rosaceae*	*Fragaria*
Kansuensis	NC_023956	157736 bp	*Rosaceae*	*Prunus*
Persica	NC_014697	157790 bp	*Rosaceae*	*Prunus*
Pyrifolia	NC_015996	159922 bp	*Rosaceae*	*Pyrus*
Rupicola	NC_016921	156612 bp	*Rosaceae*	*Pentactina*
Indica	NC_008359	158484 bp	*Moraceae*	*Morus*
Mollissima	NC_014674	160799 bp	*Fagaceae*	*Castanea*

Table 1: **Genomes information of *Rosales* species under consideration**

In case where some branches are not well supported, we can wonder whether a few genes can be incriminated in this lack of support, for a large variety of reasons encompassing homoplasy, stochastic errors, undetected paralogy, incomplete lineage sorting, horizontal gene transfers, or even hybridization. If so, we face an optimization problem: *to find the most supported tree using the largest subset of core genes*. Obviously, a brute force approach investigating all possible combinations of genes is intractable (2^N phylogenetic trees for N core genes, with $N = 82$ for *Rosales*).

More precisely, genes of the core genome are supposed to be lexicographically ordered. Each subset s of the core genome is thus associated with a unique binary word w of length n: for each i, $1 \leq i \leq n$, w_i is 1 if the i-th core gene is in s and 0 otherwise. Any n-length binary word w can be associated with its percentage p of 1's and the lowest bootstrap b of the phylogenetic tree we obtain when considering the subset of genes associated to w. Each word w is thus associated with a fitness score value $b + p$.

Let us be back in the PSO context. The search space is then $\{0, 1\}^N$. Each node of this N-cube is associated with the set of following data: its subset of core genes, the deduced phylogenetic tree, its lowest bootstrap b and the percentage p of considered core genes, and, finally, the score $b + p$. Notice that two close nodes of the N-cube have two close percentages of core genes. We thus have to construct two phylogenies based on close sequences, leading to a high probability to the same topology with close bootstrap. In other words, the score remains essentially unchanged when moving from a node to one of its neighbors. It allows to find optimal solutions using approaches like PSO.

Initially, the L (set to 10 in our experiments) particles are randomly distributed among all the vertices (binary words) of the N-cube that have a large percentage of 1. The objective is then to move these particles in the cube, hoping that they will converge to an optimal node. At each iteration, the particle velocity is updated according to the fitness and its best position. It is influenced by constant weight factors according to Equation (1). In this one, we have set $c_1 = 1$, $c_2 = 1$, while r_1, r_2 are random numbers between (0.1,0.5), and w is the inertia weight. This latter determines the contribution rate of a particle's previous velocity to its velocity at the current time step. To increase the number of included components in a particle, we reduced the interval of Equation (2) to [0.1, 0.5]. For instance, if the velocity Vi of an element is equal to 0.511545 and $r = 0.83$, then $Sig(0.51) = 0.62$. So $r > Sig(Vi)$ and this will lead to 0 in the vector elements of the particle. By minimizing the interval we increase the probability of having $r < Sig(Vi)$, and this will lead to more 1s, which means more included elements in the particle. A large inertia weight facilitates a global search while a small inertia weight tends more to a local

Algorithm 1: **PSO algorithm**

```
population ← 10, maxiter ← 10
for each particle in population do
  particle[position] ← [randint(0, 1) for each gene in core genome]
  particle[velocity] ← [rand(0, 1) for each gene in core genome]
  particle[score] ← 0
  particle[best] ← Empty list
end for
while fitness < b + p and iter < Maxiter do
  for each particle in population do
    Calculate new_fitness
    if new_fitness > fitness then
      particle[score] ← new_fitness
      particle[best] ← particle[position]
    end if
  end for
  fitness ← max(particle[score])
  Gbest ← position[Max(Particle[score]in population)]
  for each particle in population do
    Calculate particle velocity according to Equation (1)
    Update particle position according to Equations (3) and (2)
  end for
end while
```

investigation [9]. A larger value of w facilitates a complete exploration, whereas small values promote exploitation of areas. This is why Eberhart and Shi suggested to decrease w over time, typically from 0.9 to 0.4, thereby gradually changing from exploration to exploitation. Finally, each particle position is updated according to Equation (2), see Algorithm 1 for further details. In this algorithm, the particle is defined by its position (a binary word) in the cube together with its velocity (a real vector).

4 Experimental results and discussions

We have implemented the proposed DPSO algorithm on the *Mésocentre de calculs* supercomputer facilities of the University of Franche-Comté. Investigated *Rosales* species are listed in Table 1. 10 swarms having a variable number of particles have been launched 10 times, with $c_1 = 1, c_2 = 1$, and w linearly decreasing from 0.9 to 0.4. Obtained results are summarized in Table 2 that contains, for each 10 runs of each 10 swarms: the number of removed genes and the minimum bootstrap of the best tree. Remark that some bootstraps are not so far from the intended ones (larger than 95), whereas the number of removed genes are in average larger than what we desired.

7 topologies have been obtained after either convergence or $maxIter$ iterations. Only 3 of them have occurred a representative number of time, namely the Topologies 0, 2, and 4, which are depicted in Figure 2 (see details in Table 3). These three topologies are almost well supported, except in a few branches. We can notice that the differences in these topologies are based on the sister relationship of two species named *Fragaria vesca* and *Fragaria bracteata*, and of the relation between *Pentactina rupicola* and *Pyrus pyrifolia*. Due to its larger score and number of occurrences, we tend to select Topology 0 as the best representative of the *Rosale* phylogeny.

To further validate this choice, consel [10] software has been used on per site

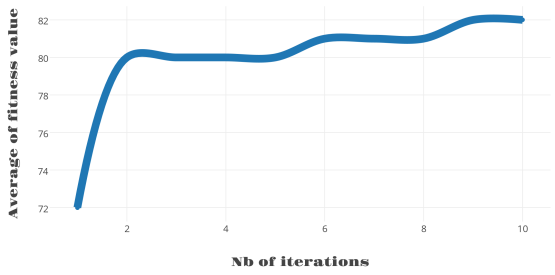


Figure 1: **Average fitness of *Rosales* order**

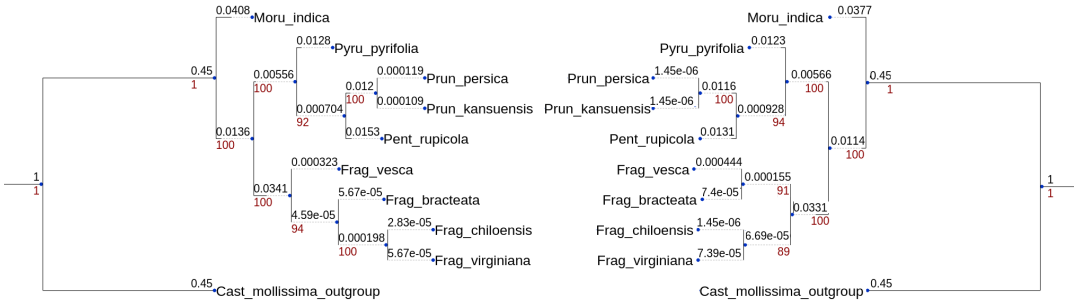
Table 2: Best tree in each swarm

Swarm	Removed genes	$(p+b)/2$	b
1	4	75.5	73
2	6	75.5	76
3	20	75	88
4	52	59.5	89
5	3	75.5	72
6	19	77.5	92
7	47	63.5	92
8	9	73.5	74
9	10	72.5	73
10	13	76.5	84

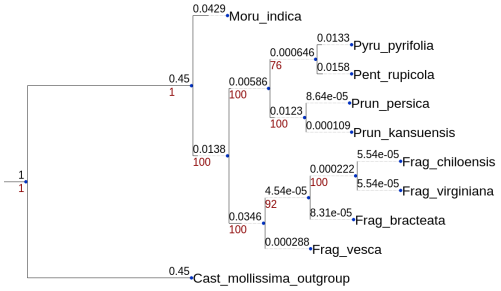
Topology	Swarms	b	p	Occurrences
0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	92	63	568
1	1, 2, 3, 4, 5, 6, 10	63	45	11
2	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	76	67	55
3	8, 1, 2, 3, 4	56	41	5
4	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	89	30	65
5	1, 3, 4, 5, 6, 9	71	33	9
6	5, 6	25	45	2

Table 3: **Best topologies obtained from the generated trees.** b is the lowest bootstrap of the best tree having this topology, while p is the number of considered genes to obtain this tree.

likelihoods of each best tree obtained using RAxML [1]. Consel ranks the trees after having computed the p-values of various well-known statistical tests, like the so-called approximately unbiased (au), Kishino-Hasegawa (kh), Shimodaira-Hasegawa (sh), and Weighted Shimodaira-Hasegawa (wsh) tests. Obtained results are provided in Table 4, they confirm the selection of Topology 0 as the tree reflecting the best the *Rosales* phylogeny.



(a) *Topology*$_0$ (b) *Topology*$_4$



(c) *Topology*$_2$

Figure 2: **The best obtained topologies for *Rosales* order**

Rank	item	obs	au	np	bp	pp	kh	sh	wkh	wsh
1	0	-1.4	0.774	0.436	0.433	0.768	0.728	0.89	0.672	0.907
2	4	1.4	0.267	0.255	0.249	0.194	0.272	0.525	0.272	0.439
3	2	3	0.364	0.312	0.317	0.037	0.328	0.389	0.328	0.383

Table 4: **Consel results regarding best trees**

5 Conclusion

A discrete particle swarm optimization algorithm has been proposed in this article, which focuses on the problem to extract the largest subset of core sequences with a view to obtain the most supported phylogenetic tree. This heuristic approach has then been applied to the 82 core genes of the *Rosales* order.

References

[1] Alexandros Stamatakis. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2014.

[2] James Kennedy and R.C. Eberhart. Particle swarm optimization. In *Proceedings of IEEE International Conference on Neural Networks*, volume 4, pages 1942–1948, 1995.

[3] Mojtaba Ahmadieh Khanesar, Hassan Tavakoli, Mohammad Teshnehlab, and Mahdi Aliyari Shoorehdeli. Novel binary particle swarm optimization. *www.intechopen.com*, (978-953-7619-48-0):11, 2009.

[4] Bassam AlKindy, Christophe Guyeux, Jean-François Couchot, Michel Salomon, Christian Parisod, and Jacques M. Bahi. Hybrid genetic algorithm and lasso test approach for inferring well supported phylogenetic trees based on subsets of chloroplastic core genes. *International Conference on Algorithms for Computational Biology, AlCoB 2015*.

[5] Stacia K. Wyman, Robert K. Jansen, and Jeffrey L. Boore. Automatic annotation of organellar genomes with dogma. *BIOINFORMATICS, Oxford Press*, 20(172004):3252–3255, 2004.

[6] Bassam Alkindy, Jean-François Couchot, Christophe Guyeux, Arnaud Mouly, Michel Salomon, and Jacques M. Bahi. Finding the core-genes of chloroplasts. *Journal of Bioscience, Biochemistery, and Bioinformatics*, 4(5):357–364, 2014.

[7] Bassam Alkindy, Christophe Guyeux, Jean-François Couchot, Michel Salomon, and Jacques Bahi. Gene similarity-based approaches for determining core-genes of chloroplasts. November 2014. Short paper.

[8] Jeffrey Rizzo and Eric C. Rouchka. Review of phylogenetic tree construction. *University of Louisville Bioinformatics Laboratory Technical Report Series*, (TR-ULBL-2007-01):2–7, 2007.

[9] Tim Blackwell Riccardo Poli, James Kennedy. Particle swarm optimization. *Springer Science + Business Media*, 1(10.1007/s11721-007-0002-0):33–57, 2007.

[10] Hidetoshi Shimodaira and Masami Hasegawa. Consel: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12):1246–1247, 2001.

A voxel-based tool for protein surface representation

Sebastian Daberdaku and Carlo Ferrari

Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131 Padova, Italy, sebastian.daberdaku@dei.unipd.it, carlo.ferrari@unipd.it

Keywords: macromolecular surface, high-resolution voxel surface, EDT.

Abstract. We present VoxSurf, a voxel-based tool for protein surface calculation, which can produce discrete representations of molecules at very high resolutions. The procedure can calculate the three main molecular surfaces, namely van der Waals, Solvent-Accessible and Solvent-Excluded, at high resolutions, by employing compact data-structures and implementing a spatial slicing protocol.
Fast Solvent-Excluded surface generation is achieved by adopting an approximate Euclidean Distance Transform algorithm. The algorithm exploits the geometrical relationship between the Solvent Excluded and the Solvent Accessible surfaces, and limits the calculation of the distance map values to a small subset of the overall voxels representing the macromolecule.
A parallelization scheme for the slicing procedure is also proposed and discussed.

1 Scientific Background

Different representations of the molecular surface can capture diverse aspects of the three-dimensional geometry of proteins and macromolecules in general. Currently, the most used methodologies are: the van der Waals surface (vdW), the Solvent-Accessible surface (SAS) and the Solvent-Excluded surface (SES) or Connolly surface. Protein surface calculations from given three-dimensional structures have been used extensively in modern molecular biology studies, and different methods to compute the three macromolecular surfaces have been proposed [1, 2, 3].

Among the explicit representations, the voxelized ones (also known as dot-surface or grid-based representations) are the most simple, and yet widely appreciated for their accuracy and applicability in various contexts. A voxel is the tiniest distinguishable element of a 3D object. It represents a single data point on a regularly spaced 3D grid, and can be thought as the three-dimensional equivalent of a pixel. Voxels can contain multiple scalar values (vector data) and have been extensively used for visualization and analysis of scientific and medical data.

Voxelized protein surfaces are currently being employed in descriptor-based protein docking and protein shape comparison. Kihara et al. propose protein docking, shape comparison and interface identification methods based on 3D Zernike descriptors (3DZD) [4], which are calculated over circular surface patches of voxelized macromolecular surfaces. In [5], dot-surfaces are used in the development of an invariant descriptor for the characterization of protein surfaces, suitable for various analysis tasks, such as protein functional classification or search and retrieval of protein surfaces over a large database. Invariant surface fingerprints have been introduced in [6] in order to compare local protein surface similarities rapidly and efficiently. The creations of these fingerprints employs a dot-surface representation of the molecular surface. Grid representations of protein surfaces have also been used in cavity detection, binding-pockets identification and evaluation techniques [7].

Although macromolecular structural data repositories, i.e. Protein Data Bank (PDB) [8], have long been available, they provide only limited surface representations, primarily aimed for visual purposes. Computing surface representations is still an application-

dependent task, resulting in different methodologies and parametrizations according to user requirements. Typically, protein surfaces have to be computed on the fly, adding high computational cost to the final application. To the extent of our knowledge, there are no tools available which can produce voxelized surface representations of macromolecules starting from their structural data (PDB files). Thus, the idea of developing a specific tool for the computation of voxelized macromolecular surfaces at arbitrary resolutions, starting from the macromolecular structure data derived from X-ray diffraction and NMR studies. By employing compact data-structures for the 3D grid representation, and implementing a spatial slicing strategy, this tool can calculate the three main molecular surfaces at very high resolutions with very little memory usage.

2 Materials and Methods

The first step of the proposed methodology consists in reading the three-dimensional representation of a macromolecule from its corresponding Protein Data Bank entry. The atomic coordinates of each atom composing the macromolecule are extracted and stored in a dedicated data-structure. The algorithm calculates the axis-aligned bounding-box enclosing the whole molecule by determining the minimal and maximal coordinates of each of the atoms in the molecule.

Given a desired grid resolution parameter, the dimensions of the voxel grid which will contain the molecule, are calculated. All atomic coordinates previously imported are translated, scaled and quantized to the new coordinate system defined by the voxel grid: each atom center is mapped in its corresponding voxel in the voxel grid.

By implementing a space-filling algorithm, all voxels surrounding a given atom center, are marked as occupied by that atom if their distance from its center is less or equal to the corresponding atomic radius. After all the atoms composing the macromolecule have been mapped into the grid, we obtain a voxelized representation of what is known as CPK model (also known as calotte model or space-filling model).

To obtain the van der Waals or the Solvent Accessible surfaces, we extract the surface voxels from the voxelized representation of the CPK volumetric model of the macromolecule. The Solvent Excluded surface is trickier to calculate because it includes the re-entrant surface portions. We have implemented a method based on the Euclidean Distance Transform (EDT) algorithm for surface smoothing.

The resulting voxelized surface is exported in an output file. We have chosen the Point Cloud Data file format of the Point Cloud Library (PCL), because of its simplicity, compactness and compatibility with different scientific visualization programs.

2.1 *Macromolecular Surfaces*

In the CPK volumetric model a molecule is represented as a set of spheres that can overlap, each one having a radius equal to the van der Waals radius of the atom it represents. For proteins and other macromolecules it is clear that the most of their van der Waals surface is buried in the inside of the molecules and is not accessible to the solvent or possible ligands. Thus the need to define the Solvent Accessible and the Solvent Excluded surfaces.

The Solvent Accessible surface (SAS) is defined as the geometric locus of the center of a probe sphere (representing the solvent molecule) as it rolls over the van der Waals surface of the molecule. The Solvent Excluded surface (SES) (or molecular surface or Connolly surface) is defined as the locus of the inward-facing probe sphere as it rolls over the van der Waals surface of the molecule. This surface can be considered as a continuous sheet consisting of two parts: the contact surface and the re-entrant surface. The contact surface is part of the van der Waals surface that is accessible to a probe sphere. The re-entrant surface is the inward-facing surface of the probe when it touches two or more atoms. There is a clear relation between the SAS and the SES, as the

Solvent Accessible surface is displaced outward from the Solvent Excluded one by a distance equal to the probe radius (figure 1).

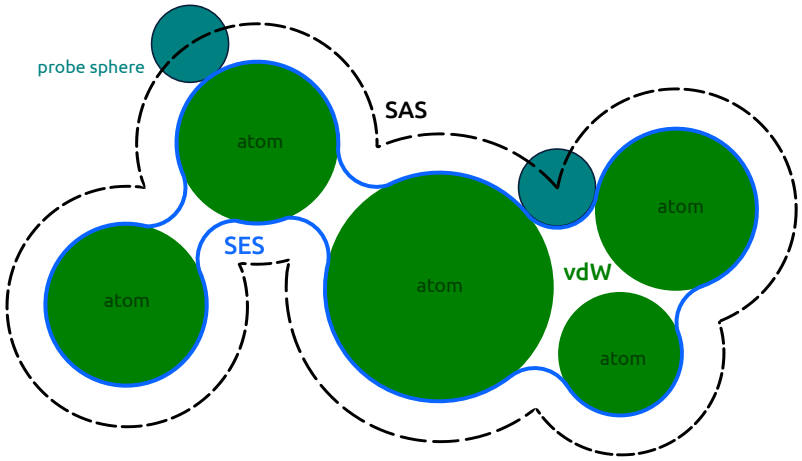


Figure 1: The three main molecular surfaces: van der Waals (green), Solvent Accessible (dashed/black) and Solvent Excluded (blue).

Macromolecules can have solvent-excluded cavities and voids, which might generate spurious surfaces inside the real molecular surface. To overcome this issue we have implemented a simple three-dimensional flood-fill algorithm, which "colors" the solvent-accessible voxels, starting from one of the eight vertices of the voxel grid.

2.2 *The Euclidean Distance Transform Method*

The Solvent Excluded surface computation is based on the Euclidean Distance Transform (EDT). The employment of the Euclidean Distance Transform for macromolecular surface computation was first introduced in [9]. A distance transform (also known as distance map or distance field), is a derived representation of a digital image (usually a binary image). Distance maps are images where the value of each voxel of the foreground is the distance to the nearest voxel (pixel) of the background.

Let the SAS be the set of all background voxels, and let us consider the EDT of the voxel grid $EDT(SAS)$. Because the Solvent Accessible surface is displaced outward from the Solvent Excluded one by a distance equal to the probe radius, it is clear that the SES can be obtained from the $EDT(SAS)$ by extracting all the voxels inside the Solvent Accessible volume with a distance map value equal to the probe radius.

Distance map values need to be calculated only for voxels inside the Solvent Accessible volume within one probe-sphere radius from surface voxels. This is achieved by the proposed implementation of the EDT calculation algorithm (based on the Region Growing method proposed in [10]) which uses masks like the Chamfer DT ($3\times3\times3$ and $5\times5\times5$ masks) and scans voxels by increasing distance value. This is implemented with a data-structure called Hierarchical Queues (HQ), made of a collection of FIFO queues, where in-going elements may enter any of the queues while outgoing elements are taken from the non-empty queue with the smallest label. The queue labeled i in the HQ contains the voxels for which i is the square of the distance to their nearest boundary voxel. For each voxel in the HQ, its location and nearest boundary voxel are stored, and, a map data-structure is also created in order to store the squared distances for each voxel. Varying the number of queues in the HQ, distance map values can be calculated only for the minimum necessary number of voxels required for the Solvent Excluded surface computation.

2.3 *The slicing procedure*

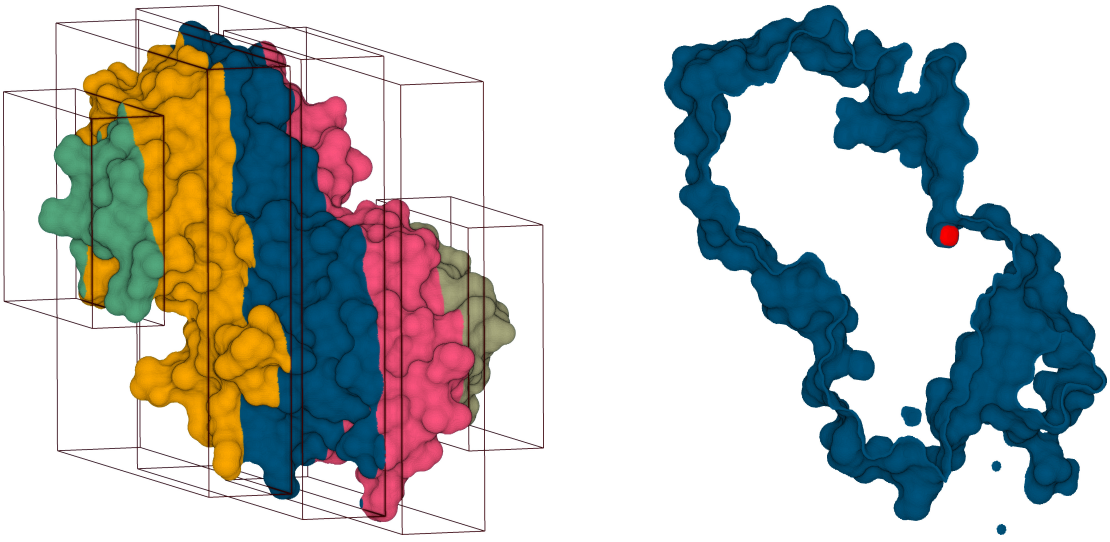
To enable the computation of high resolution surfaces in spite of memory limitations we have developed a slicing protocol for the macromolecule. The molecule is sliced in a user-defined number of parts, and the surface is calculated separately for each part, in

a sequential fashion. The slicing is done with planes perpendicular to the x-axis of the Cartesian coordinate system, as specified by the PDB (figure 2a).

Atom coordinates parsed from the PDB file are translated, scaled and quantized to the coordinate system defined by each slice. For each slice, we subtract the slice-length to the x coordinate of the translation vector $k-1$ times, where k is the current slice index ($k = 1, 2, ..., n$). The space filling procedure is performed for each slice separately, also taking into account any portions of atoms intersecting the slice whose centers might be located outside the current slice.

The correct determination of the distance map value for a given voxel requires knowledge of all boundary voxels within one probe sphere distance from the given voxel. Voxels in the immediate proximity of the slice borders require knowledge regarding the nearby boundary voxels in the adjacent slices in order to correctly calculate their distance map values. For this reason some extra margin on the x coordinate must be considered for each slice in order for the surface computation to yield correct results and it must be greater than the scaled and quantized probe-sphere radius.

Pockets and solvent-excluded cavities running through two or more slices must be distinguished from each other. The algorithm extracts all surface voxels belonging to potential pockets from each slice and stores them in an apposite data-structure. For each pair of adjacent slices, the border voxels of the candidate pockets are matched against their neighbors on the other slice. A candidate pocket is solvent-accessible if its border voxels have free neighbors on the adjacent slice. Otherwise, if its border voxels are matched with the border voxels of another candidate pocket on the adjacent slice, the current pocket remains undetermined as it could run through two or more slices in length. A back and forth scan of all adjacent slices is necessary before clearing the voxels of undetermined pockets, as they will surely be solvent-excluded (figure 2b).



(a) *Surface calculation with 5 slices*

(b) *Slice 3 with correctly detected pocket voxels (in red)*

Figure 2: SES of 4EYL (3501 ATOM entries), calculated with 5 slices, 1.4 Å probe-radius, 10^3 voxels per Å^3 resolution.

2.4 *Parallelization*

The macromolecular surface calculation protocol with slicing introduced in the previous sections suggests an immediate parallelization scheme. The surface calculation for each slice can be executed nearly-independently from the others, as process synchronization and communication is required only for the pocket-detection and extraction procedure, in order to correctly identify pockets spanning between two or more slices.

3 Results

We have run different tests of an MPI-based implementation of the parallel algorithm on an IBM®Power®P770 Server with 6 IBM®Power7 CPU's and 640Gb of RAM, running SUSE Linux Enterprise 11, and experimentally determined the speedup values for different input molecules at various resolutions, while calculating the three molecular surfaces.

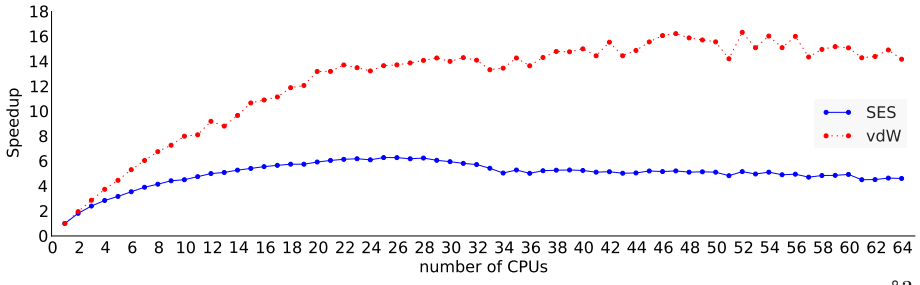
3.1 *Workload distribution*

To obtain the best results in terms of speedup, the slicing procedure should guarantee a uniform distribution of the workload between processes. A uniform distribution of the number of atoms per slice (i.e. variable-length slices) yields better speedup values than employing a constant slice length value, as the workload is split more evenly.

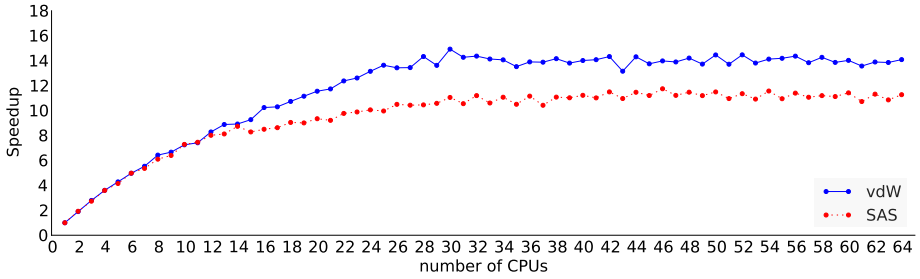
3.2 *Experimental Speedup*

The experimental speedup values that follow were calculated on the average execution time of different tests. The same configuration (PDB entry, desired molecular surface, resolution, probe radius, number of CPUs) has been executed 100 times, and the speedup was derived from the average computation time of these runs. We progressively increased the number of processors (from 1 to 64 CPUs) and evaluated the mean computation time for each configuration.

For a given PDB entry, tests have shown that the computation of the three molecular surfaces, at the same resolution, yields different speedup values for each surface (figures 3a and 3b).



(a) *vdW and SE surface computation speedup for 1GZX at 1000 voxels per Å^3.*



(b) *vdW and SA surface computation speedup for 2AEB at 1000 voxels per Å^3.*

Figure 3: Speedup values for 1GZX (T state haemoglobin - 4387 ATOM records) and 2AEB (human arginase I - 9568 ATOM records).

The vdW surface computation has a higher speedup. The CPK model creation takes most of the computation time in the vdW surface computation, which mainly depends on the number of atoms per slice. On the other hand, in the SES computation, the space-filling algorithm used in the solvent excluded cavity detection is the most time consuming task, which mainly depends on the size of the slice. The vdW surface computation also has no cavity detection step, and thus, no processes synchronization is needed.

The overall speedup is affected by the constant margin introduced in each slice during the SES computation. At some point, while increasing the number of processors

(increasing number of slices), the margin size will eventually become comparable to the effective size of the slice, thus vanishing the benefits of further parallelization.

4 Conclusion

We have effectively developed a tool which can produce voxelized macromolecular surfaces at very hight resolutions, even when faced with limited memory availability. For instance, the calculation of the surface for 1GZX at a resolution of 9000 voxels per $Å^3$ and while dividing the molecule in 10 slices, needs nearly 5GB of RAM, against the 2.6GB used while dividing the molecule in 20 slices (tests were made on a desktop computer with an Intel®Core™i7 860 CPU and 8GB of RAM (4×2GB DDR3-1333 banks) running Ubuntu 13.10 x64), which is an easily affordable amount of memory nowadays in desktop computers. By tuning the resolution and number-of-slices parameters various memory utilization rates can be achieved, depending on the users' needs.

Our parallel implementation introduces advantages in terms of the overall speedup, however the uniform distribution of atoms per slice may not necessarily yield a balanced workload between processes. On the other hand, the constant slice margin represents the main limitation to this parallelization scheme as it introduces constant overhead regardless of the slice size. These issues are left for future work.

Linux binaries and test results are available at: `http://www.dei.unipd.it/~daberdak/VoxSurf`

4.1 *Future directions*

We plan to employ the proposed tool in a bioinformatics laboratory context for *in silico* experimentations that require on the fly computation of molecular surfaces.

Acknowledgments

This work has been partially supported by the University of Padova ex60% grant "Advanced Applications in Computer Science".

References

[1] M. L. Connolly, "The molecular surface package," *J. Mol. Graph.*, vol. 11, no. 2, pp. 139–141, 1993.

[2] J. Weiser, P. S. Shenkin, and W. C. Still, "Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO)," *J. Comput. Chem.*, vol. 20, no. 2, pp. 217–230, 1999.

[3] M. F. Sanner, A. J. Olson, and J.-C. Spehner, "Fast and Robust Computation of Molecular Surfaces," in *Proceedings of the Eleventh Annual Symposium on Computational Geometry*, ser. SCG '95. New York, NY, USA: ACM, 1995, pp. 406–407.

[4] J. Esquivel-Rodriguez, V. Filos-Gonzalez, B. Li, and D. Kihara, "Pairwise and multimeric protein-protein docking using the LZerD program suite," *Protein Structure Prediction*, vol. 1137, pp. 209–234, apr 2014.

[5] Z. A. Deeb, D. A. Adjeroh, and B.-H. Jiang, "Protein surface characterization using an invariant descriptor," *Int. J. Biomed. Imaging*, vol. 2011, p. 15, 2011.

[6] S. Yin, E. A. Proctor, A. A. Lugovskoy, and N. V. Dokholyan, "Fast screening of protein surfaces using geometric invariant fingerprints," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 39, pp. 16 622–16 626, sep 2009.

[7] M. Weisel, E. Proschak, and G. Schneider, "PocketPicker: analysis of ligand binding-sites with shape descriptors," *Chem. Cent. J.*, vol. 1, no. 1, p. 7, 2007.

[8] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Mol. Biol.*, vol. 10, no. 12, p. 980, dec 2003.

[9] D. Xu and Y. Zhang, "Generating triangulated macromolecular surfaces by Euclidean Distance Transform," *PLoS One*, vol. 4, no. 12, p. e8140, dec 2009.

[10] O. Cuisenaire, "Region growing Euclidean Distance Transforms," in *Image Analysis and Processing*, ser. Lecture Notes in Computer Science, A. Bimbo, Ed. Springer Berlin Heidelberg, 1997, vol. 1310, pp. 263–270.

A workflow for the comparative analysis of MALDI-TOF mass spectrometric data in proteomics

E. Del Prete, D. d'Esposito, M. F. Mazzeo, R. A. Siciliano, A. Facchiano

Istituto di Scienze dell'Alimentazione, CNR, Via Roma 64, 83100, Avellino, Italy, {eugenio.delprete, fmazzeo, rsiciliano, angelo.facchiano}@isa.cnr.it, diegodesposito@alice.it

Keywords: proteomics, mass spectrometry, MALDI-TOF, similarity measures, R environment

Abstract. Mass spectrometry is a well-known technology used for the analysis of compounds in pure as well as in mixture form, and widely applied in large-scale studies as in the case of proteomics. Mass spectrum is a typical result, that is, a profile composed of an intensity value for each mass/charge value. In the case of large-scale analyses, raw mass spectra comparisons are difficult, as a consequence of different drawback typologies: data defects, unusual distributions, underlying disturbs and noise, bad data calibration. A bunch of data elaborations is essential, from data processing to feature extraction, in order to obtain a list of peaks from different mass spectra. In this work, a workflow has been developed to process raw mass spectra and compare the new tidy ones with the aim of defining a robust procedure, suitable for real applications and reusable for different kind of studies. Different similarity measures have been used for comparison purposes, in order to verify similarity among replicates and differences among analyzed samples. A case study is shown with the application of the processing method to data obtained from the analysis of different fish *genera*.

1 Scientific Background

Proteomics is a scientific discipline, involved in the study of large-scale datasets obtained from the analysis of entire protein mixtures (from the structure to the function), produced and modified by an organism, a tissue or a biological system. Along with other "omics" disciplines, it has a fundamental role in biomedical research, as identifying proteins linked with a physiological state and making a comparison with a pathological state (compared to a control). Mass spectrometry is an analytical technology, commonly used in proteomics and metabolomics, which performs analyses of proteins, peptides and metabolites: the basics are the possibility of splitting an ion mixture and the capability of discriminating the ions by their mass/charge value. This technology gives the benefit of getting large amounts of data in a short time and with a high-resolution, accuracy and sensitivity of molecular mass measurements. The most common data representation is a mass spectrum, an indented profile where mass/charge value is in the abscissa and the intensity is in the ordinate. The intensity is usually shown in percentage, in relationship with the tallest (base) peak. The comparison of mass spectra data is a common task in proteomics, in order to detect signals that can represent a signature of each group. The comparison can be performed among replicates from the same sample or, more interestingly, among different sample groups.

2 Materials and Methods

Data and tools

Mass spectra used to set up the procedure have been obtained from the analysis of protein extracts of fish muscle by MALDI-TOF-MS, as previously described in [1]. In particular, mass spectra have been chosen from eight different kinds of fish: four from the same *genus* (*Diplodus vulgaris*, *Diplodus sargus*, *Diplodus puntazzo*, *Diplodus annularis*), four from other *genera* (*Pagellus erythrinus*, *Lophius piscatorius*, *Trachurus trachurus*, *Auxis thazard*). Each fish has been represented by six mass spectra, three replicates for two samples. The work has been performed in *R* environment, mainly using three packages: *MALDIquant*, *MALDIquantForeign* [2] and *OrgMassSpecR* [3].

Data Processing

The workflow of our mass spectrometry data analysis is shown in Fig.1. The software related to the bio-spectrometry workstation supplies a data matrix, where the first column represents the m/z values (or m/z value, or mass) and the second one represents their intensity; a header is also present, with the information about the id number and the base (tallest) peak for each experiment. Metadata from header have been stored by means of regular expressions, mass spectrum data have been extracted from a .txt file and transformed in a special object for a simpler access. This object is constituted by class type, number and range of m/z values, range of intensity values, memory usage and mass spectrum name (if available). After raw data storage, two prearranged controls have been verified: if the mass spectrum is empty and if the distances between two consecutive mass points are equal or monotonically increasing.

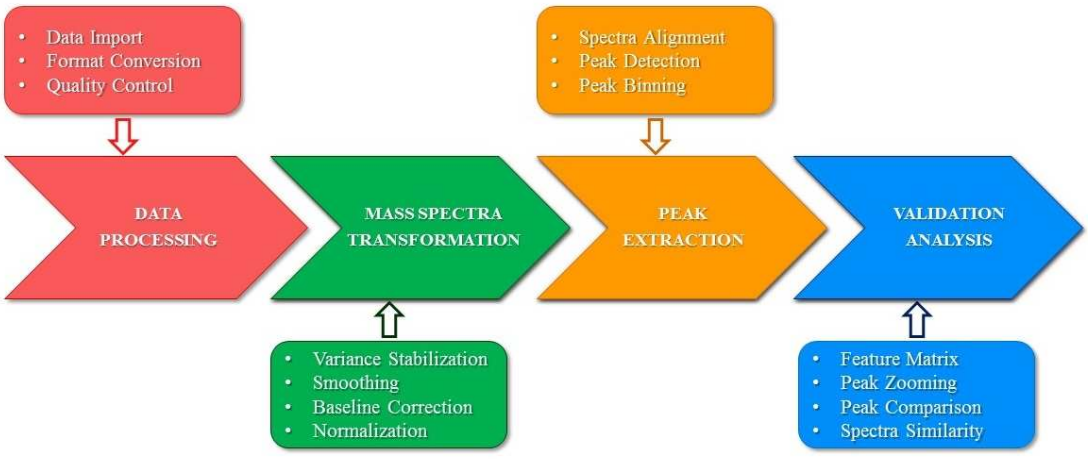


Fig.1 Mass spectrum data analysis workflow.

Mass Spectra Transformation

Due to specific features of MALDI-TOF-MS [4], peak intensity values could vary between different measurements on the same substance or sample. Hence, a strengthened normalization is compulsory for comparing several mass spectra from the same sample and different

experiments. Global and local transformations have been performed on mass spectra: one for a normalization on each entire mass spectrum, one for a normalization on different parts of the mass spectrum. In particular:

a) variance stabilization, in order to shift the data for a better graphical result and to avoid a dependency between variance and mean, for example, using a square root transformation;
b) smoothing, in order to reduce noise coming from artefacts or other underlying disturbs and, consequently, to improve the signal-to-noise ratio, for example, using the Savitsky-Golay filter;
c) baseline correction, in order to control the amplification of chemical noise in the low mass range, for example, using the Statistics-sensitive Non-linear Iterative Peak-clipping (SNIP) algorithm;
d) normalization, in order to preserve the proportionality between the intensity of different peaks of the mass spectrum, for example, using Total Ion Current (TIC) method, the most common normalization technique for MALDI-TOF data.

Peak Extraction

The set of the most important peaks from a mass spectrum represents a simpler fingerprint for the sample. At first, an alignment procedure is required to obtain a preliminary correspondence between the highest peaks from mass spectrum replicates and to preserve an average mass spectrum, useful for following comparison purposes. Local Weighted Scatterplot Smoothing (LOWESS) technique has been chosen as warping algorithm for the alignment, because phase errors need a correction due to their non-additive nature: it is different from other regression method (linear, polynomial), because it is applied on data subsets, with better performances in extracting local variability of data. After having estimated an overall noise on the mass spectra, peaks have been locally detected with the Median Absolute Deviation (MAD) method: a priori knowledge of mass spectra can help in selecting a suitable window size for data subsets and signal-to-noise ratio. Peak extraction shows that mass values are very similar, but not the same. A binning step has been executed, thus only one abscissa value represents the sets of aligned peaks from each mass spectrum.

Validation Analysis

Feature matrix can be useful for further statistical analysis: missing values have been interpolated from all the mass spectra and only the peaks over a selected frequency threshold have been kept. It is possible that only one interval of m/z values can be characteristic for a sample, thus a zoom on the mass spectrum or on the peaks can be more descriptive than the entire dataset. Moreover, numerical and graphical analyses of different mass spectra from the same sample have been performed. A good similarity measure is the cosine correlation, also called dot product, which considers two lists of peaks as vectors, and the cosine of the angle establishes how much they are similar. [5,6]. Cosine correlation has been calculated with *OrgMassSpecR* package, and a head-to-tail plot between a reference and a target mass spectrum has been shown for a visual comparison about peaks and a consequent mass spectrum discrimination. Because of the strict relationship between Pearson's correlation and cosine correlation [7], it is possible:

a) to extract a dissimilarity measurement (similar to Pearson's distance) and to create a distance matrix among the mass spectra, for studying their clusterization;
b) to provide a statistical validation on cosine correlation, using a t-Student's test with a significance level of 0.95.

3 Results

Starting from raw or partially processed mass spectra, it is possible to obtain a refined one with a smoothed trend, a low level of noise and evident peaks, which characterize the chosen dataset. In Fig. 2, five replicate mass spectra from *D. puntazzo* have been assembled together, in order to get a clear profile for this sample. The peaks have been marked with a cross: 249 intensity values have been recognized as peaks after the processing, approximately 1.4% of the total mass spectrum length, and ten of them are the highest, over a threshold of 25% of intensity. Thus, three m/z regions in this average mass spectrum include the most intense peaks: the first region around 4000, the second around 6000, the third around 12000, and they can be assumed as identifiers.

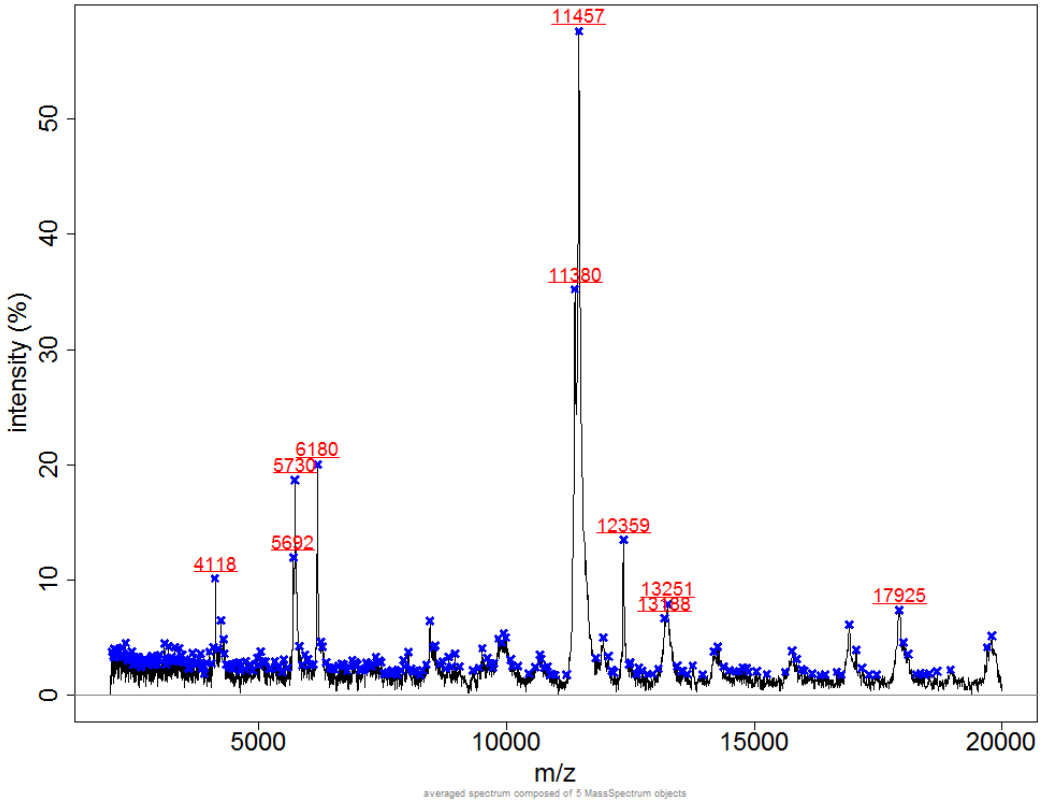


Fig. 2 Average mass spectrum extracted by five mass spectra from *D. puntazzo* samples. Crosses show all the detected peaks, numbers highlight the most important ten of them (more than 25% of intensity).

Two comparisons have been performed over the average mass spectrum, as shown in Fig. 3. After randomly choosing a mass spectrum from *D. puntazzo* to make a comparison and using the others to construct the average mass spectrum, head-to-tail graph shows that sample mass spectrum is very similar (taking in account the peaks) to the average one. Moreover, a quantitative measurement assures a similarity of 83.8%, thus, as expected, it is possible to say that sample mass spectrum is coherent with the average one. At the same time, the average mass

spectrum can be used for performing a control over a mass spectrum extracted from a different dataset. The external control has been made with an *A. thazard* mass spectrum: both graphical effect - different peak profile - and cosine correlation calculation - a similarity of 64.3% - reveal that the new mass spectrum is not coherent with the average one. In Fig.4, dataset dendrogram, built from the distance matrix, has been shown. Each fish *genus* has a well-defined cluster, under a threshold of 0.2. The only exception is recognizable on the fourth-fifth cluster, where *D. vulgaris* and *D. annularis* are clustered together. This result could be expected: these two fish, with same *genus* and different *species*, are not so easy to distinguish, that is, their mass spectra must be studied more in detail. Furthermore, *T. trachurus* is by the side of the *Diplodus genus*, very far from the other genera in term of mass spectrum similarity.

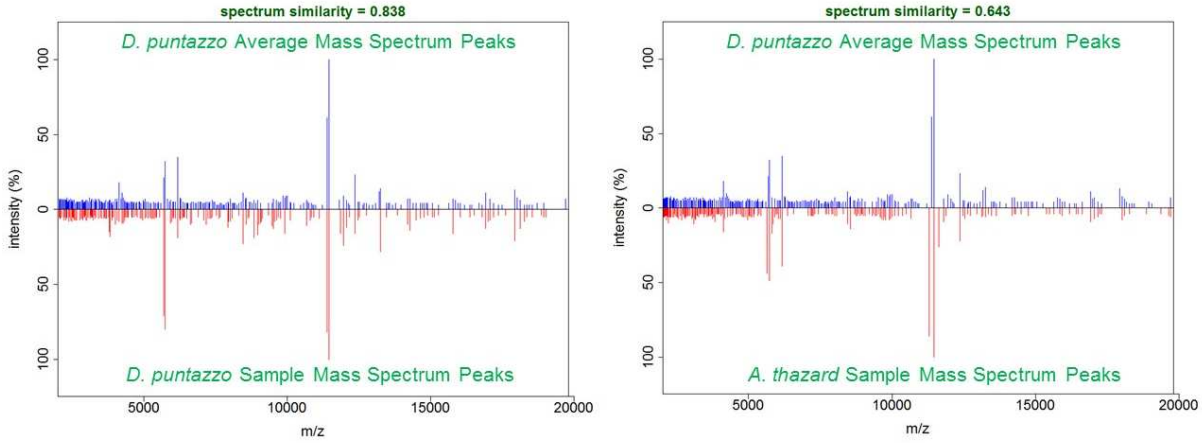


Fig. 3 Head-to-tail plots with spectrum similarity values. Left: plot between *D. puntazzo* average mass spectrum peaks (upper) and *D. puntazzo* sample mass spectrum peaks (lower). Right: plot between *D. puntazzo* average mass spectrum peaks (upper) and *A. thazard* sample mass spectrum peaks (lower).

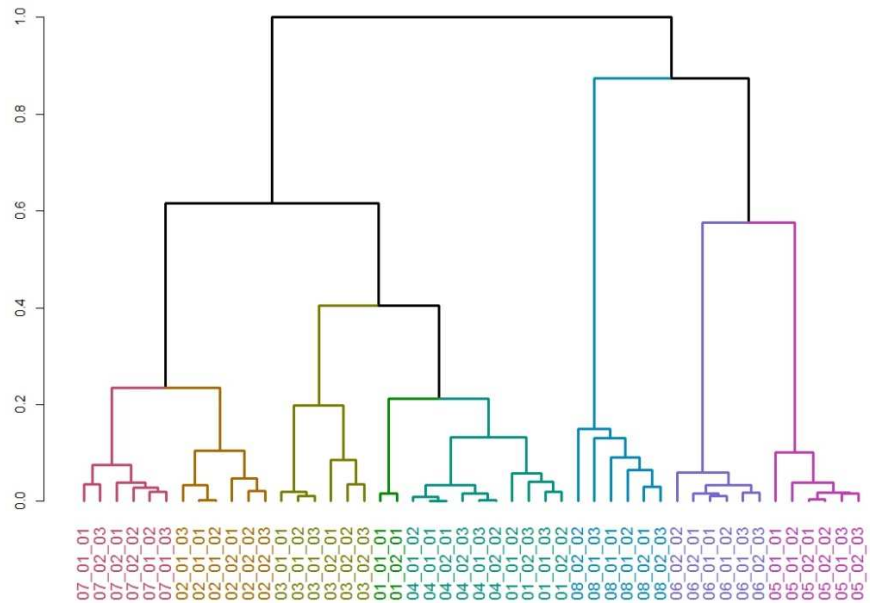


Fig. 4 Dataset dendrogram, in which the code represents fish_sample_replicate. Legend for fish: **01** *D. vulgaris*, **02** *D. sargus*, **03** *D. puntazzo*, **04** *D. annularis*, **05** *P. erythrinus*, **06** *L. piscatorius*, **07** *T. trachurus*, **08** *A. thazard*.

4 Conclusion

The proposed workflow starts from importing raw mass spectra data and it ends with analyzing and comparing tidy mass spectra data: intermediate steps concern mass spectra and peak processing. Peak extraction allows to implement a comparison between an average mass spectrum and a sample one, in order to perform a comparison and determine if the sample mass spectrum is graphically coherent with the average one, as shown in the first example, or not, as shown in the second one. Similarity measure (cosine correlation) and Pearson's distance quantify the difference among mass spectra. Possible future applications of this work include the creation of a tool for sample identification, on the basis of comparison to reference datasets, as already performed in the case of bacteria identification through the analysis of intact cells by MALDI-TOF [8], or for sample classification, especially useful in biomedical application, for instance to classify pathological or non-pathological samples [9].

Acknowledgments

This work is partially supported by the Flagship InterOmics Project (PB.P05), funded and supported by the Italian Ministry of Education, University and Research and Italian National Research Council organizations. This work is also partially supported by a dedicated grant from the Italian Ministry of Economy and Finance to CNR and ENEA for the Project "Innovazione e Sviluppo del Mezzogiorno e Conoscenze Integrate per Sostenibilità ed Innovazione del Made in Italy Agroalimentare (CISIA)" Legge n.191/2009.

References

[1] Mazzeo M F, De Giulio B, Guerriero G, Ciarcia G, Malorni A, Russo G L, Siciliano R A, "Fish authentication by MALDI-TOF mass spectrometry", *Journal of Agricultural and Food Chemistry*, 56(23):11071-6, 2008.

[2] Gibb S, Strimmer K, "MALDIquant: a versatile R package for the analysis of mass spectrometry data", *Bioinformatics*, 28(17):2270–2271, 2012, URL: http://strimmerlab.org/software/maldiquant/

[3] Stein S E, Scott D R, "Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification", *Journal of the American Society for Mass Spectrometry*, 5, 859-866, 1994.

[4] Duncan M W, Roder H, Hunsucker S W, "Quantitative matrix-assisted laser desorption/ionization mass spectrometry", *Briefings in Functional Genomics and Proteomics*, 7(5):355-370, 2008.

[5] Seongho K, Xiang Z, "Comparative Analysis of Mass Spectral Similarity Measures on Peak Alignment for Comprehensive Two-Dimensional Gas Chromatography Mass Spectrometry", *Computational and Mathematical Methods in Medicine*, Hindawi Publishing Corp., 2013.

[6] Seongho K, Aiqin F, Bing W, Jaesik J, Xiang Z, "An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure", *Bioinformatics*, 27(12):1660-1666, 2011.

[7] Gniazdowski Z, "Geometric interpretation of a correlation", *Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki*, 9 (7):27-35, 2013.

[8] Mazzeo M F, Sorrentino A, Gaita M, Cacace G, Di Stasio M, Facchiano A, Comi G, Malorni A, Siciliano R A, "MALDI-TOF mass spectrometry for the discrimination of foodborne microorganisms", *Applied and Environmental Microbiology*, 72, 1631-1638, 2006.

[9] Siciliano R A, Mazzeo M F, Spada V, Facchiano A, d'Acierno A, Stocchero M, De Franciscis P, Colacurci N, Sannolo N, Miraglia N, "Rapid peptidomic profiling of peritoneal fluid by MALDI-TOF mass spectrometry for the identification of biomarkers of endometriosis", *Gynecological Endocrinology*, 30, 872-876, 2014.

Ethno-pharmacology Based *In Silico* Approach Tracing Chymase Inhibitors from Herbal Nutraceutical Resources

Amit Dubey[1,2,3], Anna Marabotti[2,4], Pramod W. Ramteke[1] and Angelo Facchiano[2]

[1]Jacob School of Biotechnology and Bioengineering, Sam Higginbottom Institute of Agriculture, Technology and Sciences, Allahabad- 211007 (India).
[2] Istituto di Scienze dell'Alimentazione – CNR, via Roma 64 – Avellino-83100 (Italy).
[3]International Centre for Genetic Engineering and Biotechnology, AREA Science Park Padriciano 99, Trieste-34149 (Italy).
[4]Dipartimento di Chimica e Biologia, Università degli Studi di Salerno, Via Giovanni Paolo II 132, Fisciano-84084 (SA), (Italy).
(amit.dubey@isa.cnr.it, amarabotti@unisa.it, pwramteke@yahoo.com, angelo.facchiano@isa.cnr.it)

Keywords: Pharmacophore, Ethno-Pharmacology, Chymase, Herbal-Nutraceutical.

Abstract

Inhibition of chymase from herbal nutraceuticals is expected to reveal therapeutic approaches for the treatment of atherosclerosis disease, and fibrotic disorders. Ethno-pharmacology based chymase inhibitors search have better potential to provide a novel idea for herbal drug discovery. X-ray crystallographic protein data file of chymase in complex with different inhibitors were used to generate structure–based pharmacophore models. Two ligand–based pharmacophore models were developed from experimentally known inhibitors. The pharmacophore models were generated with the predefined feature types considered: hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), hydrophobic (HY), negative ionizable (NI), positive ionizable (PI), ring aromatic (RA). The structure-based pharmacophore models of chymase (PDB code: 1T32) with predefined features were tested for herbal nutraceutical compounds screening against ZINCPharmer online tool and afterwards, we validated pharmacophore models with Discovery Studio. The 4 pharmacophores models were used for screening of herbal nutraceutical databases, We also studied the chymase and its inhibitors by analyzing its identified pharmacophore features and the key amino acids Lys40, His57, Lys192, Phe191, Val146, Ser218, Gly216 and Ser195, Gly193, Ser214 and Ser195, which play a major role in chymase inhibition activity. Our study suggests that ethno-pharmacology based multiple pharmacophore approach is useful in determining structurally herbal nutraceutical diverse hits, which may fit to all possible bioactive conformations present in the active site of chymase enzyme. The scheme used in the current study could be appropriate to discover drugs for other enzymes as well.

Scientific background

Chymase enzyme (EC 3.4.21.39) belongs to hydrolase class, catalyzes the hydrolysis of peptide bonds and it is abundant in secretory granules of mast cells. Chymase is the considerable extravascular source of vasoactive angiotensin II (Ang-II), which is formed very readily through the hydrolysis of the Phe8–His9 bond of angiotensin I (Ang-I) [1]. Chymase is gathered in mast cells in a passive form and it is released as an active enzyme when mast cells are excited by injury or inflammation. Chymase shows enzymatic activity instantly after its release into the interstitial tissues at pH 7.4, following various stimuli in tissues (Figure 1). As chymase has no enzymatic activity in normal tissues, but its inhibitors shows non-toxic/safe features, specific chymase inhibitors may not have effects on any other target in normal tissues [2].

The epidemic of cardiovascular disease is a global phenomenon. One particular demonstration of cardiovascular diseases, heart failure (HF), is perilously increasing in regularity. A link between heart failure and chymase has been proved, and there is an enthusiasm to develop a specific chymase inhibitor as a new therapeutic treatment for the disease [3]. The quantity of cardiac mast cells is remarkably increased in patients with heart failure, and cardiac chymase may play an important role in the improvement of several cardiovascular diseases [4].

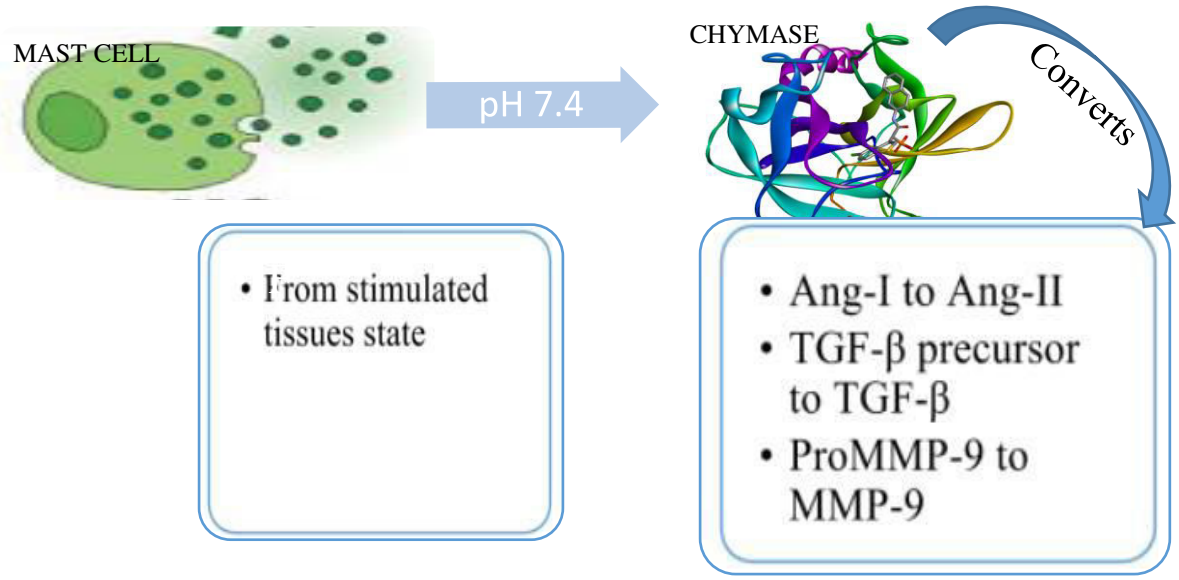


Figure 1. Active role of human mast cell chymase in cardiovascular diseases. Chymase shows enzymatic activity immediately after its release into the interstitial tissues at pH 7.4 following various stimuli in tissues. It catalyzes the conversion of angiotensin I (Ang-I) to angiotensin II (Ang-II) in vascular stimulated tissues. Chymase also converts precursors of transforming growth factor-β (TGF-β) and matrix metalloproteinase (MMP)-9 to their active forms, thus facilitating vascular response to injury. Both TGF-β and MMP-9 are involved in tissue inflammation and fibrosis, resulting in organ impairment.

The use of herbal drugs and formulations has always been an integral part of different ailments treatment in many communities across the world. The earliest recorded evidence of their use in Indian, Greek and Roman texts dates back to about 5000 years. Many plants with potential therapeutic activity were first widely used as herbal nutraceutical medicines with negligible undesired effects [5].

The objective of our ethno-pharmacology based study is to search for Chymase enzyme novel inhibitors from different herbal nutraceuticals, by computational screening and docking validation. Excellent candidates have then been selected for testing the hypothesis over *in vitro* and *in vivo* model experiments, Therefore, these molecules will provide a broad therapeutic window, and may become novel potential future drugs for the treatment of Atherosclerosis disease [6]. This approach represents a meeting point between traditional medicine bases and drug screening techniques. The objectives of the work includes:

- Identification and generation of novel pharmacophore models in chymase enzyme
- Screening and validation of active novel herbal nutraceutical chymase inhibitors from chemical databases for herbal nutraceutical compounds, with reference of plant species in use both in India and Italy

Material and Methods

Work Flow in Figure 1 shows the *in silico* Ethno-pharmacology aspects we have exploited in our study to identify herbal molecules potentially active in chymase inhibition.

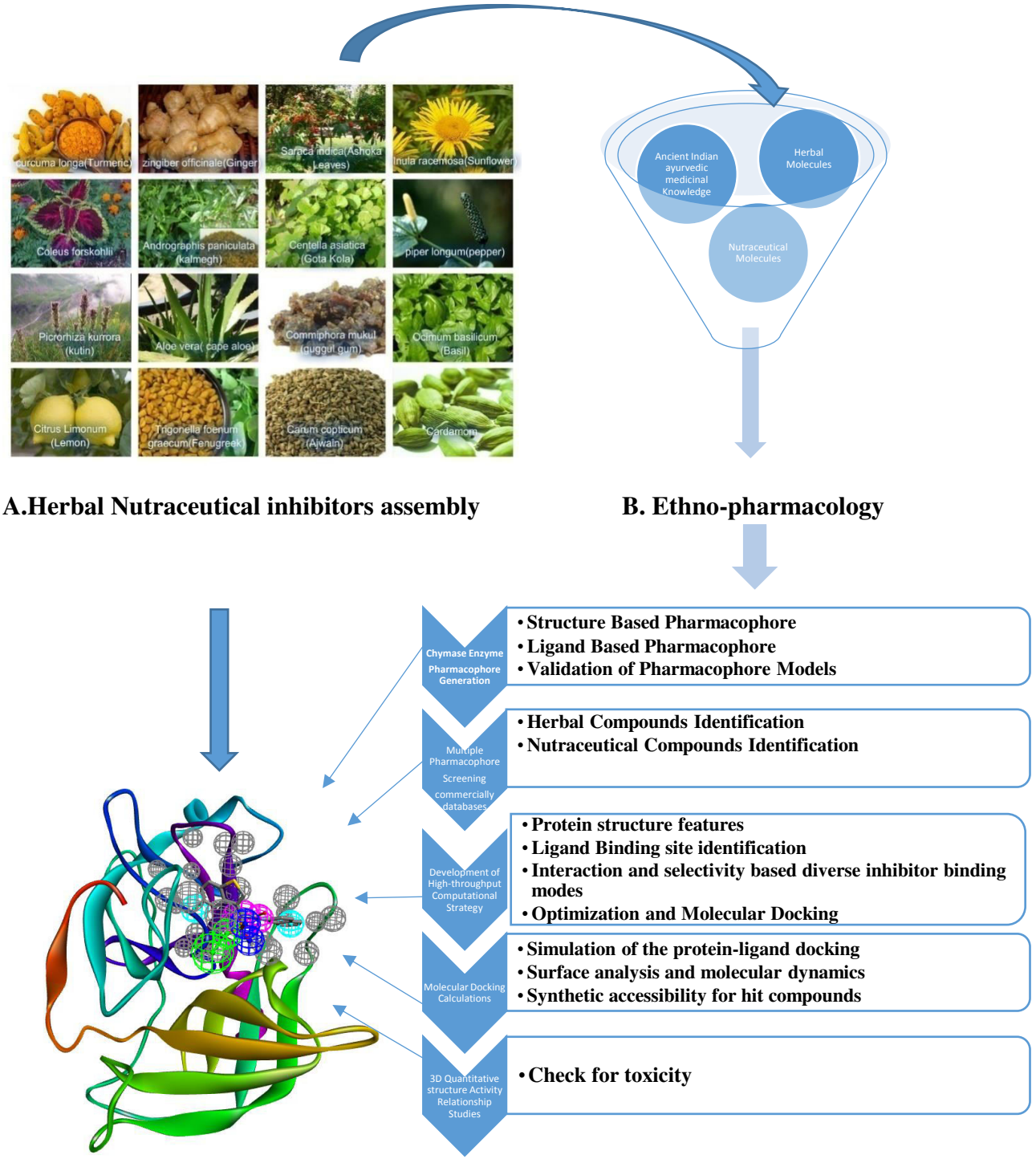


Figure 2. A. Herbal nutraceutical assembly chart (top left panel) shows different Indian-Italian herbs with medicinal properties; B. ethno-pharmacology (top right panel) comprising Indian Ancient Ayurveda approach for the selection of herbs and nutraceutical molecules; C. Chymase pharmacophore model (bottom left panel) used for tracing its inhibitors from herbal nutraceutical molecules; D. Computational framework (bottom right panel) showing the pipeline of the work.

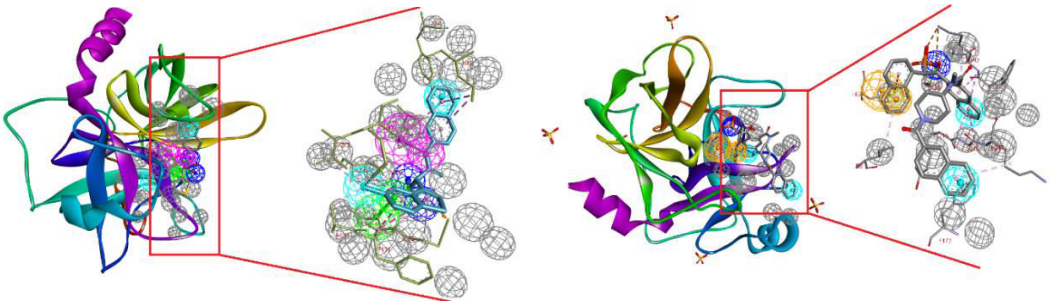
Crystal structures of chymase complexed to ligands were obtained from the Protein Data Bank [7] (PDB codes 2HVX and 1T32) and used for development of structure-based pharmacophore models.

The Receptor-Ligand Pharmacophore Generation protocol of Accelrys Discovery Studio v3.0, Accelrys, San Diego, USA, was used to perform this task with default parameters. This protocol generates selective pharmacophore models based on receptor-ligand interactions.

First, a set of features from the bound ligand was identified. With this method, we studied and developed both ligand-based and structure-based pharmacophore models on chymase inhibition, which were used for employing the virtual screening from chemical database for tracing herbal nutraceutical inhibitors (Figure 2). However, structure-based pharmacophore methodology involves generation of pharmacophore models directly from complex crystal structures and it is more reliable because it imposes the necessary constraints required for interaction and selectivity. The protein-ligand interactions and active site residues were studies by Discovery studio.
The structure-based pharmacophore models of 1T32 with predefined features were tested for herbal nutraceutical compounds (A special subset of Zinc database includes 1,434,178 natural product and 1, 93,269 natural product derivatives) screening against ZINCPharmer online tool [8] and the binding mode of the compounds inside chymase enzyme was investigated by rigid protein docking using AutoDock version 4.2. and setting up the system with ADT 1.5.6 software.

Results

In this study, two different 3D structures of chymase (1T32 and 2HVX) bound with two ligands were selected as input [9, 10] for structure-based pharmacophore generation. Results are shown in Figure 3. The following predefined feature types are considered: hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), hydrophobic (HY), negative ionizable (NI), positive ionizable (PI), ring aromatic (RA).



A. **B.**

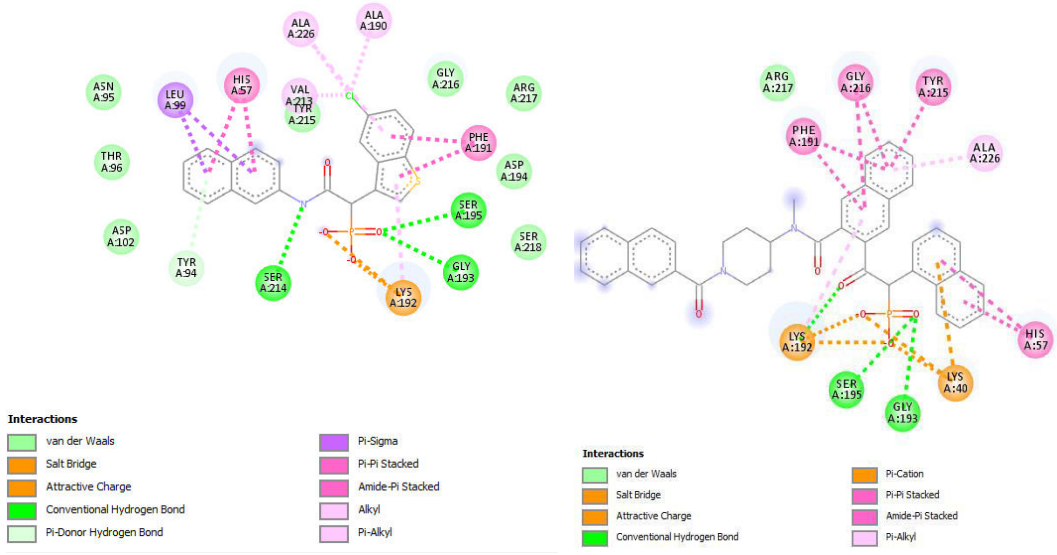
Figure 3. A. Chymase crystal structure (PDB ID: 2HVX), B. Chymase crystal structure (PDB ID: 1T32). Both structures shows active ligand interaction inside chymase. These ligand bound conformations were used for Receptor-ligand pharmacophore generation. Zoomed view clearly shows the arrangement of residues at the active site.

Chymase and its ligand interaction (Figure 4A,B) shows that pharmacophore features of the models were directed towards key amino acids like Lys192, Gly193, and Ser214 and Ser195, which play a major role in chymase inhibition activity [10].

We have performed the validation of pharmacophore models with Discovery Studio and found 39 features (HB_ACCEPTOR: 20, HYDROPHOBIC: 5, NEG_IONIZABLE: 2, RING_AROMATIC: 12) in ligand: 1T32, 6 features match with the receptor-ligand interactions: HHHNNR and 4 pharmacophores generated (Figure 5A) with 4 features (HHNR) and selectivity score value 8, 1653 were predicted.

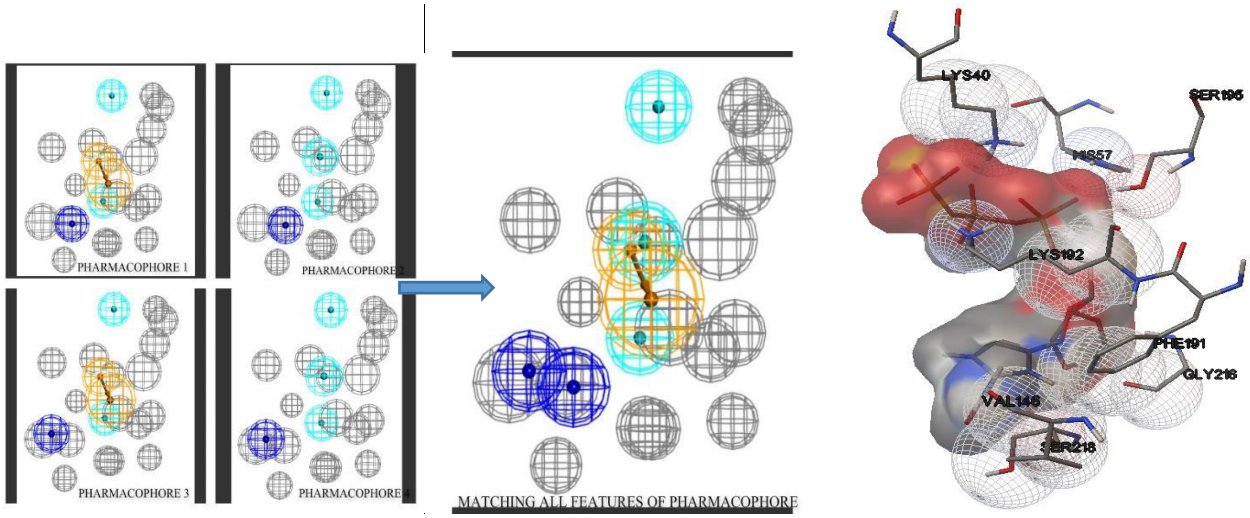
With the 4 pharmacophore models features (Figure 5A), we performed the pharmacophore screening and found 1,170 hits compounds among total 16, 27, 447 compounds (from natural products and their derivatives). Afterwards, further filtration screening were done on the basis of

their features, finally we found 7 potential lead compounds which were used for the rigid protein docking procedure.



A. **B.**

Figure 4. Chymase-ligand Interaction in PDB file 2HVX (panel A) and in PDB file 1T32 (panel B).



A. **B.**

Figure 5. Pharmacophores models with their features (Panel A) and Autodock Results shows binding of chymase and ZINC12402861 (Panel B)

In molecular docking experiments, we found that several molecules derived from herbal nutraceutical compounds are predicted to tightly bound to the active site of chymase. As an example, we show ZINC12402861 compound (Figure 5B) that is predicted to bind to chymase (the best result shows 15 different conformations in the same cluster and a predicted lowest binding energy of -12.41 kcal/mol). ZINC12402861 interacts with chymase key amino acids Lys40, His57, Lys192, Phe191, Val146, Ser218, Gly216 and Ser195. Moreover, it exhibited better predicted binding energy against chymase with respect to its co-crystallized inhibitor.

Conclusion

An established pharmacophore model from the three-dimensional structure of a target chymase protein illustrates helpful information for predicting protein-ligand interactions and further

advancement of ligand binding affinity. Finding novel and potent chymase inhibitors based on ethno-pharmacology will provide new ideas for drug design. We have utilized both ligand-based and structure-based processes to perform the pharmacophore screening to find herbal nutraceutical molecules from publically available databases. Distinct pharmacophore models generated from chymase crystal structures may show distinct inhibitor binding modes. So, we utilized a distinct features based on multiple pharmacophore model for virtual screening approaches that provides potent hits which was used to utilize various bioactive conformations accessible to fit in the active site of chymase enzyme. After successful validation, all pharmacophore models were employed for the Zinc database screening to reclaim hits with novel herbal nutraceutical chemical scaffolds, and finally we performed the molecular docking and come up with the prediction that some herbal nutraceutical compounds may be better inhibitors than the synthetic compounds already known. We are planning to incorporate surface analysis and molecular dynamics calculations for the investigation of electrostatic characteristics and conformational flexibility of selected herbal compounds.

Acknowledgements

This work has been partially supported by the Flagship InterOmics Project (PB.P05, funded and supported by the Italian Ministry of Education, University and Research and Italian National Research Council organizations). A. D. is supported by ICGEB Smart Fellowship program.

References

[1] Caughey GH, Raymond WW, Wolters PJ. "Angiotensin II generation by mast cell a-and b-chymases". *Protein Structure and Molecular Enzymology*, 1480: 245–257, 2000

[2] Amano N, Takai S, Jin D, Ueda K, Miyazaki M. "Possible roles of mast cell-derived chymase for skin rejuvenation". *Lasers in medical science,* 4: 223–229, 2009

[3] Amir RE, Amir O, Paz H, Sagiv M, Mor R, et al. "Genotype-phenotype associations between chymase and angiotensin–converting enzyme gene polymorphisms in chronic systolic heart failure patients". *Genetics in Medicine*, 10: 593–598. 2008

[4] Pejler G, Ronnberg E, Waern I, Wernersson S. "Mast cell proteases: multifaceted regulators of inflammatory disease". *Blood*, 115: 4981–4990. 2010

[5] Kamboj VP. "Herbal medicine". *Current Science*, vol. 78(1), 2000

[6] Balasuriya BWN, Rupasinghe HPV. "Plant flavonoids as angiotensin converting enzyme inhibitors in regulation of hypertension". *Functional Foods in Health and Disease*; 5; 172-188. 2011

[7] Berman HM, Henrick K, Nakamura H. "Announcing the worldwide Protein Data Bank". *Nature Structural Biology*, 10 (12): 98. 2003

[8] Koes DR, Camacho CJ. "ZINCPharmer: pharmacophore search of the ZINC database". *Nucleic Acids Research,* 40, W409-W414. 2012

[9] de Garavilla L, Greco MN, Sukumar N, Chen ZW, Pineda AO, et al. "A Novel, Potent Dual Inhibitor of the Leukocyte Proteases Cathepsin G and Chymase Molecular Mechanisms and Anti-Inflammatory Activity in Vivo". *Journal of Biological Chemistry*, 280: 18001–18007. 2005

[10] Greco MN, Hawkins MJ, Powell ET, Almond HR Jr, de Garavilla L, et al. "Discovery of potent, selective, orally active, nonpeptide inhibitors of human mast cell chymase". Journal of medicinal chemistry, 50: 1727–1730. 2007

A COMMENTARY ON A CENSORED REGRESSION ESTIMATOR

A. Eleuteri [(1)]

(1) Department of Medical Physics and Clinical Engineering, Royal Liverpool and Broadgreen University Hospital Trusts, Daulby Street L7 8XP, Liverpool, United Kingdom
Email: antonio.eleuteri@liv.ac.uk

Keywords: censoring, quantile regression, survival analysis, support vector machines

Abstract. In this note we evaluate the properties and performance of a censored regression estimator, as presented by different authors in the context of support vector regression. The estimator is based on minimisation of an inequality constrained loss in a linear program formulation. Using a theoretical argument, we conjecture that the estimator is not consistent, and we compare its performance with the Kaplan-Meier estimator on simulated and real data.

1 Scientific Background

Let us consider a sample of pairs $\{(T_i, C_i) : i = 1, \cdots, n\}$, $T_i \sim F$, T_i and C_i conditionally independent (though in the following we will assume the one-sample case.) Let us also consider the case of right censoring, so what we observe are actually the variables $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i < C_i)$, where $I(.)$ is the set indicator function. We consider the case of median estimation, and as mentioned before, for simplicity we focus our attention on the one-sample problem. We will denote by θ the median to be estimated.

The basic idea behind median (and generally, quantile) regression derives from observing that minimisation of the ℓ_1 loss for location estimates results in the median [3]. We denote the residual for the i-th observation in the uncensored case with $r_i = T_i - \theta$. The median loss function ℓ_1 (see Fig. 1) can then be written:

$$\rho(r) = r\{1/2 - I(r < 0)\}. \tag{1}$$

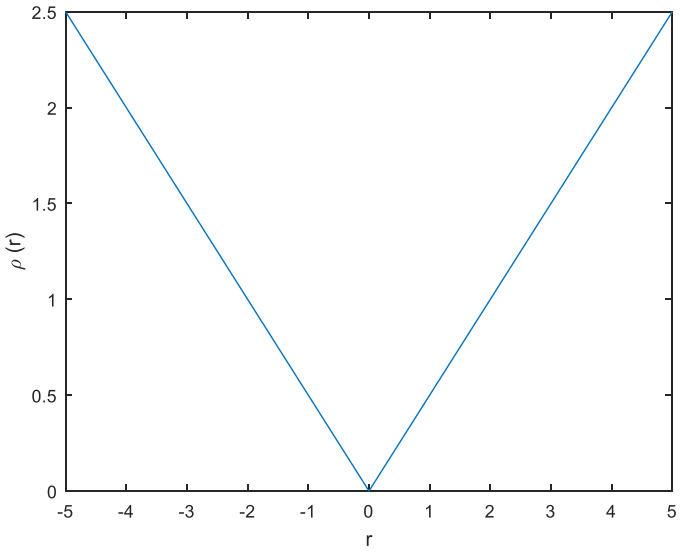


Figure 1: Median loss

Estimation of the median given a sample of observed points leads to minimisation of a piecewise linear empirical risk function:

$$\min_{\theta \in \mathbb{R}} \sum_i \rho(r_i) . \tag{2}$$

Due to the discontinuous nature of the median loss, a linear programming problem formulation is used in practice, by introducing $2n$ slack variables [3]:

$$\begin{aligned} &\min_{(\theta,u,v)\in\mathbb{R}\times\mathbb{R}_+^{2n}} \frac{1}{2}\sum_i u_i + \frac{1}{2}\sum_i v_i \\ &\text{s.t.} \quad \theta + u_i - v_i = Y_i \,, \forall i = 1 \ldots n \end{aligned} \tag{3}$$

In the censored case, we denote the residual for the i-th observation with $r_i = Y_i - \theta$. The loss function proposed in [1, 2] in this case is defined as (see Fig. 2):

$$\rho_I(r) = \delta\rho(r) + \tfrac{1}{2}(1-\delta) r I(r > 0) . \tag{4}$$

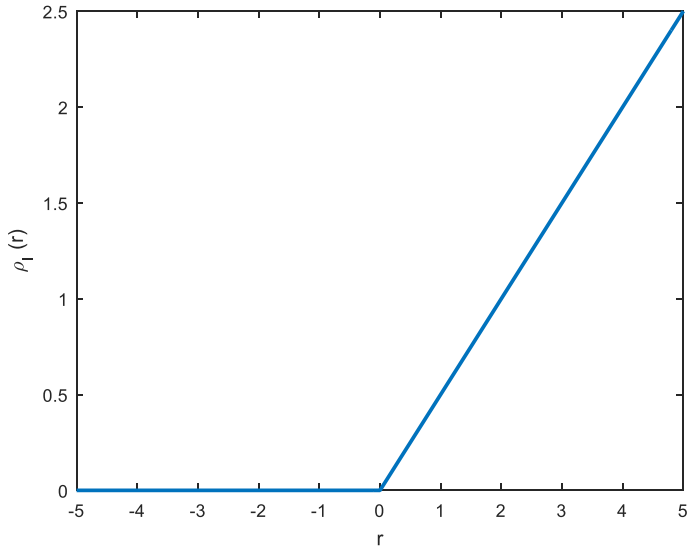


Figure 2: Inequality loss (for $\delta = 0$)

Note that this loss when evaluated on censored observations (i.e. when $\delta = 0$ in Eq. 4) is one-sided, and it reaches its minimum zero when $\theta > Y_i = C_i$ (i.e. the residual is negative, resulting in a zero loss.) In this way, estimates larger than the censored observations are "encouraged".

Similarly to the uncensored case, estimation of the median requires minimisation of a piecewise linear empirical risk function:

$$\min_{\theta \in \mathbb{R}} \sum_i \rho_I(r_i) , \tag{5}$$

that translates into the following linear programming problem:

$$\begin{aligned} &\min_{(\theta,u,v)\in\mathbb{R}\times\mathbb{R}_+^{2n}} \frac{1}{2}\sum_i u_i + \frac{1}{2}\sum_i v_i \\ &\text{s.t.} \quad \theta + u_i - v_i = Y_i \,,\ \forall i:\delta_i = 1 \\ &\qquad \theta + u_i - v_i \geq Y_i \,,\ \forall i:\delta_i = 0 \end{aligned} \tag{6}$$

Note the set of inequality constraints in correspondence to censored observations. The intuition behind this approach is simple: try to estimate the median by taking into account that censored observations provide a lower bound for the "true" unobserved points.

2 Materials and Methods

We will show now how the intuition behind the inequality constraint approach ignores some intrinsic and not readily evident aspects of the censoring process.

First note that in ordinary median regression the contribution of each point to the subgradient condition only depends on the sign of the residuals $r_i = T_i - \theta$ [3]. So a correct evaluation of the sign of the residuals is fundamental for any estimation procedure to work. Let us consider the two cases of uncensored and censored points.

For uncensored data we can observe both $Y_i = T_i < C_i$ and $I\left(r_i < 0\right)$; note that in this case the residuals can be either negative or positive.

For censored data, in the case $\theta < Y_i = C_i$, by the definition of right censoring we have $T_i > C_i$, hence we can observe $I\left(r_i < 0\right) = 0$.

However, if $\theta > Y_i = C_i$ there is an ambiguity: *we cannot observe the sign of the residual at all*, since we can have either $\theta > T_i$ *or* $\theta \leq T_i$, i.e. the residual can be negative *or* positive. In contrast, in the inequality loss formulation in Eq. 4, this case *always* results in a negative residual. As we will see with simulations, this fact has an impact on the performance of the estimator.

What can we say about a residual when we cannot observe it? We can evaluate the following conditional expectation (with respect to the measure *F*):

$$\mathbf{E}\left[I(r_i < 0) \,|\, T_i > C_i\right] = \frac{\Pr\left\{C_i < T_i < \theta\right\}}{\Pr\left\{C_i < T_i\right\}} = \frac{F(\theta) - F(C_i)}{1 - F(C_i)} = \frac{½ - F(C_i)}{1 - F(C_i)}. \tag{7}$$

The above quantity (calculated for $F(C_i) < ½$ since we are interested in the median) gives a measure of the "weight" attached to ambiguous observations. This suggests a weighting scheme originally proposed by Efron [4] and adapted by Portnoy [5] to quantile regression. Note that the weights depend on knowledge of the true distribution of the observations, which is usually not available; however, as shown in [5] these can be estimated nonparametrically using the Kaplan-Meier estimator of *F*. In fact, Kaplan-Meier quantiles can be framed as solutions of a quantile regression problem [6], in which censored observations are correctly weighted according to Eq. 7. The result of Eq. 7 suggests that the inequality loss estimator may not be consistent.

3 Results

In the following sections we report results on simulated and real data.

3.1 Simulations

We performed a series of simulation experiments to compare the finite sample performance of some estimates of the median (denoted by $\hat{\theta}$) in a censored one-sample setting.

We assume the distribution of events as standard lognormal with median $\theta = 1$, and the censoring distribution as exponential with mean 4. This results in approximately 30% censored observations. We follow the experimental setup in [6]. For each problem instance the estimate was calculated 1000 times and the results averaged. We also report the performance of the (infeasible) sample median (estimated on uncensored data) and the naïve estimator (i.e. the sample median ignoring the censored observations.)

In Tab. 1 we report the bias $\hat{\theta} - \theta$ of the estimates, and in Tab. 2 the mean squared error of the estimates (scaled to the sample size, to conform to the asymptotic variance calculations [6] and denoted by "n=∞" in the last row.)

Table 1: Bias

	Sample Median (infeasible)	Kaplan-Meier	Inequality loss	Naïve
n=50	0.0138	-0.0503	-0.0503	-0.214
n=200	0.00641	-0.00892	-0.0603	-0.221
n=500	-0.00159	-0.00750	-0.0674	-0.223
n=1000	-0.000160	-0.00206	-0.0659	-0.224

Table 2: Scaled Mean Squared Error

	Sample Median (infeasible)	Kaplan-Meier	Inequality loss	Naïve
n=50	1.674	1.756	1.555	3.424
n=200	1.780	2.023	2.268	10.860
n=500	1.565	1.902	3.693	25.955
n=1000	1.445	1.716	5.612	50.421
n=∞	1.571	1.839	-	-

From the tables we can see that the inequality estimator behaves qualitatively in a similar way as the naïve estimator, in that the bias is roughly constant independently of the sample size; and the scaled MSE increases with sample size (although at a different rate.) We extended the experiment for the inequality loss up to a sample size of 50000. The results support our conjecture that the estimator is not consistent.

Table 3: Inequality loss estimator performance

	Bias	MSE
n=2000	-0.0653	9.904
n=5000	-0.0668	23.715
n=10000	-0.0662	45.295
n=50000	-0.0673	228.027

In Fig. 3 we compare the naïve empirical risk and inequality loss-based empirical risk with the true (infeasible) empirical risk.

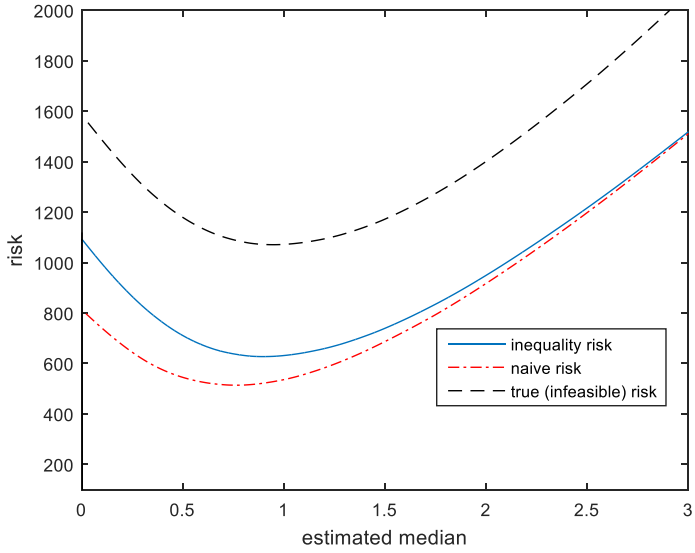


Figure 3: Empirical risk functions for lognormal data (median 1) with exponential censoring (mean 4). n=2000

3.2 Stanford Heart Transplant data

We assessed the performance of the median estimators on the Stanford Heart Transplant data (available in the R *survival* package as stanford2.) The sample size is 184, with 71 censored observations (38.6% censoring).

In Tab. 4 we report the median estimates of Kaplan-Meier, inequality and naïve estimators. In agreement with the simulation results, the inequality estimator provides a smaller estimate than the Kaplan-Meier one.

Table 4: Stanford Heart Transplant data

	Kaplan-Meier	Inequality loss	Naïve
Median Survival Time (days)	626	541	138

3.3 German Breast Cancer Study Group 2 data

As a further test, we analysed the data from the German Breast Cancer Study Group 2 [7]. The sample size is 686, with 299 censored observations (43.6% censoring).

In Tab. 5 we report the median estimates of Kaplan-Meier, inequality and naïve estimators. Again, and in agreement with the previous analysis and the simulation results, the inequality estimator provides a smaller estimate than the Kaplan-Meier one.

Table 5: German Breast Cancer Study Group 2 data

	Kaplan-Meier	Inequality loss	Naïve
Median Survival Time (days)	1806	1352	646

4 Conclusions

In this paper we have shown that a censored regression estimator independently proposed in literature by different authors doesn't appropriately take into account all the aspects of the censoring phenomenon. Comparison results with the Kaplan-Meier estimator (which is known to be consistent) through simulation and on real data suggest the estimator is not consistent, so discretion should be applied when using it to analyse data.

References

[1] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, B. De Moor. Risk Scores, Empirical Z-estimators and its application to Censored Regression. Technical Report kp06-105 (2006).

[2] P. Shivaswamy, W. Chu, M. Jansche. A Support Vector Approach to Censored Targets. Proceedings of the 2007 Seventh IEEE International Conference on Data Mining (2007).

[3] R. Koenker. Quantile Regression. Cambridge University Press (2005).

[4] B. Efron. The Two Sample Problem with Censored Data. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Prentice-Hall, New York (1967).

[5] S. Portnoy. Censored Quantile Regression. Journal of the American Statistical Association, 98, 1001-1012 (2003).

[6] R. Koenker. Censored Quantile Regression Redux. Journal of Statistical Software, 27-6 (2008).

[7] M. Schumacher, G. Basert, H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R.L.A. Neumann and H.F. Rauschecker for the German Breast Cancer Study Group. Randomized 2x2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. Journal of Clinical Oncology, 12, 2086–2093 (1994).

PrimatesDB: a functional resource on skeletal muscle tissue specific transcriptome of the *Pan troglodytes*

Daniela Evangelista*, Mariano Avino*, Kumar P. Tripathi, Mario R. Guarracino

Laboratory for Genomics, Transcriptomics and Proteomics (LAB-GTP), High Performance Computing and Networking Institute (ICAR), National Research Council of Italy (CNR), Via Pietro Castellino, 111, Napoli, Italy
*These authors contributed equally to this work

Keywords: *Pan troglodytes*, Annotations, Skeletal muscle, Transcriptome, Database, MySQL.

Abstract. Skeletal muscle represents a very well organized anatomical tissue in animals and its appearance might have predated the divergence of vertebrate and arthropods lineages about 700MYA. This diversified structure is very well visible in Primates where it differentiates according to their life styles and environmental conditions. This study focuses on *Pan troglodytes* - known as common chimpanzee - which belongs to a genus that is the most closely related to human species by which also shares a high similarity in the DNA composition. Our aim is to test the level of similarity between chimpanzee and human DNA - diversified to a functional phenotypic level to better adapt in different environmental conditions - by collecting skeletal muscle transcriptomic data from ENA (European Nucleotide Archive) database and performing its functional annotation analysis. We developed *PrimatesDB*, a freely available web resource which contains 30,944 sequences belonging to *Pan troglodytes* skeletal muscle transcriptomic data and from which it is possible to retrieve all the information related to 12,222 transcripts.
URL: www-labgtp.na.icar.cnr.it/PrimatesDB

1 Scientific Background

Pan troglodytes Blumenbach (common chimpanzee) belongs to the mammalian order Primates of Hominidae family [1]. With *Pan paniscus* (bonobo), it represents the only living species of genus *Pan* and it is the most closely related species to humans, sharing the last common ancestor circa 6 millions years ago (MYA) [2] [3] [4]. This is the reason why humans and chimpanzee DNA share at least 98% similarity. Altough common chimpanzees are not model organisms for biomedical human research area, recent advances in study of hepatitis C have shown that they sometimes become the only available source to test vaccines in humans [2]. Research into how the evolution of this primate is influenced by viruses, like HIV-1, which is found in chimpanzee as well, may have important implications in human health advances [3]. Moreover, all the genetic and phenotypic shared information between chimpanzee and humans is critical in order to elucidate the evolutionary scenario giving rise to humans [5]. Here we are approaching the first functional annotation analysis of chimpanzee to date believing that elucidating the localization, the biological and functional characterization of chimpanzee protein profile might be useful to discern the main phenotypic differences we see with humans, and, in general among most of the primates. We are selecting, as a case study, skeletal muscle tissue. Muscles might have evolved independently at least twice in animals from common ancestor contractile cells in sponge-grade organisms [6], once in cnidarians and cnetophores and another time in bilaterian (for example, vertebrates and insects). Specialized forms of skeletal and cardiac muscles predated the divergence of vertebrate/arthropode lineage circa 700 MYA while smooth muscle seemed to be evolved independently to other muscles. Vertebrate muscles are

contractile tissues that actively move body parts and form portions to several internal organs like the heart [7]. We are interested in striated muscles, which make up the skeletal musculature. In vertebrate, muscles are the active partners of bones, which represent their passive support. In primates, muscles are anatomically adapted of their particular life style. Postcranial skeletal is, for example, adapted to a great variety of locomotor, postural and feeding activities and it is not very well specialized, like in other non-primates vertebrate, such as horses and other ungulates or even whales. However, major changes among primates compared to other vertebrate are seen in locomotor morphology [7] being, the first ones more adapted to vertical leaping and clinging. Most nonhuman primates spend at least some time during the day in trees, therefore, grasping when climbing in arboreal environment is an essential component of their life. This is seen in feet and hands muscles morphology. Humans locomotors muscles are, on the other hand, adapted to a vertical postural position with evident changes in the vertebral column, pelvis, legs and feet. Thus, from [5] transcriptomic analysis, we retrieved the assembly database corresponding to the skeletal muscle myoblasts transcripts data (GABE01000001 - GABE01030945) of common chimpanzee stored in ENA. This database contains all the transcripts already assembled by the authors to their genome reference and it does not include any functional annotation. In order to obtain the biological information of the *Pan troglodytes* specific transcriptome, in our lab, we ran it into an existing computational pipeline [8]. This transcriptomic dataset was annotated and stored in *PrimatesDB*, which is a comprehensive web resource - driven on a relational database - with a user-friendly interface built to facilitate the retrieval of its functional annotations and other related integrated information. Currently, *PrimatesDB*, includes around 12,222 transcripts of the gene ontology and functional annotation of common chimpanzee skeletal muscles transcriptome. Moreover, hyperlink services are available for Ensembl and UniProt, so that users can gain diverse insights about the transcripts of interest from these publicly available resources.

2 Materials and Methods

2.1 Transcriptomic data retrieval

Starting from a recent study [5] carried out on *Pan troglodytes* transcriptomic analysis, we downloaded the assembly dataset corresponding to the skeletal muscle myoblasts transcripts data stored in ENA repository. This database contains all the transcripts already assembled by the authors to their genome reference. Unfortunately, this study does not include any gene neither functional annotation. In order to obtain biological information of this muscle specific *Pan troglodytes* transcriptome, we ran it into an existing computational pipeline [8] and carried out our downstream analysis.

2.2 Data Processing

Presently *PrimatesDB* not only accommodates the gene ontological information, but also other available functional annotations about domains, metabolic pathways, as well as, relevant biological information from SwissProt and PIR protein databases with respect to each and every muscle specific *Pan troglodytes* transcript. With the help of *PrimatesDB*, end users can obtain a comprehensive biological information for the differentially expressed transcripts IDs. Some biological information incorporated for each transcript within *PrimatesDB* regards: i. Gene Ontology: controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner; ii. Domain annotation: modular structure of the gene product, and evolutionary and molecular functional aspects of the transcripts, annotations for COG-Ontology, InterPro, PFAM and SMART domains are stored; iii. Metabolic pathway annotation: biological pathway information from KEGG, BBID and Panther resources.

2.3 Annotation analysis

Blastx program from locally installed ncbi-blast.2.2.23 stand alone package [10] is used (with threshold e-value 0.001) to identify the best hits for query sequences on locally installed SwissProt and UniProt-trEMBL databases [http://www.uniprot.org/]. The main goal of the first step is to find the similar sequences within SwissProt and UniProt-trEMBL databases for the unannotated query from the user. The output of BlastX run is an alignment file in a tsv format. This latter, using a bash script, is transformed into the protein list, which is the required input file for DAVID and QuickGO web services. Python client source code for DAVID web services employed in our pipeline, retrieves the functional and gene ontology annotation for every single transcript in a query data set. These python scripts take the input protein list file from previous step and utilize DAVID database to obtain information in the form of ChartReport, ClusterReport, TableReport and SummaryReport. For a given query data set, Python source code is implemented with default parameter for DAVID database search to obtain the TableReport obtained through DAVID webservices, it is a gene centric view which lists the genes and their associated functional and gene ontological annotation terms.

2.4 The pipeline

There are several ongoing projects [9] [10] which are useful for storing and for facilitating the information recovery of transcriptomic data without browsing the jungle of online web repositories. We implemented both the *PrimatesDB* database (DB) and a user-friendly web interface for a simpler content visualization. It is organized in tables with a relational structure containing all the items handled with proper data type, for a better performance of the database with respect to speed and deployment. The Graphical User Interface (GUI) of the web resource has been implemented as reported in figure 1, where the Home Page is specifically shown with its hierarchical structure. Currently here, it is possible to access to skeletal muscle data of *Pan troglodytes* which are organized in extremely easy-to-read tables for helping external users to quickly visualize and download the information.

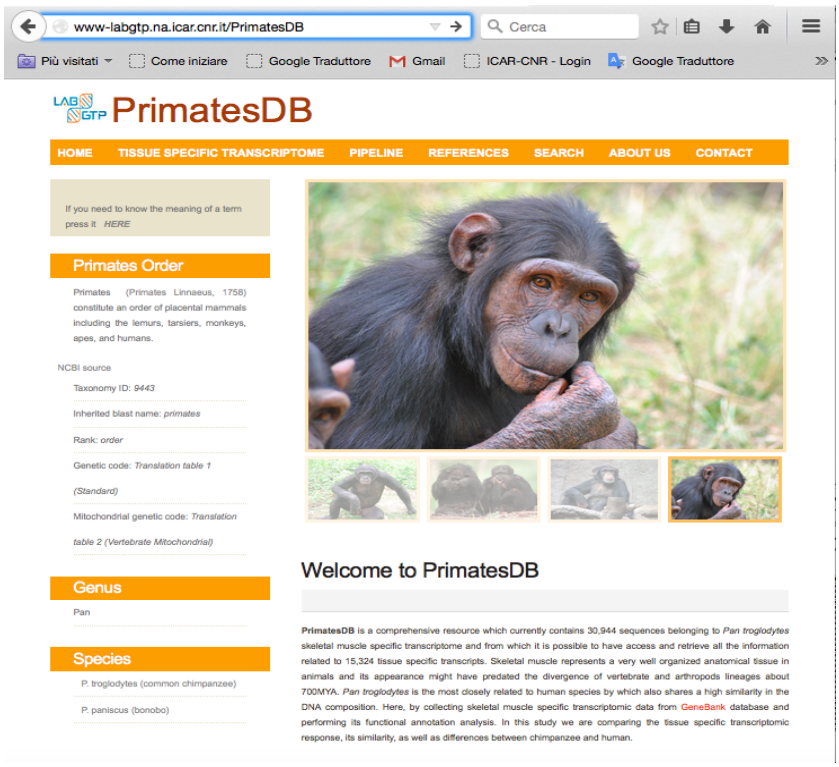


Figure 1: Screenshot of the web resource's Home Page

2.5 Database development and description

PrimatesDB resource grabs information from the database and inserts these latter into the proper web page each time it is requested. Updates in the database are reflected by the web page, which is dynamically queried on user request. *PrimatesDB* has been developed using web server Apache/2.2.26; MySQL client version 5.3.28 - 10.04.1 (Ubuntu) and the free tool phpMyAdmin version 3.3.2 deb1 Ubuntu 0.2.

2.6 PrimatesDB use case

PrimatesDB offers, in different ways, the opportunity to retrieve: i) skeletal muscle transcriptomic data from ENA database; ii) functional annotation from Transcriptator [8]; iii) information related to the transcripts ID from ENA and Ensembl repositories and; iv) protein ID knowledges from UniProt project. Here, we are going to describe the two developed ways to access to the data from the web resource, by *Tissue specific transcriptome* and *Search* section (Fig. 2).

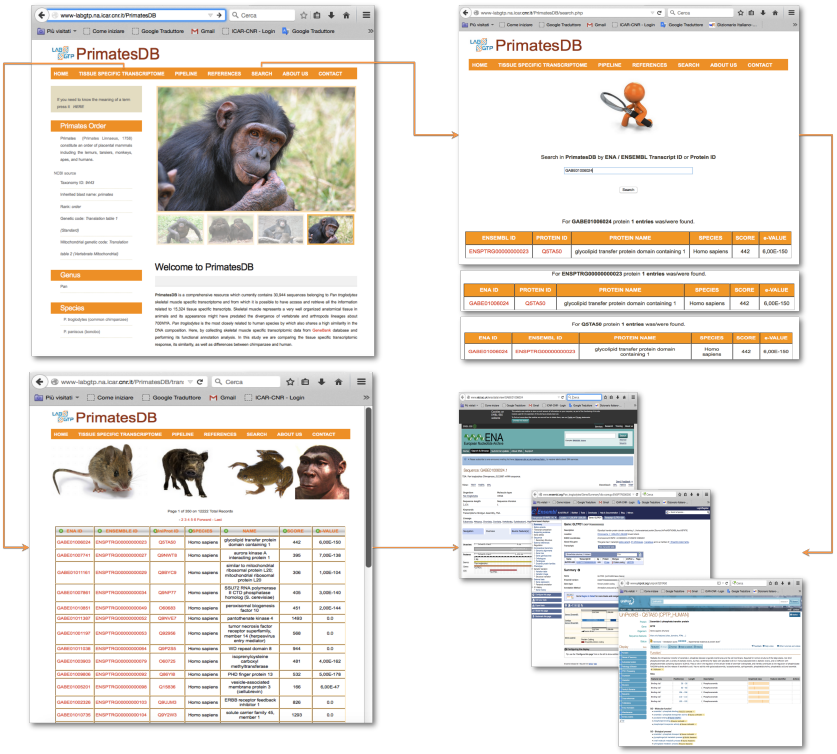


Figure 2: Possible ways to data recovery in *PrimatesDB*, starting from "Tissue specific transcriptome" and "Search" sections. The accessible integrated repositories are shown in the lower right corner.

By selecting the *Tissue specific transcriptome* section, from the navigation menu, user is able to visualize the complete list of the *Pan troglodytes* stored transcripts. The header is divided into seven columns (Fig. 3): the first corresponds to the skeletal muscle ENA IDs transcripts; the second one is composed of the ENSEMBL IDs, retrieved from BioMart ENSEMBL tool (giving ENA IDs as input in a search against Ensembl Gene 80 database and Pan troglodytes CHIMP 2.1.4 dataset). These first two columns give the user an important information about the species of interest showing which of those tissue specific transcripts have been already automatically annotated into the ENSEMBL specific organism database. The last five columns are particularly interesting for comparative studies, because they provide information about the best hit *UniProt* protein (the third column) with their relative *Species, Name, Score and e-value* (respectively fourth, fifth, sixth and seventh column) attributes. The lower right corner of figure 3 shows the top 15 species list - out of about 58 - for which BlastX program obtained the proteins hits. It is evident that most of the protein hits belonging to our non reference organism is present in Homo sapiens as well confirming that the two species are very

well related. By selecting the *Search* section, user can seek information by typing the ENA ID, the Ensembl ID or the UniProt ID. In each of these cases, he can access to all the information stored in *PrimatesDB* and also be redirected to the repositories themselves. In figure 2, we show a case study for the retrieval of the transcript information of the ENA GABE01006024, which is related to the Ensembl ENSPTRG00000000023 and Q5TA50 UniProt protein.

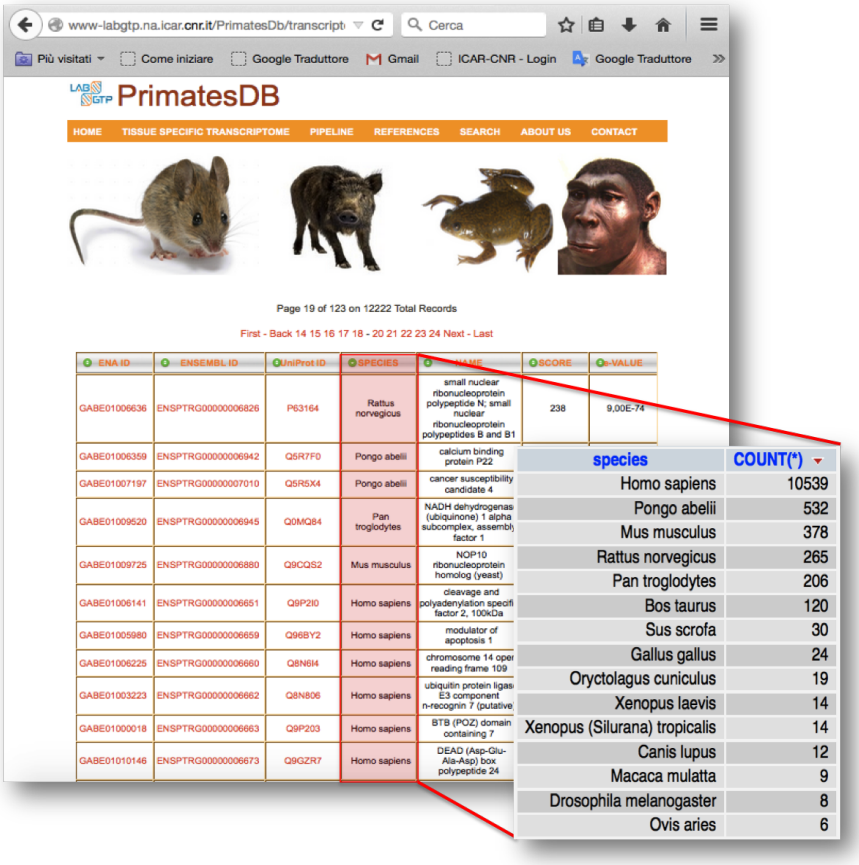


Figure 3: Page of the transcripts list. The columns ENA ID/ ENSEMBL ID/ UniProt ID are clickable buttons with redirect at the specific information of the related repositories. The enlarged picture shows the number of occurrences returned for each species from Swiss-Prot and UniProt-trEMBL databases.

3 Results

PrimatesDB is an open access and searchable database of complete annotation of the predicted tissue specific transcriptome of the non-reference organism *Pan troglodytes*. Its versatile and easily expandable structure accepts data from different sources which are automatically processed and integrated into the platform. The web interface allows the end-user to access to several sections. *PrimatesDB*, indeed, consists of seven sections which core of the web portal is represented by the *Tissue Specific Transcriptome* page. This section hosts the whole transcript list of the *Pan troglodytes* transcriptome identified by the analysis of our Python scripts. Currently, we are in the process of implement and increase the flexibility of dynamic content in the database through five database sets, which we have suitably merged in: Domain, Ontology and Pathways. All other sections were designed for all those users who want to deepen the understanding of this web application. The *PrimatesDB* web resource allows to structure the data and to display it in sorted and filtered tables accompanied by thorough explanations. The data were collected from the literature and external database, then appropriately handled with ad hoc scripts. Overall, information currently contained in *PrimatesDB* are related to 12 different functional terms of 12,222 transcripts (Fig. 3). The dataset contains

30,944 sequences, with relative lenghts ranging from 280 bp to 3100 bp, deposited into the ENA repository starting from GABE01000001 - GABE01030945.

4 Conclusion

The creation of a dedicated databases for non-reference model organisms is an important issue and always desirable. In our laboratory, we developed *PrimatesDB* a web resource for retrieving functional annotations on skeletal muscle specific transcriptome of the *Pan troglodytes*, the common chimpanzee, species closely related to the human one. The choice of this organism reflects the idea to start shading light of a life style, comparing it to other species phylogenetically related - but with different morphological characteristics - in order to understand how these might have allowed them to better adapt in their specific environments. Our analysis, begun with the retrieval of the specific transcriptomic data obtained from ENA and, the usage of an home-made computational pipeline [8] was used to process these data to place the functional annotations.To date, *PrimatesDB* represents a pilot study and it is useful to provide a comprehensive knowledge about the tissue specific transcriptome of the *Pan troglodytes* non-reference model organism. *PrimatesDB* is a very easy-to-use web resource, freely available and without login requirements. As a modular platform, *PrimatesDB* can easily be extended and customized to future demands and developments. Indeed, we are in the process of updating *PrimatesDB* resource to make it more informative and we aim to provide functional annotation for all other transcripts.

Acknowledgments

This work was funded by INTEROMICS flagship Italian project, PON02-00612-3461281 and PON02-00619-3470457. Mario R. Guarracino work has been conducted at National Research University Higher School of Economics and supported by RSF grant 14-41-00039.

References

[1] D.E. Wilson, D.M. Reeder. "*Pan troglodytes* in Mammal Species of the World". *Johns Hopkins University Press* A Taxonomic and Geographic Reference, 3 ed., 2005.

[2] J. Bukh. A critical role for the chimpanzee model in the study of hepatitis C. *Hepatology* vol. 39:6, pp. 1469-75, 2004.

[3] N.G. de Groot, R.E. Bontrop. The HIV-1 pandemic: does the selective sweep in chimpanzees mirror humankind's future?. *Retrovirology* vol. 10:1, pp. 53, 2013.

[4] The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* vol. 437:7055, pp. 69-87, 2005.

[5] M.D. Maudhoo, J.D. Madison, R.B. Norgren. "De novo assembly of the chimpanzee transcriptome from NextGen mRNA sequences". *GigaScience* vol. 4:1, pp. 1-4, 2015.

[6] P.R.H Steinmetz, J.E.M. Kraus, C. Larroux, J.U. Hammel, A. Amon-Hassenzahl, E. Houliston, G. Wrheide, M. Nickel, B.M. Degnan, U. Technau. "Independent evolution of striated muscles in cnidarians and bilaterians". *Nature* 487:7406 (2012): pp. 231-234.

[7] F. Ankel-Simons. Primate anatomy: an introduction. *Academic Press*, 2000.

[8] K.P. Tripathi, D. Evangelista, R. Cassandra, M.R. Guarracino. Transcriptator: a computational pipeline to annotate transcripts and assembled reads from rna-seq data. *Springer (ed.) Lecture Notes*Bioinformatics, XI International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics: 19-21 October 2011, Cambridge (UK), 2014.

[9] D. Evangelista, K.P. Tripathi, V. Scuotto. M.R. Guarracino "HvDBase: A Web Resource on Hydra Vulgaris Transcriptome". *Bioinformatics and Biomedical Engineering. Springer International Publishing* pp. 355-362, 2015.

[10] M. Scarpato, R. Esposito, D. Evangelista et al. "AnaLysis of expression on human chromosome 21, ALE-HSA21: a pilot integrated web resource." *Database*, 2014.

ALIGNMENT FREE DISSIMILARITIES FOR SEQUENCE CLASSIFICATION

Lo Bosco, Giosué[(1),(2)], La Neve, Dario [(3)]

(1) Universitá di Palermo
Dipartimento di Matematica e Informatica, Palermo, ITALY, giosue.lobosco@unipa.it

(2) I.E.ME.S.T.
Istituto Euro Mediterraneo di Scienza e Tecnologia, Palermo, ITALY

(3) Arancia ICT S.r.l, Palermo, ITALY, dario.laneve@gmail.com

Keywords: k-mers, L-tuples, DNA sequence similarity, DNA sequence classification, Knn classifier

Abstract. One way to represent a DNA sequence is to break it down into substrings of length L, called *L-tuples*, and count the occurence of each L-tuple in the sequence. This representation defines a mapping of a sequence into a numerical space by a numerical feature vector of fixed length, that allows to measure sequence similarity in an alignment free way simply using disssimilarity functions between vectors. This work presents a benchmark study of 4 alignment free disssimilarity functions between sequences, computed on their L-tuples representation, for the purpose of sequence classification. In our experiments, we have tested the classes of *geometric-based*, *correlation-based* and *information-based* dissimilarities, incorporating them into a nearest neighbor classifier. Results computed on three dataset of nucleosome forming and inhibiting sequences, shows that the geometric and correlation disssimilaritiess are more suitable for nucleosome classification. Finally, their use could be a valid alternative to the alignment-based similarity measures, which remains yet the preferred choice when dealing with sequence similarity problems.

1 Scientific Background

A fundamental biological question is to understand the function of the genome, and nowadays, despite the development of computational models for functional annotation, it still remains a daunting task. In particular, initial biological hypotheses about sequence similarity has been traditionally generated by using sequence alignment methods. Several algorithms that target specific goals such as global alignment, local alignment, with or without overlapping have been proposed [1, 2, 3]. The main issue of alignment methods is that their computational complexity escalates as a power function of the length of the related sequences. Despite the recent efforts in improving their computational efficiency [4, 5], the applications of alignment methods are not unlimited. They are based on the main assumption that functional elements are related to sequence substrings whose relative order is also conserved. Unfortunately there are cases showing that this can be violated, such as the cis-regulatory element sequences where there is little evidence suggesting that the order between different elements would have any significant effect in regulating gene expression. As such, the recently developed alignment-free methods [6] have emerged as a promising approach to investigate the regulatory genome. One of the methods belonging to this latter class is based on substring counting of a sequence, and is generally named as *k-mers* or *L-tuples* representation. Informally, L-tuples representation associates a sequence with a feature vector of fixed length, whose components count the frequency of each substrings belonging

to a finite set of words. The main advantage is that the sequence is represented into a numerical space where a particular dissimilarity function between the vectors can be adopted to reflect the observed similarities between sequences. L-tuples have shown their effectiveness in several in-silico analysis applied to different genomics and epigenomics studies. The interested reader can find the basic ideas of L-tuples based methods to different biological problems in the following review [7].

2 Materials and Methods

A generic DNA sequence s of length M can be represented as a string of symbols taken from a finite alphabet. We can think to a particular mapping function that project s into a vector $\mathbf{x}_s$ (the feature vector), allowing to represent s into a multi-dimensional space. One of the most common way of defining such mapping, is to consider a feature vector $\mathbf{x}_s$ that enumerates the frequency of occurrence of a finite set of pre-selected words $W = \{w_i, .., w_m\}$ in the string s. The simplest and most common definition of W is by using *L-tuples* (or *k-mers*), i.e. a set containing any string of length L whose symbols are taken in the nucleotide alphabet $\Sigma = \{A, T, C, G\}$. In this case, each sequence s is mapped to a vector $\mathbf{x}_s \in \mathbb{R}^m$ with $m = 4^L$, such that the component $(\mathbf{x}_s)_i$ counts the occurrence of the $i-th$ L-tuple into the string s. The counting process uses a window of length L that is run by step of 1 through the sequence, from string position 1 to $M - L + 1$.
Let $\mathbb{X}$ be a set. A function $\delta : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a *dissimilarity* on $\mathbb{X}$ if, $\forall\ \mathbf{x}, \mathbf{y} \in \mathbb{X}$, it satisfies the following three conditions:

1. $\delta(\mathbf{x}, \mathbf{y}) \geq 0$ (*non-negativity*);
2. $\delta(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{y}, \mathbf{x})$ (*symmetry*);
3. $\delta(\mathbf{x}, \mathbf{x}) = 0$;

One can categorize disimilarity functions according to three broad classes: *geometric-based*, *correlation-based* and *information.based*. Functions in the first class capture the concept of *physical* distance between two objects. They are strongly influenced by the magnitude of changes in the measured components of vectors $\mathbf{x}$ and $\mathbf{y}$, making them sensitive to noise and outliers. Functions in the second class capture dependencies between the coordinates of two vectors. In particular, they usually have the benefit of capturing positive, negative and linear relationships between two vectors. Functions in the third class are defined via well known quantities in information theory such as entropy and mutual information. They have the advantage of capturing statistical dependencies between two data points, even if they are not linear. We now formally define the functions of interest for this work, starting with the geometric ones. The *Euclidean* or *2-norm* dissimilarity is defined as follows:

$$d_e(\mathbf{x}, \mathbf{y}) = \sqrt[2]{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{1}$$

where $\mathbf{x} = (x_1, \ldots, x_m)$, $\mathbf{y} = (y_1, \ldots, y_m)$.
Among the correlation-based dissimilarities, the most known is the *Pearson* disssimilarity d_r

$$d_r(\mathbf{x}, \mathbf{y}) = 1 - r = 1 - \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_j - \bar{x})^2 \sum_{j=1}^{m}(y_j - \bar{y})^2} \tag{2}$$

where $\bar{x} = \frac{1}{m}\sum_i x_i$, $\bar{y} = \frac{1}{m}\sum_i y_i$. The *Cosine Distance*, is another example of

correlation-based dissimilarity, and can be defined as:

$$d_{cos}(\mathbf{x},\mathbf{y}) = 1 - (\frac{\mathbf{x}\cdot\mathbf{y}}{\sqrt{\mathbf{x}\cdot\mathbf{x}}\sqrt{\mathbf{y}\cdot\mathbf{y}}}) \tag{3}$$

that corresponds to 1 minus the cosine of the angle between the two vectors $\mathbf{x}$ and $\mathbf{y}$. Finally, the symmetrical Kullback-Leibler dissimilarity between two vectors $\mathbf{x}$ and $\mathbf{y}$ belongs to the class of information-based dissimilarities, and is so defined:

$$d_{kl}(\mathbf{x},\mathbf{y}) = \frac{\sum_i p_{x_i} log_2 \frac{p_{x_i}}{p_{y_i}} + \sum_j p_{y_j} log_2 \frac{p_{y_j}}{p_{x_j}}}{2} \tag{4}$$

where $p_{x_i} = \frac{x_i}{M-L+1}$. This dissimilarity is able to measure the difference between the probability distributions of the L-tuple representation of two sequences.

2.1 K-nearest neighbor classsifier

K-nearest neighbor classifier (*Knn*) is a very simple classification method that has been widely used in the realm of machine learning. It makes use of the notion of neighborhood, defined by a distance function between elements. Let R be the number of classes, T_i the training set of elements for a class i and δ a distance between elements. Let $Y(\mathbf{x})$ be the set of K elements closest to an unlabeled element $\mathbf{x}$ with respect to δ, then the Knn classifier assign to $\mathbf{x}$ the class j using the following assignment rule :

$$j = \underset{1\leq i\leq R}{\arg\max}(|Y(\mathbf{x}) \cap T_i|) \tag{5}$$

This rule means that the unlabeled element x is classified by assigning the label which is most frequent among the K training samples nearest to that point. Note that K is a parameter of the method, and its best choice depends upon the data. Generally, larger values of K are preferred since the effect of noise on the classification can be reduced.

2.2 Dataset description

In this study we have considered three datasets of DNA sequences underlying nucleosomes from the following three species: (i) *Homo sapiens (HM)*; (ii) *Caenorhabditis elegans (CE)* and (iii) *Drosophila melanogaster (DM)*. The nucleosome is the primary repeating unit of chromatin, which consists of 147 bp of DNA wrapped 1.67 times around an octamer of core histone proteins. Several studies have shown that nucleosome positioning plays an important role in gene regulation and that distinct DNA sequence features have been identified to be associated with nucleosome positioning. Details about all the step of data extraction and filtering of the three datasets can be found in the work by Guo et al [8] and in the references therein. Each of the three datasets is composed by two classes of samples: the nucleosome-forming sequence samples (positive data) and the linkers or nucleosome-inhibiting sequence samples (negative data). The HM dataset contains $2,273$ positives and $2,300$ negatives, the CE $2,567$ positives and $2,608$ negatives and the DM $2,900$ positives and $2,850$ negatives. The length of a generic sequence is 147 bp.

3 **Results**

Starting from a dataset S of n sequences , the training and test of the Knn classifier have been selected using a a 10 fold cross validation schema. We have used as numerical dataset for the classifier a matrix D_S of size $n \times 4^L$, such that $(D_S)_i^j$ stores the countings c_i^j of the $j-th$ L-tuple w_j into a sequence s_i of the dataset. We have computed the experiments for different L ranging from 1 to 6. Regarding the K value of the Knn classifier, we have decided to compute experiments for each odd value (the number of

classes is $R = 2$) ranging in the integer interval $\{1, ., 21\}$. We have computed a total of 3 indices to measure the performance of the classifier: *Accuracy (A), Sensitivity (Se) and Specificity (Sp)*. In the following, we recall their definitions:

$$A = \frac{TP + TN}{TP + FN + FP + TN},\ Se = \frac{TP}{TP + FN},\ Sp = \frac{TN}{FP + TN} \tag{6}$$

where the prefix T (true) indicates the number of correctly classified sequences, F (false) the uncorrect ones, P the positives class and N the negatives class. We have computed the accuracy results of 4 Knn classifiers, referred as *Knn-euclidean, Knn-correlation, Knn-cosine* and *Knn-Kullback*. Each one of them incorporates the corresponding distances defined in Section 2. Figures 1,2,3 show the mean accuracy, sensitivity, specificity values computed among all the considered neighbors K, for each values of $L = \{2, ., 6\}$. In Table 1 we report mean (μ) and standard deviation (σ) of the the best accuracy, sensitivity and specificity values (in percentage) reached by each one of the classifiers, for different L (L-tuple length). Results show that the Knn-euclidean is the best performer in terms of sensitivity (91% on CE and HM) and accuracy ($> 83\%$ CE, $> 76\%$ DM, $> 84\%$ HM). It is also evident (see Figures 1,2,3) that the the Knn-euclidean classifier is also invariant, in terms of accuracy, to the used L. Knn-cosine is the second best performer, in terms of accuracy ($> 83\%, L = 3$ CE, $> 76\%, L = 3$, DM). Knn-kullback seems only suitable for the DM dataset, which represent the most difficult dataset to classify (best accuracy 77%). Knn-correlation is the worst performer, showing only good specificity values for $L = 6$. Finally, it is also observable that the best choice for L is 3 and 4, and that all the classifiers' sensitivities and accuracies tend to decrease while L increases.

Table 1: In column, for each L in the range $\{2,.,6\}$ the mean (μ) and standard deviation (σ) values of the K-nn classifier accuracy (A), sensitivity (Se) and specificity (Sp) computed on 10 folds in the cases of the Caenorhabditis elegans (CE), Drosophila melanogaster (DM) and Homo sapiens (HM) dataset, for each one of the considered classifiers. In bold, the values with the best values for each dataset.

	L=2						L=3						L=4						L=5						L=6					
	A		Se		Sp		A		Se		Sp		A		Se		Sp		A		Se		Sp		A		Se		Sp	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
CE-Knn-euclidean	82	1	**90**	1	74	3	**83**	1	**91**	2	76	3	**84**	2	89	2	78	3	**83**	1	86	2	81	2	**83**	2	82	3	83	1
CE-Knn-correlation	80	1	74	2	86	2	78	1	65	2	92	1	73	1	52	3	**94**	1	69	1	43	3	**95**	1	67	1	38	2	**95**	1
CE-Knn-cosine	82	1	89	2	75	2	**83**	1	84	2	83	2	79	2	68	2	90	1	72	2	50	3	94	1	68	1	41	3	**95**	1
CE-Knn-Kullback	82	1	87	2	76	2	68	1	**90**	1	47	3	81	1	75	2	87	2	72	1	47	2	**95**	1	68	1	39	3	**96**	1
DM-Knn-euclidean	**76**	1	78	2	74	1	**77**	1	80	1	75	3	**76**	2	79	3	74	2	75	1	80	1	71	2	75	1	80	2	70	2
DM-Knn-correlation	72	1	73	2	69	2	70	2	75	2	64	2	69	1	71	2	66	2	68	1	62	2	74	2	68	2	52	3	**84**	2
DM-Knn-cosine	**77**	2	79	2	74	2	**77**	2	82	2	72	3	73	2	79	1	67	3	70	1	68	3	73	2	69	2	55	1	**83**	2
DM-Knn-Kullback	**76**	2	78	3	75	3	69	2	79	4	61	3	72	2	**85**	3	57	3	65	2	75	3	54	2	70	2	68	3	73	3
HM-Knn-euclidean	80	1	86	1	75	2	**84**	1	**90**	1	78	2	**85**	1	**91**	1	80	3	**85**	1	**91**	1	79	1	**84**	1	**90**	1	79	2
HM-Knn-correlation	79	1	82	2	76	2	83	1	82	2	84	3	83	2	79	3	87	2	81	2	74	3	87	1	79	1	70	3	**88**	2
HM-Knn-cosine	80	1	85	3	76	2	**85**	2	88	1	81	3	**85**	1	85	2	84	1	83	1	80	2	86	1	81	1	73	2	87	3
HM-Knn-Kullback	80	2	85	3	75	2	70	2	80	2	61	4	**85**	1	86	2	84	2	83	1	79	2	87	2	80	1	72	2	**88**	2

4 Conclusion

In this paper, we have presented a benchmark study of 4 alignment free distance functions between sequences, belonging to the classes of *geometric-based* , *correlation-based* based and *information-based* distances. They are computed on their L-tuples representation, for the purpose of sequence classification. Experiments have been carried out on three datasets of nucleosome sequences, using a Knn classifier paradigm incorporating each one of the considered distances. Preliminary results show that the geometric and correlation distances are more reliable than the information-based one. In the near future we plan to extend the experiments using other information-based distances, and other dataset of sequences.

Acknowledgments

Part of this work was carried out using instruments provided by the Euro-Mediterranean Institute of Science and Technology, and funded with the Italian National Operational

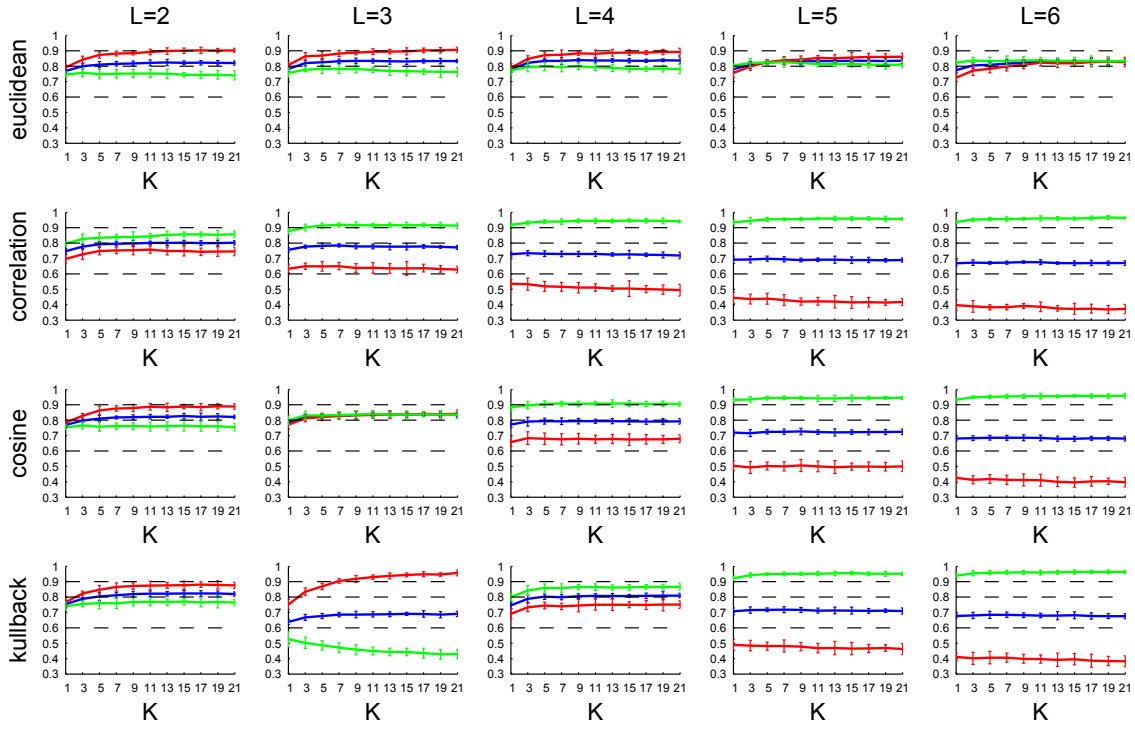


Figure 1: Accuracy (blue), Sensitivity (red) and Specificity (green) plots of the Knn classifiers, for different neighbors K in the range $\{1,.,21\}$, and for different L-tuple lengths L in the range $\{2,.,6\}$ in the case of Caenorhabditis elegans (CE) dataset. Error bars for each K are also reported.

Programme for Research and Competitiveness 2007-2013 grant awarded to the project titled "CyberBrain-Polo di innovazione" (Project code: PONa3_00210, European Regional Development Fund).

References

[1] S.B. Needleman and C.D. Wunsch,"A general method applicable to the search for similarities in the amino acid sequence of two proteins". *J. Mol. Biol.*, vol.48, pp.443453,1970.

[2] T.F.Smith and M.S.Waterman, "Identification of common molecular subsequences". *J. Mol. Biol.*, vol.147, pp.195197, 1981.

[3] O. Gotoh, "An improved algorithm for matching biological sequences". *J. Mol. Biol.*, vol.162, pp.705708, 1982.

[4] S. Altschul, W. Gish, W. Miller et al. "Basic local alignment search tool". *JMol Biol*, vol.25, N.3, pp. 403410, 1990.

[5] D. Lipman, W. Pearson, "Rapid and sensitive protein similarity searches". *Science*, vol.227, N.4693, 1985.

[6] S. Vinga, J. Almeida, "Alignment-free sequence comparisona review". *Bioinformatics*, vol.19, N.4, pp.513523, 2003.

[7] L. Pinello, G. Lo Bosco and G-C. Yuan, "Applications of alignment-free methods in epigenomics", *Briefings in Bioinformatics*, vol.15, N.3, pp.419-430, 2013.

[8] S-H. Guo, E-Z. Deng, L-Q. Xu, H. Ding, H. Lin, W. Chen, K-C. Chou, "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition", *Bioinformatics*, vol.30, n.11, pp.1522-1529, 2014.

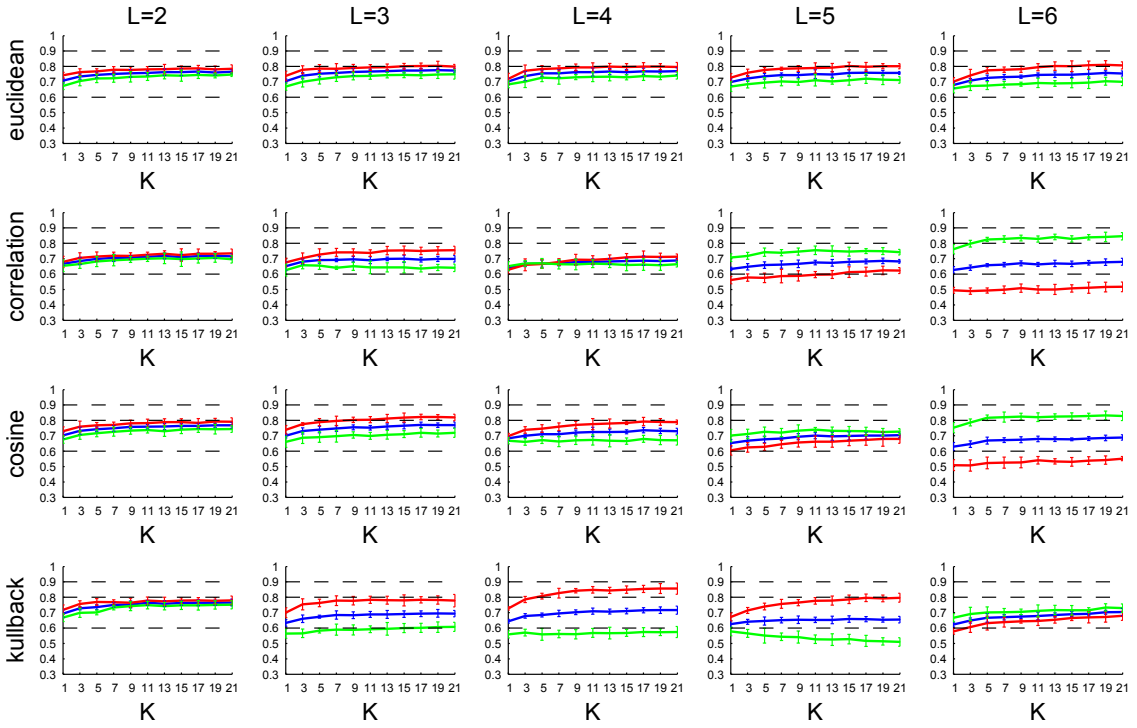


Figure 2: Accuracy (blue), Sensitivity (red) and Specificity (green) plots of the Knn classifiers, for different neighbors K in the range $\{1,.,21\}$, and for different L-tuple lengths L in the range $\{2,.,6\}$ in the case of Drosophila melanogaster (DM) dataset. Error bars for each K are also reported.

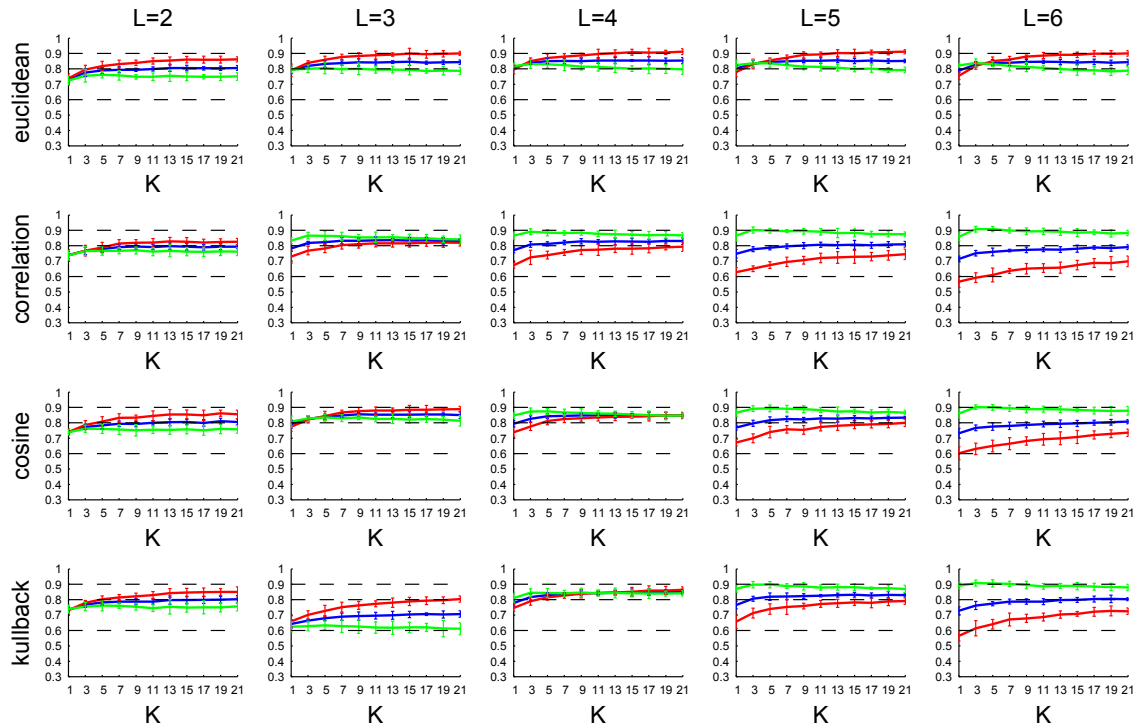


Figure 3: Accuracy (blue), Sensitivity (red) and Specificity (green) plots of the Knn classifiers, for different neighbors K in the range $\{1,.,21\}$, and for different L-tuple lengths L in the range $\{2,.,6\}$ in the case of Homo sapiens (HM) dataset. Error bars for each K are also reported.

Comparison of gene expression signature using rank based statistical inference

Kumar Parijat Tripathi[(1*)], Sonali Gopichand Chavan[(2)], Seetharaman Parashuraman [(2)], Marina Piccirillo[(1)], Sara Magliocca [(1)], Mario R. Guarracino [(1)]

(1) Genomic, Proteomic and Transcriptomic Laboratory, National Research Council of Italy (CNR),Institute for High-Performance Computing and Networking (ICAR) Via Pietro Castellino, 111 80131, Napoli, Italy, kumpar@na.icar.cnr.it

(2) Institute of Protein Biochemistry, National Research Council of Italy (CNR) Via Pietro Castellino, 111 80131, Napoli, Italy

Keywords: expression signature, perturbation, prototype rank list, RRHO, secretory pathway.

Abstract. To understand the unique characteristics of biological state or phenotype, for example, disease or cellular homeostasis, it is of vital importance to understand the behavior of global gene expression. In the field of transcriptomics, gene expression patterns under the corresponding phenotypic state could be used as a proxy to determine the physiological and chemical response from the cellular system in an organism to survive and propagate. Studying these kinds of patterns helps us to understand the response of molecular machinery of cell, to predict the survival or prognosis of an individual to a disease, and also to predict regulation of a particular metabolic pathway. To understand the biological implication of these gene expression signatures is still an open question. In the present work, we are trying to understand the behavior of gene expression signatures of 22 knock down (perturbed) genes involve in secretory pathways in more than 12 different human cancer cell lines, with the help of rank based statistical approach. The aim of our work is to study the consequence of these gene perturbations at the transcriptional level, independently from the specific cell line effects, which would certainly provide an insight into their influence on the biological system. Through comparison of gene expression signature with respect to each perturbation per cell lines, we are able to cluster the knock-down(perturbed) genes, based on their gene expression signatures to understand the combined effect of these perturbations. It helps to understand the cellular mechanism behind a macro-molecular transport system within the cell. Later in our work we also implemented rank−rank hyper-geometric overlap maps (RRHO) for the identification of statistically significant overlapping genes between gene-expression signatures with respect to 22 genes perturbation experiments. Our results show that the transcriptional response with respect to each perturbation does not have independent behavior, but somehow these perturbations put a combinatorial effects on transcriptional regulation. Based on expression signature, these 22 knock-down genes are categorized into 4 clusters and sister perturbation in each cluster have a cumulative role in shaping up the behaviour of cellular system.

1 Scientific Background

The secretory pathway is responsible for the delivery of a large variety of proteins to their proper cellular location and is essential for cellular function and multicellular development. It is composed of a series of compartment that includes endoplasmic reticulum (ER), golgi apparatus, Trans Golgi Network (TGN), through which the cargo (protein or lipid) is transported in an orderly fashion starting from the ER where the bio-synthesis of cargoes is initiated. This is followed by processing of cargoes through

the Golgi apparatus by addition of glycan groups, which are then sorted to their appropriate sites at the Trans Golgi Network [1]. At all these levels, each step, including forward and recycling pathways are controlled by regulatory modules that maintain the homeostasis of the system [2]. With the online availability of huge experimental data, especially the gene expression profiles; it is now possible to extensively predict novel functions and interaction [3, 4]. In our research work, we are keenly interested in studying the pathways and functional components regulated to the secretory pathway. In the present era of high-tech experimental opportunity, expression data provide an easy and large-scale analysis platform for understanding biological mechanism at the cellular level. In the present work, we put our focus on the list of genes, which are localized in a secretory pathway (Figure 1). Our aim is to study the consequence of these gene perturbations at the transcriptional level independently from the specific cell line effects, which would certainly provide an insight into their influence on the biological system. We employed rank based statistical approach to compare their gene expression profile to predict their global effects with respects to pathways and functions, which might be directly or indirectly linked with the gene perturbed and thus the secretory pathway.

2 Materials and Methods

2.1 Data retrieval

The gene expression profile data is a collection of 22 genes perturbation (knockdown) experiments in 12 cancer cell lines at different time points (Figure 1). The data is downloaded from LINCS, NIH program that funds the generation of perturbation profiles across multiple cells and perturbation types(genetic and chemical) (http://www.lincscloud.org/perturbagens/). It comprises of gene expression signature profiles. We used gene expression profiles of both directly measured landmark transcripts plus imputed genes. It has been normalized using invariant set scaling followed by quantile normalization. In these profiles expression data are represented in terms of fold change across the distinctive cell line at different time points.

	Cancer Cell Lines											
Gene perturbation	A375	A549	ASC	HA1E	HCC515	HEPG2	HT29	MCF7	NPC	PC3	SKL	VCAP
ARF1	✓	✓		✓		✓	✓	✓		✓		✓
COG2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COG4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COG7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COPA	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
COPB2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
COPZ1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GOLGA5												✓
M6PR	✓	✓	✓	✓	✓	✓	✓			✓		✓
PLA2G4A	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
RAB1B	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SAR1B	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
SEC24B	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SEC24C	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SEC24D	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
TMED10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TMED7	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
TMED9	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
YKT6	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BLZF1	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
AKAP9	✓	✓	✓	✓		✓	✓	✓		✓		✓
BNPI1	✓	✓		✓	✓	✓	✓	✓		✓		✓

Figure 1: Perturbed genes in corresponding cell lines for which Gene expression profiles are downloaded from LINCS.

2.2 Comparison of gene expression signature using prototype rank list

To obtain gene expression signature and compute distances between pre-processed gene expression profiles for the selected genes, we used "Gene Expression Signature Package" from Bioconductor in R [5]. This package provides the implementation of a methodology to determine the gene expression signature for the perturbation data and calculate distance between them. Gene expression signature is represented as a list of genes whose expression is correlated with a biological state of interest. The distance between the gene expression signature is calculated non parametrically using rank based pattern matching similarity based on Kolmogorov-Smirnov statistic [6]. There are four basic steps are involved;

1. Firstly, with the help of the python script, the gene expression profiles were sorted

according to the differential expression values with respect to the controls, and each gene in the profile is ranked accordingly to its expression value within the sorted list. A matrix is generated, which is composed of ranked list representing the corresponding gene expression profile in each cell line for a given gene perturbation. The graded lists of profiles known as PRLs (prototype ranked list).
2. The PRLs obtained from precise perturbation experiments in all the corresponding cell lines are aggregated by rank merging procedure to negate the effects of specific cell lines. We used built-in "krubor" function of "Gene Expression Signature Package" to carry out rank merging process. It comprises of two sub steps;

- a distance is measured between two ranked list using Spearman's footrule [7] and two or more ranked lists are merged using Borda Merging method.
- a single ranked list is obtained in a hierarchical way using Kruskal algorithm [8].

3. A "signature length" of 250 is taken into account to include 250 most up-regulated genes (near the top of the list) and the most down-regulated genes (near the bottom of the list) for the distance calculation between all the PRLs representing the individual perturbation experiment, with the help of ScorePGSEA and ScoreGSEA functions in "Gene expression signature" in R package. We take into consideration two distance measurements between PRLs;

- Average Enrichment Score Distance $D_{avg} = (TES_{x,y} + TES_{y,x})/2$;
- Maximum Enrichment Score Distance $D_{max} = Min(TES_{x,y}, TES_{y,x})/2$.

4. Affinity propagation clustering (AP) [9] is used to group these PRLs representing different perturbation experiments, inferring distance measures among respective gene expression signatures. AP iteratively searches for optimal clustering by maximizing an objective function called net similarity.

2.3 Rank-rank hypergeometric overlap test analysis

To highlight the correlation strength between two expression profiles, we carried out the rank-rank hypergeometric overlap test analysis using "RRHO package" from Bioconductor in R [10]. This algorithm compares two gene expression profiles ranked by the degree of differential expression. It is used to infer the amount of agreement between two sorted lists (PRL's) by computing the number of overlapping elements in the first $i * stepsize$ and $j * stepsize$ elements of each list, where stepsize represents the number of genes selected from the complete ranked gene list i and j, and return the observed significance of this overlap using a hyper geometric test (Fisher exact test). The output is returned as a list of matrices including: the overlap in the first $i * stepsize, j * stepsize$ elements and the significance of this overlap.

2.4 Gene Set Enrichment Analysis

The PRLs generated were further processed by Gene Set Enrichment Analysis (GSEA) [11] (http://www.broadinstitute.org/gsea/downloads.jsp) using a Molecular Signature Database (MsigDB). GSEA is a computational method that determines whether an initially defined set of genes shows statistical significant differences between two biological states. MsigDB has a collection of annotated gene sets (curated gene set, motif gene set, GO gene set, oncogene signature, immunologic signature, etc.) for use with GSEA software. In order to study the enriched pathways, we used KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway gene set that includes genes contributing to each of the pathways listed in this gene set. Subsequently, using GSEA, enriched KEGG pathways and thus enriched genes, were predicted for all the 22 PRLs and sorted for further analysis.

3 Results

3.1 Distance calculation and clustering of gene expression profiles

In order to calculate the pair wise distances among samples (PRLs representing expression profiles in response to gene perturbation experiments), gene's lists are ranked according to the gene expression ratio (fold change). We defined the "signature length" of 250 to include 250 most up-regulated genes (near the top of the list) and the most down-regulated genes (near the bottom of the list) for the distance calculation. Using scoreGSEA function, we obtained Average Enrichment Score Distance D_{avg} as a measure of pairwise distance between PRLs. The matrix table is generated using pairwise distance between all 22 PRLs versus each other. We also changed the signature length as 500, 800, 1000, and 7000 genes to obtain the best possible average distance and cluster the gene expression profile for these 22 perturbed genes (result not shown).

Taking into account the D_{avg} matrix calculated above, heat map (Figure 2) is generated using hierarchical clustering approach to show the clustering of expression profiles based on their average distances from each other.

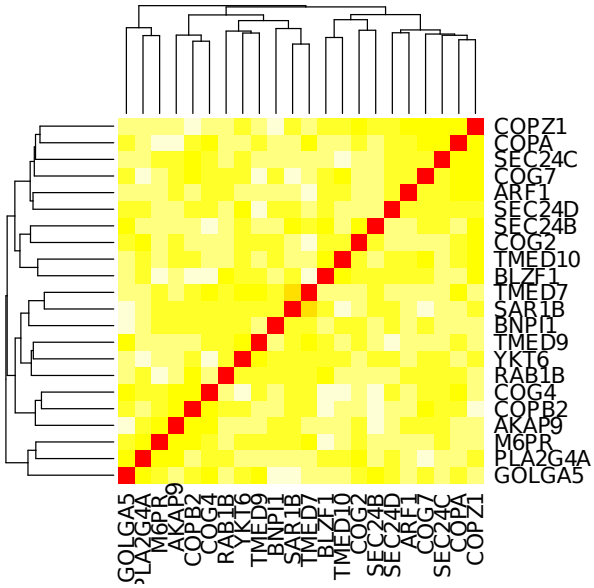


Figure 2: Heatmap based on hierarchical clustering of PRLs.

The closeness of expression profiles, was further verified with the affinity propagation clustering method. It is employed to obtain the exemplar based clusters (Figure 3). Affinity propagation clustering approach determines four clusters based on four exemplars i.e. COPB2, COPZ1, GOLGA5 and SAR1B. The clusters obtained through this approach are in concordance with the hierarchical clustering based approach with some major differences. This is due to because hierarchical clustering starts with every data point as its own cluster and then recursively merges pairs of clusters, but in this way makes hard decisions that can cause it to get stuck in poor solutions. While, on the other hand, affinity propagation is a spectral clustering algorithm that requires each cluster to vote for a good exemplar from within its data points, and thus it provides better clustering of perturbed genes in this study as far as biological inference is concerned. For example, hierarchical clustering shows M6PR closer to GOLGA5 in spite of TMED9, which is biologically closer to GOLGA5 with respect to gene ontology. These results further reveal that though the gene expression signature was obtained independently concerning distinct gene perturbation experiments in different cancer cell lines, however, still they share the same biological connectivity as far as functional regulation of cellular activity is concerned within each respective cluster. For example, in cluster 1, the gene expression profiles obtained by the perturbation of COG4, COG, COPB2, M6PR, AKAP9 and RAB1B genes in separate experiments within different cancer cell

lines, shares the common transcriptional response. It will help us to understand the basic biology in which these genes are involved. It also helps us to address and characterize the biological properties of less-known genes in comparison with its elated sister genes within the cluster, as far as their biological activity and molecular functionality in the cellular system and metabolic pathways are concerned.

APResult object	
Number of samples	22
Number of iterations	152
Input preference	-0.07108233
Sum of similarities	2.661744
Sum of preferences	-0.2843293
Net similarity	2.377415
Number of clusters	4
Cluster 1	COG4, COG7, COPB2, M6PR, AKAP9, RAB1B
Cluster 2	ARF1, BLZF1, COG2, COPA, COPZ1, SEC24C, TMED10
Cluster 3	GOLGA5, PLA2G4A, TMED9
Cluster 4	BNPI1, SAR1B, SEC24B, SEC24D, TMED7, YKT6

Figure 3: Clustering of gene perturbation on the basis of affinity propagation. It generates 4 clusters, each represented by exemplar shown in ***red***

3.2 RRHO analysis

In addition to ”Gene Expression Signature” analysis, we also carried out rank-rank hyper-geometric overlap test to measure the statistical significance of the number of overlapping genes between two expression profiles. Though the gene expression signature is a robust methodology but there is a bottleneck existed in the form of ”signature length”. In the previous analysis, we provide the result based on signature length 250 which means that we only concerned the top 250 and bottom 250 genes in our study. Though it is possible to change this criterion, and we also obtained results by altering the signature length significantly in our work. Nevertheless, previous methodology does not consider the whole 22000 genes in gene expression profile all together while calculating the average distance between profiles. To over come this bottleneck and consider all the genes simultaneously within the study, we used RRHO analysis. In RRHO analysis, we select a window of 2000 (around 10 $\%$) genes from the complete list of genes in a expression profile, which is dynamic in nature and moves throughout all the gene's identifiers listed on the ranked list. The output is returned as a list of matrices including: the overlap in the first $i*stepsize, j*stepsize$ elements and the significance of this overlap. For example, we are representing a case study of cluster 3 previously obtained from ”gene expression signature”. The rank-rank hyper-geometric overlap map (Figure 4) between GOLGA5 vs TMED9 and GOLGA5 vs PLA2G4A show more correlations and statistical significance comparing to PLA2G4A vs SEC24C, suggesting the GOLGA5, TMED9 and PLA2G4A are much closer to each other than the SEC24C. The significance of the overlap between two lists is calculated as function aggregating information from the whole overlap matrix into one summary statistic, typically the min p-value, or max on -log(pval) scale. We also calculated this summary static between each pair of knock down genes.

3.3 Gene set Enrichment analysis

GSEA assigns an enrichment score (ES) along with the false discovery rate (FDR) to each of the enriched predicted pathways. ES (Enrichment Score) is calculated by walking down the ranked list of genes increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not. Based on the ES obtained, an ES matrix was generated. The number of pathways enriched were narrowed down with FDR cut-off of 5 percent. Of the 22 PRLs analyzed; only 8 showed enriched pathways within the 5 percent FDR cut-off (result not shown).

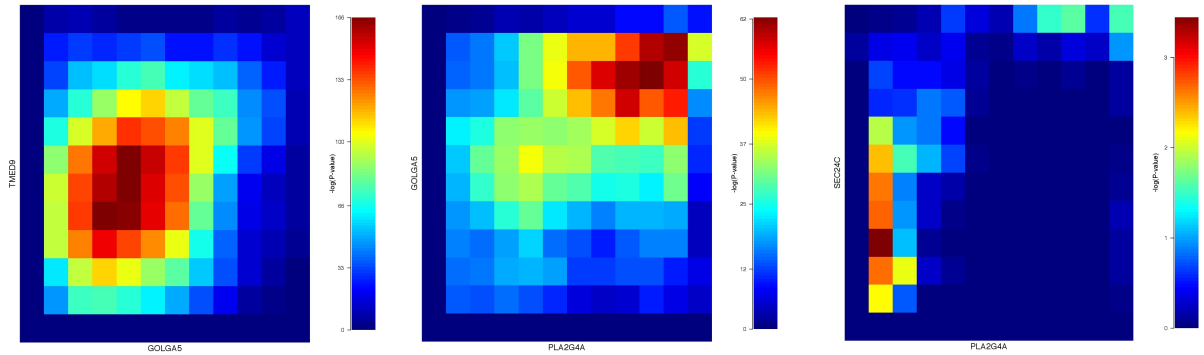


Figure 4: Rank-rank hyper-geometric overlap maps. Comparing expression profile of GOLGA5 vs TMED9 and GOLGA5 vs PLA2G4A (higher correlation); PLA2G4A vs SEC24C (lower correlation).

4 Conclusion

In the present work, we analyze the perturbation profiles of 22 knock-down genes in distinctive cell lines involved in secretory pathways using three different methodology based on non parametric rank statistics. The effect of perturbation across the different cell lines is taken into consideration and based on their differential regulation of genes, a comparison study between their gene expression signature is carried out. The gene expression profiles with respect to each perturbation are checked for their conservation across the cell lines to get an insight into novel functions. The outcome from the methods (PRL based GSEA and Gene's expression signature and RRHO analysis) could then be compared for their accuracy in terms of significant information. We try to understand rank based approach to standardize the appropriate method for studying large-scale perturbation data.

Acknowledgments

We would like to thank the INTEROMICS flagship project, PON02-00612-3461281 and PON02-00619-3470457 for the funding support. Mario Guarracino work is conducted at National Research University Higher School of Economics and supported by RSF grant 14-41-00039.

References

[1] R.B. Kelly. "Pathways of protein secretion in eukaryotes." Science, 230(4721): 25-32, 1985.

[2] A. Luini, G. Mavelli, J. Jung and J. Cancino. "Control systems and coordination protocols of the secretory pathway." F1000prime reports, 6, 2014.

[3] A. Butte. "The use and analysis of microarray data." Nature reviews drug discovery, 1(12): 951-960, 2002.

[4] T. Werner. "Bioinformatics applications for pathway analysis of microarray data." Current opinion in biotechnology, 19(1): 50-54, 2008.

[5] Fei Li, Yang Cao, Lu Han, Xiuliang Cui, Dafei Xie, Shengqi Wang, and Xiaochen Bo. "GeneExpressionSignature: an R package for discovering functional connections using gene expression signatures ". OMICS: A Journal of Integrative Biology, 17(2): 116-118, 2013.

[6] N. Smirnov N . "Table for estimating the goodness of fit of empirical distributions". Annals of Mathematical Statistics, 19: 279281, 1948

[7] Persi Diaconis and R. L. Graham. "Spearman's Footrule as a Measure of Disarray". Journal of the Royal Statistical Society. Series B, Vol. 39, No. 2, pp. 262-268, 1977

[8] J. B. Kruskal . "On the shortest spanning subtree of a graph and the traveling salesman problem". Proceedings of the American Mathematical Society, 7: 4850, 1956

[9] U. Bodenhofer, A. Kothmeier, and S. Hochreiter. "APCluster: an R package for affinity propagation clustering ". Bioinformatics, 27(17):2463 - 2464, 2011.

[10] Plaisier, Seema B., Richard Taschereau, Justin A. Wong, and Thomas G. Graeber. "Rank-rank Hypergeometric Overlap: Identification of Statistically Significant Overlap Between Gene- expression Signatures". Nucleic Acids Research, 38, no. 17,2010.

[11] Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles ". PNAS, vol. 102, no. 43,15545 - 15550, 2005

LAF Barcoding: classifying DNA Barcode multi-locus sequences with feature vectors and supervised approaches

Emanuel Weitschek[(1,2)], Giulia Fiscon[(1,3)], Valerio Cestarelli [(1)], Paola Bertolazzi [(1)], and Giovanni Felici[(1)]

(1) Institute of Systems Analysis and Computer Science, National Research Council, Via dei Taurini 19, 00185, Rome, Italy

(3) Department of Engineering, Uninettuno International University, Corso Vittorio Emanuele II 39, 00186, Rome, Italy

(3) Department of Computer, Control and Management Engineering, Sapienza University, Via Ariosto 25, 00185, Rome, Italy

Emails: emanuel@iasi.cnr.it, fiscon@dis.uniroma1.it, v.cestarelli@gmail.com, paola.bertolazzi@iasi.cnr.it, giovanni.felici@iasi.cnr.it

Keywords: DNA Barcoding, alignment-free, classification, supervised machine learning.

Abstract.

DNA barcodes – one or multiple very short gene sequences – have been proven effective to classify a specimen to species. To handle this task in the plant and fungus kingdoms, multi-locus DNA barcode data as well as sequence analysis techniques are demanded, posing new challenges.

In this work, we describe LAF-BARCODING, a Logic Alignment Free technique that counts the number of fixed-length substrings (k-mers) of the input sequences, represents them in feature vectors, and classifies them through a rule-based approach in order to specifically assign multi-locus DNA barcode sequences to their corresponding species.

We use LAF to classify several sets of DNA barcode sequences, belonging to the plant and fungus life kingdoms, obtaining compact and meaningful classification models (*if-then rules*) with high accuracy rates. Conversely to the widespread alignment-based (e.g., character, tree, and similarity) methods, we highlight that LAF can be successfully applied to multi-locus DNA barcode sequences.

1 Scientific Background

DNA Barcoding is a technique proposed in [1] and used to automatically identify species: short and diverse portions of DNA sequences have been defined as barcodes for plants, animals, and fungi [2]. A DNA barcode enables to distinguish species and identify specimens using one or multiple very short gene sequences. In particular, in the plant and fungus life kingdoms multi-locus barcodes for each individual must be considered in order to obtain reliable classification performances.

The wide spread state-of-the-art methods to analyze DNA barcode sequences are tree-based (that rely on phylogenetic approaches), similarity-based (that use statistical measures), and character-based (that take advantage of string analysis techniques) [10].

Additionally, supervised machine learning techniques have been proposed in several works to approach the DNA barcode specimen to species assignment problem: given a reference library composed of DNA barcode specimen sequences of known species (training set) and a query set of unknown DNA barcode sequences (test set), recognize the latter into the species that are present in the library [9].

In this work, we propose LAF-BARCODING, a new supervised classification approach to identify DNA barcodes particularly suited for multi-locus sequences, i.e., coming from plants and fungi.

2 Materials and Methods

Several DNA barcode analysis methods are based on sequence alignment, i.e., sequence similarity is assessed by aligning portions of sequences that have common nucleotides. However, multi-locus DNA barcode sequences cannot be aligned. Furthermore, the alignment of sequences can be very time consuming, especially for large collections of sequences [8]. Therefore, alignment-free sequence analysis methods can solve the above-mentioned issues. Alignment free methods have been already successfully used for DNA Barcoding by [3, 4]. In greater details, alignment-free algorithms rely on counting and comparing the frequency of all the distinct k-mers that occur in the considered sequences. Such methods map two sequences S_1 and S_2 onto corresponding multidimensional feature vectors $\mathbf{S_1}$ and $\mathbf{S_2}$, which are indexed by a number of subsequences in the given alphabet, e.g., all possible subsequences of a predefined length k (k-mers) are typically used. $\mathbf{S_1}[W]$ and $\mathbf{S_2}[W]$ – the element of $\mathbf{S_1}$ and $\mathbf{S_2}$ associated with the subsequence W – contain the number of occurrences of W in S_1 and S_2, respectively. The occurrences are normalized with respect to the sequences length obtaining the frequency values.

In the following, we describe *Logic Alignment Free* (LAF) [7] and its application to DNA barcode multi-locus sequences. LAF aims to classify barcode sequences and assign them to their species by using a supervised machine learning approach [9]. The method represents the barcode sequences with feature vectors, which are then processed by rule-based classification algorithms [5]. It then calculates the occurrences and the frequencies of the k-mers in the DNA barcodes by scanning a sliding window of length k over the sequence, represents them in feature vectors, and finally extracts a classification model and the specimen to species assignments with rule-based supervised machine learning algorithms. LAF integrates the Jellyfish tool [6] for extracting the feature vectors and adopts the Weka package (`www.cs.waikato.ac.nz/ml/weka/`) for the rule-based classifiers implementations. For further details, the reader may refer to [7]. LAF is available at `dmb.iasi.cnr.it/laf.php`.

In this work, we apply LAF to a large collection of plant and fungus DNA barcode sequences reported in Table 1. The sequences have been downloaded from The BOLD data base (`www.boldsystems.org`).

Table 1: Data sets description.

	n. of samples	**n. of species**
Plants		
Chlorophyta	1915	71
Lycopodiophyta	110	5
Pteridophyta	1618	83
Rhodophyta	17240	267
Fungi		
Ascomycota	39265	1287
Basidiomycota	17954	613
Chytridiomycota	109	3
Glomeromycota	2741	50
Zygomycota	1702	87

3 Results

We test LAF on the above described DNA barcode data, setting the length of the k-mers to 4 and using the RIPPER [5] rule-based classification algorithm. RIPPER is a direct rule extraction method based on a pruning procedure, whose aim is to minimize the error on the training set. An example of extracted rule model is reported in Figure 1.

$$(GTCA \geq 1288.24) \wedge (CCTA \leq 166.94) \Rightarrow Mycena\,pura$$

Figure 1: An example of our rule-based classification model for a fungus species. The feature values refer to the relative frequencies of the k-mers multiplied by 10^5.

We obtain very promising classification results (using 10-fold cross validation) that are shown in Table 2. It is worth noting that we obtain good classification rates also for very large data sets (Rhodophyta, Ascomycota, Basidiomycota).

Table 2: Performances of LAF on plant and fungus data sets.

	Correctly classified [%]	**N. of clauses**	**Precision**	**Recall**	**F-Measure**
Plants					
Chlorophyta	80.83	89	0.81	0.81	0.80
Lycopodiophyta	86.36	5	0.89	0.86	0.87
Pteridophyta	70.83	97	0.69	0.71	0.69
Rhodophyta	83.19	537	0.83	0.83	0.83
avg plants	80.43	151	0.80	0.80	0.80
Fungi					
Ascomycota	77.43	1831	0.80	0.77	0.78
Basidiomycota	71.49	904	0.75	0.72	0.72
Chytridiomycota	98.17	3	0.98	0.98	0.98
Glomeromycota	80.70	92	0.81	0.81	0.81
Zygomycota	75.91	98	0.76	0.76	0.75
avg fungi	78.99	489	0.80	0.79	0.79
avg tot	79.65	335	0.81	0.80	0.79

We highlight that the differences in the number of inferred disjunctive clauses among the analyzed species depend on (i) the number of analyzed sequences (samples), (ii) the number of analyzed species, and (iii) the species divergence (and hence on the complexity of a data set to be classified). For example, in the Ascomycota data set we extract 1831 disjunctive clauses, which model 1287 species (less than 2 clauses per rule).

The execution times of the LAF software tested on the above mentioned data sets and executed on a workstation with an Intel i5 processor with 1.6 GHz clock frequency, 6GB RAM, and 5400rpm hard disk are reported in Table 3.

In order to asses the classification performances of LAF and the rule-based approach, we compared it with respect to other state-of-the-art supervised learning classifiers by giving as input the LAF feature vectors to Support Vector Machines (SVM), Decision Trees (C4.5, Random Forest), and Naive Bayes algorithms. We use the F-measure metric ($\frac{2P \cdot R}{P+R}$) to evaluate the classifiers as it summarizes the other two considered metrics, i.e., recall (R) and precision (P) [9].

Table 3: Execution times of LAF feature vectors extraction and classification with RIPPER [min].

	Feature Extraction	**Classification**	**Total time**
Plants	[min]	[min]	[min]
Chlorophyta	2.0	0.1	2.10
Lycopodiophyta	0.1	0.0	0.10
Pteridophyta	2.0	0.1	2.10
Rhodophyta	27	6.5	33.5
Fungi			
Ascomycota	71	60	131
Basidiomycota	28	15	43.0
Chytridiomycota	0.1	0.0	0.10
Glomeromycota	4.0	0.1	4.10
Zygomycota	2.0	0.1	2.10

Table 4: Performances of the supervised classification methods.

	F-measure			
	SVM	J48	Naive Bayes	Ripper
Plants				
Chlorophyta	0.91	0.86	0.70	0.80
Lycopodiophyta	0.90	0.85	0.81	0.86
Pteridophyta	0.83	0.79	0.72	0.69
Rhodophyta	–	0.83	0.75	0.83
avg plants	0.88	0.83	0.75	0.79
Fungi				
Ascomycota	–	0.81	0.75	0.72
Basidiomycota	–	0.74	0.76	0.72
Chytridiomycota	1.00	0.97	0.98	0.98
Glomeromycota	0.93	0.82	0.72	0.81
Zygomycota	0.83	0.80	0.79	0.79
avg fungi	0.92	0.83	0.80	0.80
avg tot	0.90	0.83	0.78	0.80

Table 4 shows the results of such comparisons confirming the validity of LAF sequences representation combined with all the considered supervised approaches. SVM outperform the other methods, but they do not provide the investigator with a readable model containing the extracted features, and on the largest data sets (i.e.,Rhodophyta, Ascomycota, and Basidiomycota) they do not succeed even after hours of computation.

4 Conclusion

In this work, we presented LAF and its application to DNA barcode sequences belonging to plant and fungus life kingdoms. LAF is an alignment-free method that takes advantage of k-mer feature vectors joint to rule-based classification algorithms for assigning specimen to species by analyzing their multi-locus DNA barcodes. Our analysis resulted in an accurate classification and in the detection of common subsequences (k-mers) in each species of the data sets.

As future directions we propose to further investigate the relationship among the classification performances, models, and number of samples. Additionally, an analysis of the most common k-mers that are shared among species, and a fine tuning of the classifier could be performed. Finally, we propose to test LAF with other classification algorithms, on additional DNA barcode multi-locus sequences, and to compare the results with other methods for analyzing multilocus sequences.

Acknowledgments

The authors have been supported by the FLAGSHIP "InterOmics"(PB.P05) and "EPIGEN" projects, and by the Italian PRIN "GenData 2020" (2010RTFWBH).

References

[1] Paul DN Hebert, Sujeevan Ratnasingham, and Jeremy R de Waard. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 1):S96–S99, 2003.

[2] Peter M Hollingsworth, Laura L Forrest, John L Spouge, Mehrdad Hajibabaei, Sujeevan Ratnasingham, Michelle van der Bank, Mark W Chase, Robyn S Cowan, David L Erickson, Aron J Fazekas, et al. A dna barcode for land plants. *Proceedings of the National Academy of Sciences*, 106(31):12794–12797, 2009.

[3] Pavel Kuksa and Vladimir Pavlovic. Efficient alignment-free dna barcode analytics. *BMC Bioinformatics*, 10(Suppl 14):S9, 2009.

[4] Massimo La Rosa, Antonino Fiannaca, Riccardo Rizzo, and Alfonso Urso. Alignment-free analysis of barcode sequences by means of compression-based methods. *BMC Bioinformatics*, 14(Suppl 7):S4, 2013.

[5] Thorsten Lehr, Jing Yuan, Dirk Zeumer, Supriya Jayadev, and Marylyn D Ritchie. Rule based classifier for the analysis of gene-gene and gene-environment interactions in genetic association studies. *BioData Mining*, 4(1):4, 2011.

[6] Guillaume Marcais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, Mar 2011.

[7] Dimitris Polychronopoulos, Emanuel Weitschek, Slavica Dimitrieva, Philipp Bucher, Giovanni Felici, and Yannis Almirantis. Classification of selectively constrained dna elements using feature vectors and rule-based classifiers. *Genomics*, 104(2):79–86, 2014.

[8] Susana Vinga and Jonas Almeida. Alignment-free sequence comparisona review. *Bioinformatics*, 19(4):513–523, 2003.

[9] Emanuel Weitschek, Giulia Fiscon, and Giovanni Felici. Supervised dna barcodes species classification: analysis, comparisons and results. *BioData Mining*, 7(1), 2014.

[10] Emanuel Weitschek, Robin Velzen, Giovanni Felici, and Paola Bertolazzi. Blog 2.0: a software system for character-based species classification with dna barcode sequences. what it does, how to use it. *Molecular Ecology Resources*, 13(6):1043–1046, 2013.

Special session on

The EDGE, enhanced definition of genomic entities for systems biomedicine in oncology

Organisers

Prof. Elia Mario Biganzoli (Università degli Studi di Milano)
Prof. Maria Clelia Di Serio (University Vita-Salute San Raffaele, Italy)

This Session has been sponsored by

Department of Clinical Sciences and Community Health, University of Milan

Transcriptomic analysis of the Trop-2 metastatic program

Emanuela Guerra[(1)], Rossano Lattanzio[(1)], Marco Trerotola[(1)], Pasquale Simeone[(1)], Valeria Relli[(1)], Patrizia Querzoli[(2)], Enzo Bianchini [(3)], Domenico Angelucci[(4)], Giuseppe Pizzicannella[(5)], Laura Antolini[(6)] , Andrea Telatin[(7)], Barbara Simionati[(7)], Mauro Piantelli[(1)] and Saverio Alberti[(1,8)]

(1) Unit of Cancer Pathology, CeSI, Foundation University G. dAnnunzio,
Via Colle dell Ara, 66100 Chieti Scalo, Italy.

(2) Institute of Pathology, University of Ferrara,
Via Luigi Borsari 46, 44121 Ferrara, Italy.

(3) Department of Pathology, Ferrara Hospital,
Viale Tre Martiri 140, 45100 Ferrara, Italy.

(4) Department of Pathology,
Chieti Hospital, 66100 Chieti, Italy.

(5) Department of Pathology,
ASL2 Chieti-Lanciano-Vasto, 66100 Chieti, Italy.

(6) Biostatistics Center, Department of Clinical Medicine, Prevention and Biotechnology,
University of Milano-Bicocca, 20900 Monza, Italy

(7) BMR Genomics srl,
Via Redipuglia, 22 - 35131 Padova, Italy.

(8) Department of Neuroscience, Imaging and Clinical Sciences, Unit of Physiology and Physiopathology, University G. dAnnunzio,
Via Colle dell Ara, 66100 Chieti Scalo, Italy.

Keywords:

Abstract. Tackling metastatic disease is key to improve cancer patients survival [1]. Thus, better knowledge of metastatic disease-driving mechanisms is urgently needed. Hundreds of genes have been linked to the metastatic phenotype [2-5]. However, no unique marker of cancer aggressiveness and metastatic spreading has been identified. Trop-2 is a transmembrane signal transducer [6-10], which drives the renewal of normal and neoplastic stem cells [11,12] and stimulates cancer growth [6,13,14], in quantitative relationship to its expression levels [6,15]. Metastatization requires cellular functions that are additional to those that initiate tumors and drive transformed cell growth progression [3,16]. Hence, to search for genes instrumental to metastatic diffusion, we looked for genes concordantly disregulated across distinct cancer metastatic models. This led us to discover that TROP2 is the only gene upregulated in metastatic cells across distinct experimental models. .

1 Metastatic transcriptome profiling

Metastasis-associated genes are expected to include primary drivers of the metastatic phenotype, but also secondary events, together with adaptive, counterbalancing changes [17]. To distinguish among transcript subclasses with distinct function, we thus profiled

the transcriptomes of distinct colon cancer metastatic systems [18], with the goal to identify concordantly dysregulated genes. Analysis of independent data-sets showed that most expression changes were mutually exclusive, thus illustrating the unexpected finding that transcriptomic scenarios are profoundly different even in parallel colon cancer models. Only two genes, TROP2/TACSTD2, a determinant of stem cells renewal [11,12], that drives tumor growth [6,14], and vimentin, a gene associated with epithelial-mesenchymal transition [4], were concordantly upregulated. We then went on to verify these findings through a meta-analysis of transcriptome data from rat pancreatic carcinoma cell lines [19], murine prostate [20] and murine breast [21]. The breast cancer cells were driven by the fms/CSF1R oncogene, which is abnormally expressed in breast cancer. Upregulation CSF1R into mammary epithelial cells renders the transfectants capable of tumorigenesis and local invasion. Remarkably, this was associated with upregulation of the TROP2 gene [21]. TROP2/CSF1R cells injected intravenously in mice, produced more lung metastases than parental cells [21]. Notably, corresponding clinical significance was shown by analysis of invasive breast carcinoma cells from patients with invasive ductal carcinoma cells, which were shown to require high abundance of CSF-1 paracrine signaling for efficient metastasis formation in all clinical breast cancer subtypes. This closely reflected in human disease, as triple negative breast cancer subtypes were shown to express high CSF-1R and additive autocrine CSF-1/CSF-1R signaling, as required for invasion.

2 Trop-2 associates with human cancer metastatization

We verified the relevance of these findings in human cancer. A first case series of colon cancers and matching metastases indicated marked upregulation of TROP2 in most metastatic lesions. These findings were then extended to breast, uterus, colon and ovary cancers. Immunohistochemistry (IHC) analysis of independent validation case-series of breast, uterus, colon and ovary metastatic cancers revealed that Trop-2 was overexpressed in matched metastases in all tumor types. Higher fractions of metastatic IHC-positive cells were found in 79% for breast, 85% for stomach, and 88% for colon carcinomas. Hence, Trop-2 appeared as a hallmark of metastatic cells in most human cancers.

3 Trop-2 drives metastatic spreading

Trop-2 had been indicated to facilitate colonization by prostate cancer cells injected in the bloodstream [22]. However, no proof was obtained for a capacity of Trop-2 to induce distant metastatization. Hence, we assessed this capacity in an orthotopic metastatic model [18]. Colon cancer transfected with a wild-type Trop-2 (wtTrop-2) were injected in the spleen of nude mice and metastatization to the liver was assessed. wtTrop-2 overexpression was found to increase the metastatic diffusion of colon cancer cells. Histopathology analyses indicated that wtTrop-2 is a driver of malignant progression. Control colon cancer tumors grew as manifold nodules, with central necrosis and peripheral fibrous capsule. On the other hand, wtTrop-2-expressing tumors showed much reduced differentiation, and loss of central apoptosis, together with thinner perinodular tumor capsule and pseudo-capsule, consistent with more invasive growth.

The cytoplasmic tail of Trop-2 is cleaved-off 11, during proteolytic processing for Trop-2 activation [8,11]. Hence, a Trop-2 mutant devoid of cytoplasmic tail was designed, to mimic constitutively-active Trop-2. The potential impact of activated Trop-2 on metastatization was assessed. Activated Trop-2-expressing tumors showed highly invasive growth patterns, with tumor budding across layers of hepatocytes. Cells acquired mesenchimal morphology, with much diminished signs of compressive growth and of pseudo-capsule formation.

4 Next-generation transcriptomic analysis

To obtain a faithful portrait of Trop-2-induced changes, differential mapping of the transcriptome of the tumor (of human origin) and of stroma (of murine origin) of colon cancer models was performed. To this end, a first step, was to quantify transcripts according to species origin. In this context, the "stroma" was expected to demonstrate features of spleen lymphoid cells, of hematopoietic resting or activated endothelium and connective tissue. The metastatic "stroma" was expected to stem from hepatocytes and bile ducts, with additional components as for the spleen.

Planned comparisons between samples included: A. Data obtained through technical replicates, i.e. independent sequencing runs from independent libraries from the same mRNA sources. This will allow to measure technical noise, as for sample preparation or from sequencing procedures bias. B. Biological replicates, i.e. tumors or metastases obtained in different experiments / mice inoculated with the same cells. These include: biological replicates of tumors spleen of control cells biological replicates of tumors of the spleen cells transfected with Trop-2 biological replicates of liver metastases of cells transfected with Trop-2 biological replicates of tumors of the spleen cells transfected with constitutive activated Trop-2 biological replicates of liver metastases of cells transfected with constitutive activated Trop-2. Analysis of biological replicates will allow to quantify deviation from the theoretical model of absolute identity, to quantify reproducibility / spreading of data measurements. This will also provide a measure of variability of the mode of growth / cell interaction analyzed with the environment. C. Comparisons of biologically distinct samples will allow to profile the effect of Trop-2 on the transcriptome of primary tumors: - B (i) = tumors spleen of control cells vs B (ii) = tumors of spleen cells transfected with Trop-2. - metastasis to the liver of control cells vs B (iii) = liver metastases of cells transfected with Trop-2 will measure the 'effect of Trop-2 on the transcriptome of liver metastases - B (i) = tumors spleen of control cells vs metastasis to the liver of control cells will identify changes in the transcriptome of liver metastases compared to primary tumors. - B (ii) = tumors of spleen cells transfected with Trop-2 vs B (iii) = liver metastases of cells transfected with Trop-2 will identify changes in the transcriptome of liver metastases compared to primary tumors expressing Trop-2. Independent comparisons will be performed with mutants vs Trop-2 wild type, to identify changes in the transcriptome of primary tumors / metastases, associated with mutations of Trop-2, which alter the processing or signaling capacity.

5 Perspectives

Trop-2 metastatic upregulation was verified in breast, colon, stomach or ovary cancer patients. Trop-2 was then shown to potently drive metastatization in vivo. These findings for the first time identify Trop-2 as a master driver of metastatization. The Trop-2-driven module was dissected through next-generation sequencing/transcriptome profiling. This is allowing to explore shared features of cancer spreading in preclinical cancer models. These features will be then challenged in corresponding human cancer type case series.

References

[1] De Vita, V. T., Lawrence, T. S., Rosenberg, S. A. De Vita, Hellman, Rosenberg's Cancer: Principles & Practice of Oncology, 8th Edition. 8th edn, (Lippincott Williams & Wilkins), 2008.

[2] Hanahan, D. and Weinberg, R. A. Hallmarks of cancer: the next generation. Cell 144, 646-674, 2011.

[3] Nguyen, D. X., Bos, P. D. and Massague, J. Metastasis: from dissemination to organ-specific colonization. Nat. Rev. Cancer 9, 274-284, 2009.

[4] Polyak, K. and Weinberg, R. A. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. Nat Rev Cancer, 2009.

[5] Vanharanta, S. and Massague, J. Origins of metastatic traits. Cancer Cell 24, 410-421, 2013.

[6] Trerotola, M. et al. Up-regulation of Trop-2 quantitatively stimulates human cancer growth. Oncogene 32 222-233, 2013.

[7] El Sewedy, T., Fornaro, M. and Alberti, S. Cloning of the murine Trop2 gene: conservation of a PIP2-binding sequence in the cytoplasmic domain of Trop-2. Int J Cancer 75, 324-330, 1998.

[8] Alberti, S. et al. Biochemical characterization of Trop-2, a cell surface molecule expressed by human carcinomas: formal proof that the monoclonal antibodies T16 and MOv-16 recognize Trop-2. Hybridoma 11, 539-535, 1992.

[9] Fornaro, M. et al. Cloning of the gene encoding TROP-2, a cell-surface glycoprotein expressed by human carcinomas. Int. J. Cancer 62, 610-618, 1995.

[10] Ripani, E., Sacchetti, A., Corda, D. and Alberti, S. The human Trop-2 is a tumor-associated calcium signal transducer. Int. J. Cancer 76, 671-676, 1998.

[11] Stoyanova, T. et al. Regulated proteolysis of Trop2 drives epithelial hyperplasia and stem cell self-renewal via beta-catenin signaling. Genes Dev 26, 2271-2285, 2012.

[12] Trerotola, M. et al. CD133, Trop-2 and alpha2beta1 integrin surface receptors as markers of putative human prostate cancer stem cells. Am J Transl Res 2, 135-144, 2010.

[13] Wang, J., Day, R., Dong, Y., Weintraub, S. J. and Michel, L. Identification of Trop-2 as an oncogene and an attractive therapeutic target in colon cancers. Mol Cancer Ther 7, 280-285, 2008.

[14] Guerra, E. et al. A bi-cistronic CYCLIN D1-TROP2 mRNA chimera demonstrates a novel oncogenic mechanism in human cancer. Cancer Res 68, 8113-8121, 2008.

[15] Guerra, E. et al. The Trop-2 signalling network in cancer growth. Oncogene 32, 1594-1600, 2013.

[16] Malanchi, I. et al. Interactions between cancer stem cells and their niche govern metastatic colonization. Nature, 2011.

[17] Amit, I. et al. A module of negative feedback regulators defines growth factor signaling. Nat Genet 39, 503-512, 2007.

[18] Morikawa, K. et al. Influence of organ environment on the growth, selection, and metastasis of human colon carcinoma cells in nude mice. Cancer Res 48, 6863-6871, 1988.

[19] Tarbe, N., Losch, S., Burtscher, H., Jarsch, M. and Weidle, U. H. Identification of rat pancreatic carcinoma genes associated with lymphogenous metastasis. Anticancer Res 22, 2015-2027, 2002.

[20] Calvo, A. et al. Alterations in Gene Expression Profiles during Prostate Cancer Progression: Functional Correlations to Tumorigenicity and Down-Regulation of Selenoprotein-P in Mouse and Human Tumors. Cancer Res 62, 5325-5335. (2002).

[21] Kluger, H. M. et al. cDNA microarray analysis of invasive and tumorigenic phenotypes in a breast cancer model. Lab Invest 84, 320-331 2004.

[22] Trerotola, M. et al. Trop-2 promotes prostate cancer metastasis by modulating 1 integrin functions. Cancer Res. 73, 3155-3167, 2013.

Cancer Images: from invading hordes to pseudo-organ structures.

Romano Demicheli[(1)],

(1) Department of Medical Oncology,
Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Istituto Nazionale Tumori, via Venezian 1, Milano 20133, Italy

Keywords:

Abstract. The early perception of cancer was substantially borrowed from the paradigms of bacterial infections: cancer was regarded as pathologic phenomenon occurring at cellular level, a genome-driven disease, where the accumulation of a sufficient number of alterations in key genes results in cell transformation [1]. Transformed cells were viewed as aliens intruding a vulnerable idle microenvironment. Cell transformation was believed to be an irreversible process: Once a cancer cell, always a cancer cell. Yet, this paradigm was progressively challenged by a number of experimental and clinical findings, highlighting the crucial role of tumour stroma and providing evidence that normal cells may display cancer-like behaviour while, conversely, cancer cells may regain normal cell traits [2-5]. The novel cancer image, where tumours look like pseudo-organ structures more than invading hordes, supported reasonable explanations for clinical findings and suggested new concepts such as tumour dormancy and accelerated metastasis growth due to primary tumour removal [6]. It also advocated the occurrence of some kind of homeostatic effect of primary tumour upon distant metastases, apparently mimicking the organ homeostasis that succeeds the growth process. Seminal knowledge on tissue homeostasis emerged from investigations on damage repair, where dramatic reawakening of the tissue building machinery is required. Both parallel processes between wound healing and morphogenesis [7] and the role of growth factors and cytokines [8] have been recognized, as well as the role of bone marrow (BM), which provides inflammatory mature cells to injured tissues. Furthermore, a number of recent reports, in both animals and humans, indicate that bone marrow also supplies cells capable of producing non-hematopoietic tissue. Likewise, non-haematological tumour cells may derive from BM cells. For instance, in recipients of sex mismatched BM, peripheral blood stem cell or organ transplantation, developing a successive solid cancer, BM derived cells (DCs) were found to contribute to tumour cells. Furthermore, BMDCs may develop as a constituent of tumour cell populations, generating, in some cases, the entire tumour mass, the same as they may accomplish in wound healing. Tumour stroma may derive from bone marrow, as well. Therefore, we may acknowledge that BMDCs may play in tumours and in wound healing similar supportive roles in the form of stroma cells that care the local parenchymal cell population by reacting to microenvironmental signals. A further support to the cross-talk and cooperation between tumour and BM is the recent discovery of the so called pre-metastatic niche. It was documented that before the arrival of tumour cells, adjustments occur in metastatic sites that make them conducive for successive metastasis development. This cellular bookmarking was first reported for lung metastases from LLC and B16 melanoma in mice [9]. Intradermal injection of LLC cells resulted in BMDC cluster formation limited to the lung and liver with no clusters in other organs. In contrast, B16 melanoma tumour cells induced the formation of BMDC clusters in multiple tissues such as lung, liver, testis, spleen and kidney. Remarkably, pre-treatment with melanoma derived conditioned medium resulted in redirection of LLC metastasis to sites frequently observed

in B16-melanoma. In humans, the process of vascularisation in the metastatic versus non-metastatic versus non-cancerous inflamed axillary lymph nodes was consistent with findings in the animal model. The collective action underlying all these processes implies that cellular performers communicate with other participants. Usually it was believed that soluble factors such as cytokines, chemokines, growth factors and bioactive lipids released from a given cell type and circulating through the whole organism are able to induce responses by other cells endowed with specific receptors. Recent research, however, is elucidating a much more complex and efficient communication system, the core of which is a busy trafficking of microvesicles. Microvesicles (MVs), for long time considered cellular debris, have been recently recognized as functionally relevant [10]. MVs are spherical membrane fragments containing a cargo of cytosol including a distinct and definite combination of lipids, proteins and nucleic acids (mRNA, miRNA and DNA), i.e. a non-random sample of the molecular repertoire of the originating cell [11]. MV surfaces express the adhesion molecules of the cell of origin, allowing specific capture by target cells that recognize them, which may be modified by surface interaction. Most important, MVs may induce epigenetic changes in target cells by transferring selected arrays of mRNA and miRNA associated with ribonucleoproteins. The above reported findings indicate that BM routinely provides a contribution of specific parenchymal cells to various tissues especially after tissue damage. Remarkably, the homing of BMDCs into the damaged tissue is associated, to some extent, with the emergence of gene expression patterns corresponding to phenotypes of stroma and parenchymal cells of the invaded tissue (e.g. myofibroblasts and keratinocytes in the skin). This phenotypic change apparently relies on dominant effects of the tissue microenvironment upon imported cells. This concept is strongly supported by focused investigations. There is, therefore, a growing body of evidence that homeostasis involves both local factors and cooperation of distant partners, i.e. the whole organism. Progenitor cells with different commitment (e.g., HSCs, MSCs and ESCs) in addition to pluripotent cells (e.g., VSELS) may reside in virtually all organs, among which BM is a main reservoir. When homeostasis alterations occur, signalling pathways may activate and, if needed, mobilize them. Activated progenitor or pluripotent cells (PPCs) secrete a variety of growth factors, cytokines, chemokines and bioactive lipids that regulate their biology and orchestrate interactions with the surrounding microenvironment. In addition to soluble factors, activated PPCs also secrete MVs, conveying packaged signalling factors, including genetic information, which may change the phenotype of the target cells, locally or at distance. Thus, one can conceive that, while cellular populations may be relatively stable, transcriptional regulation, a key determinant of the phenotype of a particular cell, may shift between different cell types [12]. What remains stable and, if damaged, induces cellular conversions to re-achieve the original condition is apparently the whole tissue architecture. It should be emphasized that all sub-components of both normal and tumour tissue act as a team whose members cooperate in space and time. For instance, in prostatic carcinoma [13], cancer associated fibroblasts (CAFs) elaborate active factors, which recruit monocytes toward tumour cells. SDF-1delivery promotes their trans-differentiation toward the tumour associated macrophages (TAM) phenotype. The relationship between TAMs and CAFs is reciprocal. TAMs affect fibroblasts enhanced reactivity. On the other side, cancer cells themselves participate in this cross talk through secretion of monocyte chemotactic protein-1, facilitating monocyte recruitment as well as macrophage differentiation and M2 polarization. The role of stroma cells may be central: the microenvironment (cells and ECM) promotes tumorigenesis independent of initiating genetic events in tumour cells. The stroma itself is a crucial target for cancerogenesis, as it can induce adenocarcinoma development among apparently fully normal resident epithelial cells [14]. Taken together, these experimental findings suggest that ontogenetic processes, by which cells of different embryonic

lineages actively induce each other to cooperate in the formation of tissues, organs and body regions, are ongoing within tumours. The fact that the behaviour of each cell (i.e. the phenotype) is regulated by other cells and by additional ECM factors within a complex network of signals implies that cells cannot change their activity freely all by themselves, because of the multiple regulatory interactions within which they live. Therefore, only certain tissue architectures are permitted (i.e., are stable). This architectural constraint recalls what occurs within nucleus, where genes are subject to mutual actions and their global architecture may be studied by the Gene Regulatory Network approach [15,16]. In this approach, usefully adopted, to a limited extent, in analyses of complex dynamic systems, the global architecture of the network creates a landscape of mutually exclusive attractors (stable states) surrounded by their basin of attraction (towards the stable state) and separated by areas of unstable states, at a particular instant in time. At tissue level, attractors correspond to distinct normal tissue dynamic architectures, while unstable states are transiently occupied during critical bifurcations, as during embryo development, wound healing or cancer initiation or spread. According to this conceptual approach, cancers are emerging self-stabilizing tissue states (i.e. cancer attractors) where cells/tissues display characteristic traits of tumor phenotype (proliferation, migration, invasion, angiogenesis etc.), which become accessible as a consequence appropriate perturbing factors. According to this conceptual approach, cancers are emerging self-stabilizing tissue states (i.e. cancer attractors) where cells/tissues display characteristic traits of tumor phenotype (proliferation, migration, invasion, angiogenesis etc.), which become accessible as a consequence appropriate perturbing factors. As recalled, cancer researchers have focused merely upon epithelial cells displaying morphological and genetic abnormalities (epithelial cancer cells), which were considered the origin of all neoplastic disease (a cancer cell-cantered explanation). Likewise, profound scientific and commercial success of molecular biology, the progress of cancer gene investigation technologies, together, pushed forward the uncritically accepted postulate that genes explain everything. This static mono-dimensional explanation diffused and strengthened classical Somatic Mutation Theory (SMT), as the sole acceptable explanation for cancer [17]. However, this epithelio-centric somatic mutational view of tumorigenesis is clearly in conflict with most findings and specifically with those proving the modulatory role of stroma in tumorigenesis. In particular, it is now clear that the stroma may be a crucial target of the carcinogen [14]. The irreversibility of the neoplastic phenotype is no more sustainable as it is conflicts strongly with experimental, as well as clinical and epidemiologic findings. In Laminin-1 rich basement membrane extracts, tumorigenic cells, usually growing in disorganized masses, may display a reorganization of both intracellular and intercellular structures under the action of targeted signalling modifications, resulting in phenotypic changes and acini-like polarized organization [18]. Moreover, in a series of elegant experiments with rodents [2], where coat color and isoenzyme analysis were used as markers, it was demonstrated that Embryonic Carcinoma cells, which form malignant tumors upon subcutaneous injection, contribute to the development of chimerical normally developed mice if injected into a blastocyst. It is truly astonishing the fact that these seminal findings, where cells bearing neoplastic genome display perfectly normal behaviour when put into a normal neighbourhood have been ignored by the cancer cell community. More recently, new paradigms were proposed that provide appealing explanations to known findings. Their main communal trait is the shifting of the pathological process origin from the cell level to the tissue level [19,20]. These novel approaches view cancers as developments gone awry or maladjusted living entities with parasitic properties. A similar perception of tumours as organ-like structures emerged from clinical investigations as well [21]. In summary, tissue-based cancer explanations assume that, while structural lesions in the DNA coding sequences, or epigenetic disorders in control of gene expression, undoubtedly have important contributory roles in

cancer, disturbed tissue interactions among cell populations in a given area are critical to cancer causation, growth and spread. They are a communal cancer cause, shifting the cancer tissue from steady-state cell kinetics to progressive malignant behaviour, to the materialization of a cancer attractor.

References

[1] Hanahan D, Weinberg RA: The hallmarks of cancer. Cell 100, 57-70, 2000.

[2] Mintz B, Illmensee K: Normal genetically mosaic mice produced from malignant teratocarcinoma cells. Proc Natl Acad Sci U S A. 72, 3585-3589, 1975.

[3] Maffin MV, Soto AM, Calabro JM, Ucci AA, Sonnenschein C: The stroma as a crucial target in rat mammary gland carcinogenesis. J Cell Sci. 117, 1495-1502, 2004.

[4] Bissell MJ, Labarge MA: Context, tissue plasticity, and cancer: are tumour stem cells also regulated by the microenvironment? Cancer Cell. 7, 17-23, 2005.

[5] Podsypanina K et al: Seeding and propagation of untransformed mouse mammary cells in the lung. Science 321, 1841-1844, 2006.

[6] Demicheli R, Retsky MW, Hrushesky WJ, Baum M: Tumour dormancy and surgery-driven interruption of dormancy in breast cancer: learning from failures. Nat Clin Pract Oncol 4, 699-710, 2007.

[7] Martin, P., Parkhurst, S.M. Parallels between tissue repair and embryo morphogenesis. Development 131, 3021-3034, 2004.

[8] Werner, S., Grose, R. Regulation of wound healing by growth factors and cytokines. Physiol Rev 83, 835870, 2003.

[9] Kaplan, R.N. et al. VEGFR1-positive haematopoietic bone marrow progenitors initiate the premetastatic niche. Nature 438, 820-827, 2005.

[10] Ratajczak, J., Wysoczynski, M., Hayek, F., Janowska-Wieczorek, A., Ratajczak, M.Z. Membrane-derived microvesicles: important and underappreciated mediators of cell-to-cell communication. Leukemia 20, 14871495, 2006

[11] Valadi H, Ekstrm K, Bossios A, Sjstrand M, Lee JJ, Ltvall JO (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol 9, 654659, 2007.

[12] Theise, N.D., Krause, D.S. Suggestions for a new paradigm of cell differentiation potential. Blood Cells Mol Dis 27, 625-631, 2001.

[13] Comito G, Giannoni E, Segura CP, Barcellos-de-Souza P, Raspollini MR, Baroni G, Lanciotti M, Serni S, Chiarugi P: Cancer-associated fibroblasts and M2-polarized macrophages synergize during prostate carcinoma progression. Oncogene, 2013

[14] Maffini MV, Soto AM, Calabro JM, Ucci AA, Sonnenschein C: The stroma as a crucial target in rat mammary gland carcinogenesis. J Cell Sci;117:1495-502, 2004.

[15] Huang S, Ingber DE: A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks. Breast Disease; 26:27-54, 2006

[16] Demicheli R, Coradini D: Gene regulatory networks: a new conceptual framework to analyse breast cancer behaviour. Ann Oncol; 22:1259-1265, 2011.

[17] Hanahan D, Weinberg RA: Hallmarks of cancer: the next generation. Cell.; 144:646674, 2011.

[18] Weaver VM, Petersen OW, Wang F, Larabell CA, Briand P, Damsky C, Bissell MJ: Reversion of the malignant phenotype of human breast cells in three-dimensional culture and in vivo by integrin blocking antibodies. J Cell Biol.;137:231-245, 1997.

[19] Sonnenschein C, Soto AM: Theories of carcinogenesis: an emerging perspective. Sem Cancer Biol;18:372-377, 2008.

[20] Tarin D: Cell and tissue interactions in carcinogenesis and metastasis and their clinical significate. Sem Cancer Biol;21:72-82, 2011.

[21] Demicheli R, Retsky MW, Hrushesky WJM, Baum M: Tumor dormancy and surgery-driven dormancy interruption in breast cancer: learning from failures. Nature Clin Pract Oncol; 4:699-710, 2007.

Dissecting the biological complexity of breast cancer

Christine Desmedt[(1)]

(1) Breast Cancer Translational Research Lab
Institut Jules Bordet, Universit Libre de Bruxelles, Brussels, Belgium
christine.desmedt@bordet.be

Keywords:

Abstract. From the clinical point of view, there is evidence that breast cancer patients can have different clinical presentations and evolutions of the disease, which cannot always be explained by the standard clinical and pathological parameters, such as for example age at diagnosis, tumor size, histological grade and axillary lymph nodes involvement. It is therefore essential to better understand the biology of the disease. During my presentation, I will present, in a non-exhaustive way, some elements which contribute to the biological complexity of breast cancer.

1 Histological diversity of breast cancer:

Ductal breast cancer has been the most widely studied histological subtype since it is the most frequent one. Lobular breast cancer represents the second most common histological subtype after ductal breast cancer, and accounts for 5 to 15% of all invasive breast cancers. From a clinical point of view, lobular tumors are generally associated with an indolent but progressive clinical behavior and tend to exhibit a peculiar metastatic behavior compared to ductal breast cancers (1). Although patients affected by lobular breast cancer have lower response rates to conventional chemotherapeutic agents, recent results suggest that they might show an increased benefit from a certain type of endocrine treatment, the aromatase inhibitors (1). Nevertheless, despite these histological and clinical differences, no specific treatment recommendation exists for lobular breast cancer, due to the relative paucity of the research dedicated to this disease. In recent study, we identified recurrent, cancer-specific genomic alterations with potential implications in the clinical and pathological determinants of lobular tumors. We for example identified a higher frequency of tumors carrying oncogenic AKT1, ERBB2 and ERBB3 mutations in lobular compared to ductal breast cancer (2). Given the existence of drugs targeting these alterations, this works suggests that lobular tumors should be interrogated for the presence of these mutations in order to start individualizing the treatment of this disease.

2 Importance of the tumor microenvironment:

The tumor microenvironment is defined as the cellular environment in which the tumor exists, including surrounding blood vessels, immune cells, fibroblasts, other cells, signaling molecules, and the extracellular matrix (3). The tumor and the surrounding microenvironment are closely related and interact constantly. Several gene expression studies investigated the interaction between these cells with the tumor epithelial cells and their role with regard to prognosis and prediction of anti-cancer treatment efficacy. High immune scores have been associated for example with good prognosis and response to neo-adjuvant chemotherapy, especially in triple-negative and HER2-positive breast tumors (4). What is now important is to understand how to best quantify these immune infiltrates and to understand their composition and their exact role in breast cancer.

3 Inter-tumor genomic heterogeneity:

The application of next-generation sequencing (NGS) to breast cancer has revealed a very large genetic diversity among different breast tumors (reviewed in 5). However, although breast tumors are heterogeneous with regard to mutated genes, a part from the mutated genes can be grouped into the deregulation of similar pathways. This means that although the tumors are genetically different, some could be phenotypically similar due to mutations in the same pathway, which is very important in terms of treatment strategy. Comparing also the types of mutations present in different tumors has also allowed deriving the so-called mutational signatures, which reflect the different mutational processes which have been operational in the tumors. These studies have for example demonstrated that tumors from patients carrying hereditary BRCA1 or BRCA2 mutations are associated with different mutational signature than those from sporadic breast tumors (6).

4 Intra-tumor genomic heterogeneity:

NGS has also allowed to further explore intratumor genetic heterogeneity in primary breast cancer. These studies showed that although subclonal mutations were present in all tumors, there was always a dominant clone which comprised at least 50In approximately one fourth of invasive breast tumors, tumors are characterized by the presence of multiple synchronous unilateral lesions of invasive breast cancer, referred to as multifocal breast cancer. Although we have recently demonstrated that in all the investigated cases, these lesions always had a common somatic genetic ancestor, we showed that even when the lesions of a multifocal breast cancer present similar pathological characteristics, in one third of the cases the lesions can present different oncogenic alterations which can be clinically relevant (7).

5 Dynamicity of the disease:

An additional level of complexity resides in the fact that breast cancer is dynamic. Further evolution of the tumor and anti-cancer treatment can modify the presence and proportion of the different genetic alterations and/or clones of the tumor (8). In a recent study, we reconstructed the phylogeny of the genomic alterations occurring in the primary tumor and metastases from ten patients, which led to several observations (9). First, we noticed that reversions to the wild-type genotype in metastases are due to underlying copy number changes. Second, whole genome duplications seem to be early events, since they were conserved across metastases. Third, we noticed a striking dissimilarity of dissemination patterns of metastases in treatment nave patients compared to those undergoing primary surgery followed by adjuvant chemotherapy. Lastly, the level of heterogeneity appeared to be proportional to the time elapsed between the diagnosis of the primary tumor and emergence of the metastases. Taken together, these discoveries show that metastases of breast cancers are genomically different from their primary tumors. Altogether, this means that repeated sampling of the tumor at different time points might be needed for a more effective strategy to investigate potential treatment targets.

References

[1] Guiu S, Wolfer A, Jacot W, et al. Invasive lobular breast cancer and its variants: how special are they for systemic therapy decisions? Crit Rev Oncol Hematol. 2014

[2] Desmedt C, Gundem G, Zoppoli G, et al. San Antonio Breast Cancer Conference 2014

[3] NCI Dictionary of Cancer Terms - Tumor Microenvironment

[4] Ignatiadis M, Singhal SK, Desmedt C, et al. Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. J Clin Oncol. 2012 Jun 1;30(16):1996-2004.

[5] Desmedt C, Voet T, Sotiriou C, Campbell PJ. Next-generation sequencing in breast cancer: first take home messages. Curr Opin Oncol. 2012 Nov;24(6):597-604.

[6] Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 22;500(7463):415-21.

[7] Desmedt C, Fumagalli D, Pietri E, et al. Uncovering the genomic heterogeneity of multifocal breast cancer. J Pathol. 2015 Apr 6.

[8] Ellis MJ, Ding L, Shen D, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature 2012; 486: 353-60.

[9] Desmedt C, Brown D, Smeets D, et al. Annual meeting of the American Association for Cancer Research 2014

New Techs and oncology in System Medicine

Michele Libutti[(1)],

(1) ASL NA1- Naples, Italy
mthemail@gmail.com

Keywords:

Abstract. Today Medicine is still reactive, with a focus on developing therapies for diseases, especially in late stages of neoplastic diseases. Over the next 10 to 20 years, medicine will move toward a predictive and preventive model. New technologies will allow individuals to have the relevant portions of their genomes sequenced, and multi-parameter molecular diagnostics via blood analysis ,circulating tumour cells (CTC) and DNA or miRNA or cytokines and imaging systems will become a routine procedure for assessing the disease status. There will also be the chance to assess correlations of genetic variations with disease (differences between a primary and a metastatic disease in the same patient) and this combination of advances will allow the determination of prognosis. Predictive and preventative medicine will lead naturally to a personalized medicine that will revolutionize health care. Molecular diagnostics will allow drug-companies the opportunity for more effective means of drug discovery The paradigm will divide patients with a particular disease into a series of therapeutic windows, each with smaller patient populations but higher therapeutic chances. Health care providers will move from dealing with disease to also promoting wellness (prevention). Finally, the public must be educated to their roles in a very different type of medicine, as must the physicians who practice it. There will be enormous scientific and engineering challenges to achieve this vision far greater than those associated with the Human Genome Project.

Predictive, preventive, and personalized medicine will transform science, industry, education, and society in ways that we are only starting to figure it out.

The brainwaves associated with an individual's reaction to a certain word can be used for identification, according to a NewScientist article by Bas den Hond. The finding comes from a research team at the Basque Center on Cognition, Brain, and Language led by Blair Armstrong. The researchers monitored the brainwaves of 45 volunteers as they were reading a list of acronyms, and then ran the brainwave data through specialized software to analyze the differences between individuals. Their system was actually able to identify each volunteer based on brainwave differences - which themselves arise from the different memories and associations that each volunteer subconsciously connects to each acronym - and it did so with an accuracy of 94 percent. While that is not quite accurate enough for this system to be used as standalone security, den Hond points out that it does offer a unique capacity for continuous identification. And paired with another form of identification, that prove a valuable addition to the security apparatus. Moreover, Armstrong thinks it could be refined to provide even stronger results. It is certainly an unusual form of identification, but such intangible biometric technologies are increasingly being explored. Behavioral biometric systems, for example, are starting to catch fire, and even the identification of idiosyncratic movement patterns is being implemented into multimodal biometric systems. Brainwave authentication might sound like the stuff of science fiction right now, but it is clearly a feasible means of identification, and it could have a big role to play in the security systems of the future

Apple's new HealthKit app, released along with its iPhone 6, could offer the tipping point for the healthcare industry to shift into mobile-first telehealth services.

Apple's HealthKit collects and users health information through the new iPhone's advanced biometric sensors, which can detect steps, distances traveled, and even changes in elevation calculating calories . These sensors's metrics into the HealthKit platform's standardization of data are already a game-changing development in the management of big data.

With 90 million wearable health devices projected to be sold this year, the ability to agglomerate their myriad datasets into one uniform set of data that allows for meaningful analysis could be an enormously useful development for healthcare professionals and their patients.

Medical researchers in the academic world are already leaping into this field, with Stanford University teaming up with Apple to monitor the blood sugar levels of diabetic children, and Duke University working on an app that can track important cardiac measures for users at risk of heart disease.

Meanwhile, the private sector is already seeing companies like HealthLoop and American Well leap into action, developing healthcare apps of their own to take advantage of the new mobile biometric technologies.

Medisafe, a medication management application has begun to incorporate biometric data into its platform, the company has announced. Users of the iOS and Android apps will now be able to track metrics such as blood pressure and glucose levels via their devices biometric sensors and other digital equipment.

A biometrics developer has showed off a unique modality: ear biometrics. Developed by Descartes Biometrics Inc., the HELIX system can identify individuals based on the appearance of their unique ear shapes. HELIX boasts of a cross- platform library and is compatible with most camera-equipped devices, including webcams, smartphones. AMC Health, a chronic patient management specialist, has announced that it is going to showcase its solutions at ATA 2015, a major telemedicine conference. The company will do a joint presentation with Geisinger Health Plan (GHP) entitled "Using IVR and Telemonitoring to Lower Hospital Admissions, Readmissions and Costs for Heart Failure (HF) Patients". The presentation will revolve around the results of a study that monitored the effects of AMC Health's biometric remote care technologies over the course of two years. The 541 participating patients received care integrating Bluetooth-enabled biometric sensors and an interactive voice response system for capturing symptoms and behavioural data. What it found was that the patients, all of whom had previously been diagnosed with heart failure, were 23 percent less likely to need in-hospital treatment, and those who were hospitalized were 44 percent less likely to be re-hospitalized within 30 remote care solutions

Provider Sentrian has announced that it is embarking on a major study on the benefits of its technology. The company has teamed up with the Scripps Translational Science Institute (STSI) and the CareMore Health System to set up a study involving a thousand patients suffering from chronic obstructive pulmonary disease (COPD). The main goal of the year-long study is to see if Sentrian's Remote Patient Intelligence (RPI) platform can detect the onset of a patient's acute event, and to see how accurately it is able to do so. Positive findings would validate the RPI platform as a major tool in the battle to fight COPD; according to Sentrian, almost 13 million Americans suffer from COPD, and one in 11 such patients need to be readmitted to hospital within 30 days of being discharged. And the company says its platform could prove useful in the management of other diseases, as well. By working with scientific and health care leaders like Scripps and CareMore we can significantly reduce hospitalization and move toward the triple aim of health care to simultaneously reduce cost, improve outcomes and improve patient satisfaction.

WEARABLE TECHNOLOGIC GADGETS: examples.

1. Apple watch:

   -patients can monitor movement, steps, calories consumed

   -blood pressure

   -cardiac rhytm to enhance treatment of hypertension by healthkit pulse sensing

2. Google Glass: The device is already tested to catch and share data and situations that an Individual is actually being experiencing.

3. EYE Lens with glycemic control:

   -testing a smart contact lens that's built to measure glucose levels in tears using a tiny wireless chip and miniaturized glucose sensor that are embedded between two layers of soft contact lens material. this will lead to a new way for people with diabetes to manage their disease.

4. Data mining on millions of persons

5. Creation of maps dynamic maps will allow precise monitoring of data and behaviors (predictions of behavior of a phenomenon)

6. Send data direct to a phone

7. Theranos micro analysis with microfluid technology:

   -needs only micrometers of blood for tests- can keep patients out of hospitals for blood analysis

   -can help people take medicines on time and avoid use of alcool and calories

   -lower costs example: troponin test in acute hearth failure will cost 6.77 vs 1 4 dollars.

   -data sent to a wearable device.

ONCOLOGY The development of cost-effective technologies able to comprehensively assess DNA, RNA, protein, and metabolites in patient tumors has fueled efforts to tailor medical care. Indeed validated molecular tests assessing tumor tissue or patient germline DNA already drive therapeutic decision making. However, many theoretical and regulatory challenges must still be overcome before fully realizing the promise of personalized molecular medicine. The masses of data generated by high-throughput technologies are challenging to manage, visualize, and convert to the knowledge required to improve patient outcomes. Systems biology integrates engineering, physics, and mathematical approaches with biologic and medical insights in an iterative process to visualize the interconnected events within a cell that determine how inputs from the environment and the network rewiring that occurs due to the genomic aberrations acquired by patient tumors determines cellular behavior and patient outcomes. A cross-disciplinary systems medicine biology effort will be necessary to convert the information contained in multidimensional data sets into useful prognostic and predictive biomarkers that can classify patient tumors by prognosis and response to therapeutic modalities and to identify the drivers of tumor behavior that are optimal targets for therapy. An understanding of the effects of targeted therapeutics on signaling networks and homeostatic regulatory loops will be necessary to prevent inadvertent effects as well as to develop rational combinatorial therapies. Systems medicine approaches identifying molecular drivers and biomarkers will lead to the implementation of smaller, shorter, cheaper, and individualized clinical trials that will increase the success rate and hasten

the implementation of effective therapies. Trials will be implemented on a big number of Patients or Individuals, shifting from hundreds to millions of persons tested with great satisfaction for Biostatisticians.

References

[1] Hood,L Heath,J. R., Phelps, M.E. Lin, B Systems Biology and New Technologies Enable Predictive and Preventative Medicine, Science 306, 640, 2004

[2] Gonzalez-Angulo A.M, Hennessy, B. T.J., and . Mills, G.B. Future of Personalized Medicine in Oncology: A Systems Biology Approach, J Clin Oncol. 2010 Jun 1; 28(16): 27772783, 2010

ASSESSMENT OF THE ROBUSTNESS OF BAYESIAN P-SPLINES ESTIMATION TECHNIQUES FOR PROGNOSTIC ASSESSMENT AND PREDICTION

Giuseppe Marano (1), Patrizia Boracchi (2), Elia M. Biganzoli (1),(2)

(1) *Fondazione IRCSS Istituto Nazionale Tumori di Milano*
via G. Venezian 1, 20133 Milan, Italy, Giuseppe.marano@istitutotumori.mi.it

(2) *Dept. of Clinical Sciences and Community Health*
University of Milan, via .Vanzetti 5, 20133 Milan, Italy, patrizia.boracchi@unimi.it

Keywords: Hazard Smoothing, Piecewise Exponential Model, Bayesian P-Splines

Abstract. In this work we extended previous results concerning the regularized estimation of the Piecewise Exponential (PE) model through Bayesian P-splines techniques. A "standard" model with a smoothed piecewise hazard function, and a varying coefficient model with a time-dependent effect of a covariate have been fitted to a survival dataset from breast cancer patients. For evaluating the robustness of estimation method, several models with different structural components have been compared. The estimates were evaluated against "benchmark" patterns derived from validated clinical studies. The results showed very appreciable performances of the estimation method, though the sensitivity with respect to the prior of the smoothing parameter could be an issue, especially in small sample studies.

1 **Scientific Background**

In bio-statistical applications, non-parametric and semi-parametric survival analysis methods have been preferred over parametric ones for assessing the prognostic role of clinical/biological variables over time. The most widely adopted model is the Cox Model, in which no assumption of the functional form of the hazard function on time is made: however, such feature becomes a drawback if the interest lies on the hazard function itself or in predictive modeling.

Several methods for fitting survival models with a smoothed hazard function have been proposed. The use of Penalized Splines (P-Splines) may be of advantage over other flexible polynomials, since it avoids an arbitrary definition of the number of spline bases (by using a high number of basis). The degree of smoothing is controlled by regularized estimation thus reducing the effective number of model degrees of freedom, and therefore protecting against overfitting.

Penalized splines to smooth the hazard function in proportional hazard models have been proposed in the frequentist framework by Cai, Hindman, Wand (2002) and Kauermann (2004) amongst others. Their proposals are based on the relationship between penalized splines and mixed model theory (cfr: Ruppert, Carroll, Wand 2009). As a consequence mixed model software routines could be used for estimation. However the implementation of such a strategy could be cumbersome, because the Log-Likelihood function includes a complex integral without a closed form solution. An alternative is fitting a piecewise exponential (PE) model by penalized GLM estimation routines. However, in such case interval estimates are generally biased under standard large sample theory. Interval estimates of model parameters and the corresponding survival functions are derived by a Bayesian approach (Wood 2006).

Bayesian P-Splines techniques for hazard smoothing have been discussed by Fahrmeir and Hennerfiend (2003) ; dedicated routines are available in the software package BayesX. In such a context, point and interval estimates of model parameters and survival functions may be obtained in a straightforward way from posterior density samples. To this aim MCMC methods are needed but they require efficient sampling algorithms in order to guarantee convergence of Markov chains, and may be computationally intensive. The estimation method in Fahrmeir and Hennerfiend's paper also deals with the approximation of a complex integral without closed form solution, and in order to solve this problem specific sampling schemes have been adopted for computation. A dedicated software must be used. An alternative could be fitting the PE model. In fact, since the relationship between the likelihood function of this model and likelihood of a regression model with Poisson error, general sampling algorithms for hierarchical GLM models could be used for estimation. For example, WinBUGS code for penalized GLM regression is given in a paper by Crainiceanu, Ruppert and Wand (2005).

In a previous report (Marano, Boracchi, Biganzoli 2014) we assessed the performances of Bayesian P-Splines method for regularized estimation of the PE model, as compared with the a frequentist approach (GAM regression). The estimates of regression coefficients and survival probabilities were not affected by the choice of the priors for the Bayesian model, and showed a good agreement with GAM estimates. Concerning the hazard function, the results were sensitive with respect to the priors, and thus different shapes of the hazard were shown. In order to further evaluate the reliability of the estimated hazard function, we considered in addition different penalties (second order and third order difference penalties). Moreover, time varying coefficient models were also fitted. Futhermore, the behaviour of estimated hazards was compared against benchmark clinical situations derived from clinical literature in (i.e. papers on tumor dormancy in breast cancer; rif: Demicheli et al 2006 and 2008). Computation was performed in WinBUGS and thus the estimates were obtained by GIBBS sampling with the computationally efficient ARS algorithm.

2 **Patients and Methods**

Data were collected from a case series of 2210 breast cancer patients out of 2232 hospitalized at Istituto Nazionale dei Tumori di Milano who underwent conservative surgery (that is, quadrantectomy) and axillary lymph node dissection followed by radiotherapy (QUART) (cfr: Veronesi et al, 1995). The end-point was event-free survival (local recurrence, distant metastases, other priomary tumor, death). The predictors were well known prognostic factors available in the dataset: age at diagnosis (year) and tumor size (cm) (continuous variables): nodal status (4 categories: N-, 1 N+, 2-3 N+, ≥4 N+) histologic type (3 categories: infiltrating intraductal component or infiltrating lobular component (IDC + ILC), extensive intraductal component, EIC; other histologies), tumor site (2 categoris: external, internal or central). The continuous covariates were categorized according to conventional clinical criteria. We considered two general models: a "standard" PE model, where only the baseline hazard function was smoothed, and a varying coefficient model with a time dependent effect of tumor size. For the former model, the time axis was partitioned in 18 intervals of 1 year length; For the latter one, the follow-up period was truncated to 15 years, due to the insufficient number of events occurred at late follow-up times to patients within each tumor size category.

Concerning the smoothing method, the hazard function of the standard PE model was expressed by a piecewise constant function (corresponding to a B-Spline of degree 0), in the above partition of time axis. In the varying coefficient model, a time-dependent effect of tumor size was specified including in the model a 0/1 dummy variable for each of the five categories of

tumor size (the intercept was thus omitted). Then, a series of 15 time-varying coefficients (one for each interval of the partition of the follow-up) was specified for each indicator variable (see, e.g., Lambert and Eilers 2005).

Non informative (diffuse normal) priors were chosen for regression parameters (n=14). Hazard parameters (n=18 for the standard PE model, n=75 for the varying coefficient model) were expressed on log-scale, and Random Walk (RW) priors of order two and three were adopted for them (for details here and thereafter see the methodological note below). In particular, in the time-varying coefficient model five RW priors were specified, one for each category of tumor size.

For the smoothness parameters: τ^2; several non-informative priors, corresponding to standard specifications in hierarchical regression and Bayesian P-splines, were adopted The priors were: 1) a uniform(0,100) prior for $\sigma = 1/\tau$; 2) four inverse-gamma(k,k) priors for τ^2; k=1, 10^{-2}, 10^{-4}, 10^{-6}. Moreover the DIC criterion was computed in order to compare the goodness of fit of the resulting models. Results shown in the next section were obtained from one-chain samples, with length 20,000, burn-in 2,000 and thinning by 5. Posterior means and 95% Credible Intervals were computed for regression coefficients and .MCMC sampling was performed by executing WinBUGS under R (by the *R2WinBUGS* package) ; subsequent calculations were done by R with *coda* and *mgcv* packages added.

Methodological note:

In the Piecewise Exponential Model the instantaneous hazard function is piecewise constant on a given partition of the time axis. The model is fitted by GLM techniques, by specifying time intervals as predictors through dummy variables. A smoothed hazard function may be obtained through penalized estimation methods. The GLM model is the following:

$$\mathbf{Y_{ih} \sim POISSON(\mu_{ih}) ; \log(\mu_{ih}) = \alpha_h + \beta X_i + \log(\Delta_{ih})} \quad \mathbf{(1)}$$

where: Y_{ih} is the censoring indicator of subject i within time interval h; α_h the log-hazard in interval h; βX_i represents predictor effects; and $\log(\Delta_{ih})$ is an offset term with Δ_{ih} = "duration time" of subject i within interval h.

Bayesian P-Splines techniques are based on a hierarchical model for expression (1) :

$$\begin{cases} \mathbf{Y_{ih} \mid \alpha, \beta \sim POISSON(\mu_{ih}) \; ; \; \log(\mu_{ih}) = \alpha_h + \beta X_i + \log(\Delta_{ih})} & \text{(likelihood)} \\ \mathbf{\beta \sim \pi_\beta \; ; \; \alpha \mid \tau^2 \propto \exp(-\tau^2/2 \, \alpha' P \alpha) \; ; \; \tau^2 \sim \pi_\tau} & \text{(priors)} \end{cases} \quad \mathbf{(2)}$$

where the conditional prior (Multivariate Normal) of the hazard parameters: $\alpha = (\alpha_1, \ldots, \alpha_H)'$; depends on the penalty matrix P and on a smoothing parameter $\tau^2 = 1/\sigma^2$. Consequently, τ^2 is related to variances/covariances of hazard parameters and at the same time governs the amount of smoothing.

Since P corresponds to a finite difference penalty matrix, the above prior is equivalent to a Random Walk prior. Notably, this is essentially an auto-regressive Gaussian prior with rank-deficient covariance matrix and thus cannot be sampled directly. To specify the prior a reparameterization of the penalty matrix based on Gaussian Markov Random Fields theory was used; for details see (Fahrmeir, Hennerfiend 2003) and related literature.

3 **Results**

The models were fitted without incurring in convergence problems. Standard diagnostic techniques showed appreciable mixing properties of the Markov Chains, both for model parameters and for their functions (survival function), and showed autocorrelations rapidly approaching zero. A slower convergence of autocorrelations was shown only by the priors of the smoothing parameters, so that thinning was set to 5 for obtaining the final estimates.

For the standard PE model we fitted 10 different models with different specifications of priors and penalties (methods section). Since the shape of the hazard function is equivalent for each combination of covariate values (in a model without time dependent effects), we reported as example the estimated hazards for a patient with a specific low-profile risk. The estimates of the hazard function (Fig. 1) were practically undistinguishable. The shape resulting from Fig. 1 showed a major peak within the third year of follow-up, followed by a non constant trend, in which two other peaks could be noticed This behavior is consistent with accepted clinical knowledge on breast cancer casistics, altough the accuracy of estimates is rather low, especially in the later follow-up period.

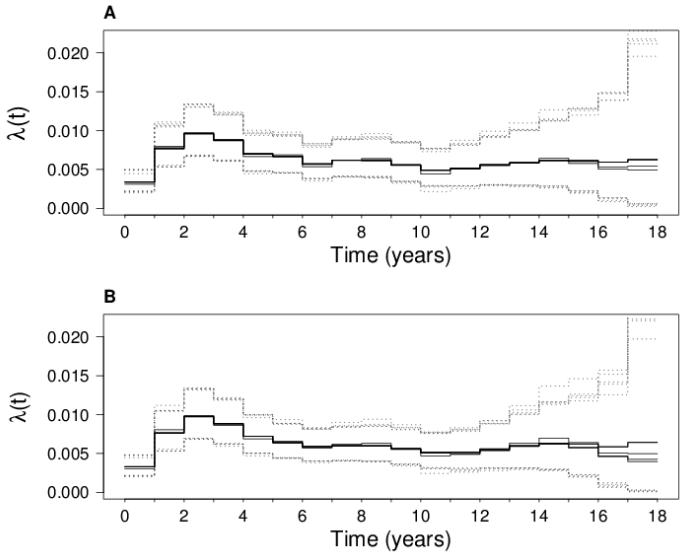


Figure 1: estimated hazard for the standard PE model. Reported in the figure is the hazard function for a patient with the following covariates: age=0-35 years, tumor size<0.5 cm, nodal status≥4 N+, quadrant=external, histology=IDC+ILC. Panel A: 5 models with a second order difference penalty and 5 distinct priors ; panel B: 5 models with a third order difference penalty and 5 distinct priors. Black solid lines: posterior means; dotted lines: 95% credibility intervals.

The statistics for model comparison were reported in table 1. No relevant differences emerged in terms of model fit (deviance), effective degrees of freedom (pD) and therefore in terms of overall adequacy (DIC). Thus, in the present example the choice of different priors and/or penalties had a negligible influence on posterior estimates.

The estimated hazard functions from the varying coefficient model were reported in Fig. 2. As in the previous case, the results were very robust with respect to priors and penalties; thus for sake of simplicity the estimates of only one model were reported in the figure. The most relevant patterns were shown for higher values of tumor size: from 1.1-1.5 cm to >2.0 cm. A high peak after the second year of follow-up may be seen in each case. For tumor size between 1.6 and 2.0 cm a second evident peak may be seen at the 10^{th} year of follow-up. These characteristics are,

again, consistent with the clinical experience. even though the estimates for tumor size >2.0 cm were inaccurate, due to the low number of patients in the respective sub-group.

Prior	Second order difference penalty			Third order difference penalty		
	Dbar	pD	DIC	Dbar	pD	DIC
Uniform(0,100)	2844.3	23.6	2868.0	2845.2	22.0	2867.2
I.Gamma(1,1)	2842.9	25.5	2868.4	2844.0	24.7	2868.6
I.Gamma(10^{-2},10^{-2})	2844.2	23.4	2867.5	2844.9	22.1	2867.0
I.Gamma(10^{-4},10^{-4})	2844.7	23.3	2868.1	2845.2	21.8	2867.0
I.Gamma(10^{-6},10^{-6})	2844.9	23.3	2868.2	2845.4	21.9	2867.2

Table 1: Assessment of model adequacy for standard PE models. Dbar = deviance. Pd = effective degrees of freedom; DIC = Deviance Information Criterion.

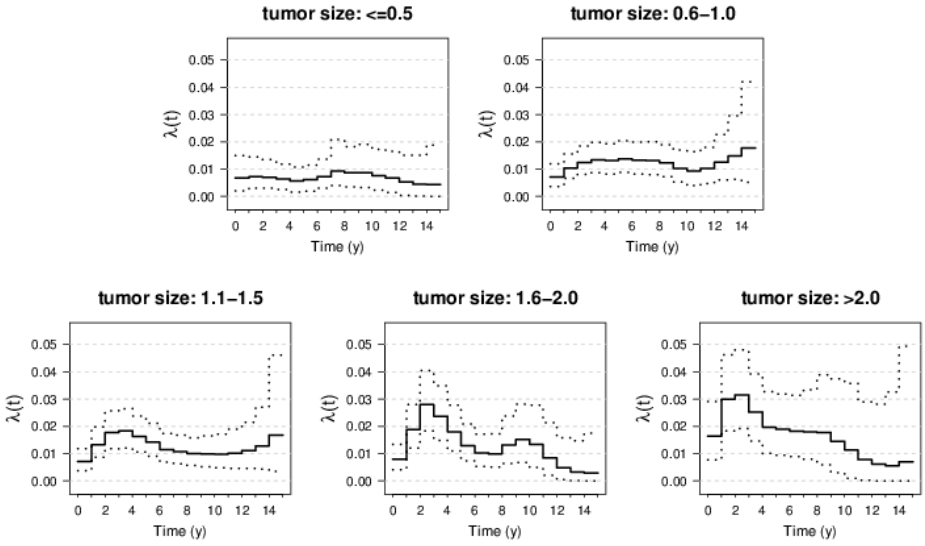


Figure 2: estimated hazard s for the varying coefficient model. Reported in the figure is the hazard function for a patient with the same covariates as in fig. 1, except for tumor size. Black solid lines: posterior means; dotted lines: 95% credibility intervals.

4 Discussion

As discussed in the previous work (Marano, Boracchi, Biganzoli 2014), the estimation of the PE model through Bayesian P-spline techniques provides a practical alternative for regularized estimation of survival regression models. The major advantages over other methods are:

1) a unified theoretical framework for assessment of prognostic covariate effects, assessment hazard functions, and prediction. In particular, in the frequentist setting, confidence intervals of model parameters and their functions are derived under non-standard theory.
2) estimation may be performed by general software routines for fitting mixed models (here: the ARS algorithm implemented in WinBUGS).

In this work we focused on the estimation of pre-specified regression models in a large case series with a long follow-up period. The results were in agreement with clinical knowledge, thus suggesting that the estimation method has a good reliability in our example.

Estimates of regression coefficients and hazard functions were very robust with respect to the specification of the smoothing components of the model (priors for the smoothing parameter, penalty matrix). By gathering these results with the previous ones we conclude that the sensitivity with respect to the priors above can affect only the estimates of the hazard function, and that biased results are more likely to occur when the sample is small.

The low frequency of events occurring at later follow-up times had some influence on the efficiency of MCMC sampling algorithm: in fact, we were forced to truncate the follow-up period to 15 years for avoiding hard convergence problems. However, MCMC diagnostics evidenced no further problems after the truncation. The low frequency of events also affected the accuracy of hazard estimates, as reported in the previous section. Thus, higher or more specific casistics are needed for a better evaluation.

Overall, the proposed method can extend in a practical way the tools available for the flexible modeling of survival functions allowing the study of the disease dynamics in complex frameworks like cancer follow-up studies.

Acknowledgements

This work was funded by the Italian Association for Cancer Research (AIRC). IG 2012 rif: 13420, 'Statistical Tools for Prognosis and Prediction in Cancer: Assessments and Application to a Sarcoma Case Series'. Elia Biganzoli was Principal Investigator. Giuseppe Marano was a fellow of AIRC.

References

[1] Cai, T., Hyndman, R. J., & Wand, M. P. Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics*, *11*(4), 784-798, 2002..

[2] Crainiceanu, C. M., Ruppert, D., & Wand, M. P. Bayesian Analysis for Penalized Spline Regression Using Win BUGS. Journal of Statistical Software 14(4)., 2005.

[3] Demicheli, R., Retsky, M. W., Hrushesky, W. J., & Baum, M. Tumor dormancy and surgery-driven interruption of dormancy in breast cancer: learning from failures. Nature Clinical Practice Oncology, 4(12), 699-710. 2007.

[4] Demicheli, R., Biganzoli, E., Boracchi, P., Greco, M., & Retsky, M. W. Recurrence dynamics does not depend on the recurrence site. Breast Cancer Res, 10(5), R83, 2008.

[5] Fahrmeir, L., & Hennerfeind, A. *Nonparametric Bayesian hazard rate models based on penalized splines* (No. 361). Discussion paper//Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München. 2003

[6] Kauermann, G. Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational statistics & data analysis*,*49*(1), 169-186, 2005.

[7] Lambert, P., & Eilers, P. H.. Bayesian proportional hazards model with time-varying regression coefficients: a penalized Poisson regression approach. Statistics in Medicine, 24(24), 3977-3989, 2005

[8] Marano, G., Boracchi, P., Biganzoli. E.M., Estimation of a piecewise exponential model by Bayesian P-splines techniques for prognostic assessment and prediction (2014); 11th international meeting on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB)

[9] Ruppert, D., Wand, M. P., & Carroll, R. J..Semiparametric regression during 2003–2007.*Electronic Journal of Statistics*, *3*, 1193, 2009.

[10] Veronesi, U., Marubini E, DelVecchio M, Manzari A, Andreola S, Greco M, Luini A, Merson M, Saccozzi R, Rilke F, Salvadori B. Local recurrences and distant metastases after conservative breast cancer treatments:partly independent events. Journal of the National Cancer Institute; 87:19 –27, 1995.

Special session on

New knowledge from old data: power of data analysis and integration methods

Organisers

Dr. Anagha Joshi (Roslin institute, University of Edinburgh)

Prof. Tom Michoel (Roslin institute, University of Edinburgh)

Transcription control in human cell types by systematic analysis of ChIP sequencing data from the ENCODE

Guillaume Devailly[1] and Anagha Joshi[1]
[1] Division of Developmental Biology, the Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, UK
guillaume.devailly@roslin.ed.ac.uk, anagha.joshi@roslin.ed.ac.uk

Keywords: ChIP sequencing, promoter, transcription, enhancer, sequence motif.

Abstract

Transcription control plays a key role during development and disease with trans-acting factors (TFs) regulating expression of genes through DNA interaction. ChIP sequencing is widely used to get the genome wide binding profiles of TFs in a cell type of interest. The reduction in cost of sequencing and the technological improvement has resulted in vast amount of ChIP sequencing data accumulating in the public domain. The ENCODE consortium alone provides 690 publicly available ChIP sequencing data sets across 91 human cell types. We performed a multi-facetted bioinformatics analysis of this data to unravel diverse properties of TFs in the cellular context. Specifically, we characterised genomic location as well as sequence motif preference of the factors. We demonstrated that the distal binding of factors is more cell type specific than the promoter proximal. We identified combinations of factors acting in concert at distinct genomic loci. Finally, we highlighted how this data is of value to associate novel regulators to disease by integrating it with disease-associated gene loci obtained from GWAS studies.

1 Scientific background

Chromatin immune-precipitation followed by high throughput sequencing (ChIP-seq) has become the standard method for identifying the binding sites of transcription factors and chromatin modifiers at genome-wide scale. As the data generation is now becoming a routine and the bottleneck has shifted to computational analysis of this data. This explosion of data has therefore led to a new path of discovery moving the field from hypothesis-driven to data-driven analysis. The ENCODE consortium [1] has been one of the leading projects which used same data generation and analysis quality control procedures across multiple labs worldwide to generate diverse genome-wide datasets across human cell types. It has so far produced 690 ChIP sequencing samples for transcription factors and transcription regulators, using 189 different antibodies (163 targeted factors) in 91 cell lines under different cell treatments. The combination of high experimental standards with extensive data release renders this ENCODE dataset invaluable for the scientific community, and serves as model for many other consortia. Notably, the ENCODE released 690 uniformly processed peak files as well as one track combining the peaks at the UCSC genome browser, frequently used by many labs world-wide. We performed a systematic analysis of this data to understand diverse aspects of transcription control across human cell types.

2 Materials and Methods

Peak lists from ENCODE ChIP sequencing experiments were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/. For each peak, distance to the closest TSS was obtained with bedtools [2] from the list of TSS provided by GENCODE (v21) [3]. Overlapping peaks in the 474 experiments done in the six cell lines with over 35 experiments were merged using bedtools, to obtain a binary matrix of 553,211 non overlapping genomic regions x 474 experiments. Two subsets of this matrix were isolated according to distance to the closest TSS: a matrix containing the 300,901 regions at less than 1 kb from the closest TSS and a matrix containing the 144,514 regions at more than 10 kb from the closest TSS. Pearson correlations were then calculated between each pairs of experiments, and the resulting correlation matrices were clustered according to the method described in [4]. From the original matrix containing 553,211 non-overlapping regions, we compute for each cell line Transcription Factor (TF) density in each regions by dividing the number of factors with a peak in that region by the total number of ChIP sequencing experiments done in that cell line, to obtain a matrix of 553.211 regions x TF density in six cell lines. K-means clustering of the regions lead to the identification of 417,597 lowly bound regions in all cell lines, while other regions shown high TF density in one, two or all cell lines. The sequence motif enrichment was performed using HOMER [5]. A list of significantly disease associated regions were downloaded from https://www.ebi.ac.uk/gwas/. P values were calculated using Bonferroni-corrected hypergeometric test. The enriched combinatorial patterns were calculated by generating random binding data keeping the same number of peaks for each factor and the significance was estimated compared to 100 randomizations.

3 Results

The ENCODE consortium has generated ChIP sequencing data for transcription factors and transcription regulators providing 690 uniformly processed peak files. These samples were generated using 189 different antibodies (163 targeted factors) in 91 cell lines, with some samples under cell treatments and each sample typically results from merging the analysis of at least two biological replicates. The number of peaks for a factor in a cell type ranges from few hundred to tens of thousands of peaks. ZNF274 in HeLa-S3 cell type has the least number of peaks called (n = 74) while CFOS in MCF10A cell line has the most number of peaks called (n= 91,953) in this compendium. Here we elaborate diverse aspects of transcription control investigated using this data as well as integrating it together with other genome-wide datasets.

3.1 Genomic location binding preference of transcription factors

We annotated all peaks from the 690 ChIP-seq datasets using the closest transcription start sites identified by GENCODE (v21) [3]. We then quantified the number of peaks overlapping a TSS as well as the fraction of peaks localised more distally from a TSS. All patterns can be grouped largely into five groups where peaks from group 1 and 2 peaks are preferentially found at distal regions while peaks from group 4 and 5 are preferentially found at promoter proximal regions. Group 4 contains experiments with at least 50% of peaks (Figure 1 B) overlapping a TSS. This includes various forms of RNA polymerase II as well as classical promoter associated transcription factors such as ELK1 or BRCA1. This group defines factors with high affinity for promoters that may play a role in TSS specification. On the other hand, group 1 constitutes of experiments where only about 10% of peaks overlapped a TSS (Figure 1 experiments at the top of the plot). This group includes factors such as CTCF, RAD21, SMC3, and CEPB known to be bound to distal regulatory regions acting as enhancers or insulators. Group 2 and group 5 experiments have less than 50% of peaks at TSS, and include factors such as EZH2, POU2F and BHLHE40. Group 3 consists of two smaller clusters constituted of experiments where peaks were not symmetrically spread between downstream and upstream of TSS.

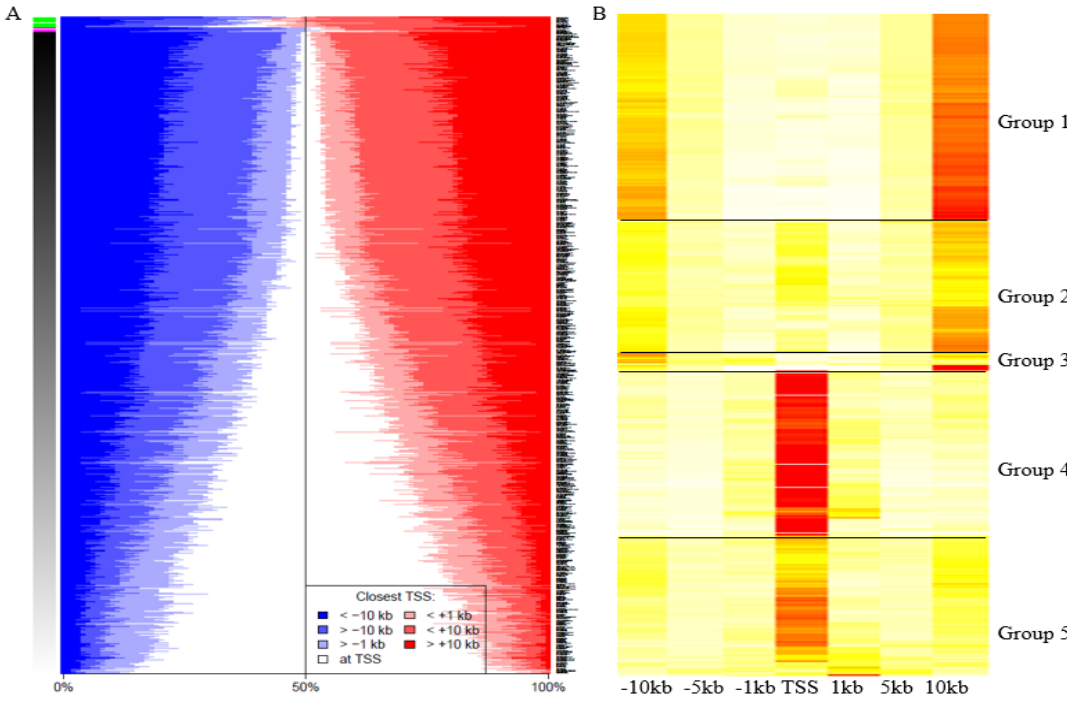


Figure 1: A. For each ChIP sequencing experiment, the proportion of peaks overlapping a TSS in represented in white, the peaks upstream the closest TSS are in blue, while peaks downstream a TSS are in red. The green side bar indicates experiments with more than 50% of peaks upstream a TSS. The magenta side bar designates experiments with more than 50% of peaks downstream a TSS. The black to white side bar correspond to experiments with respectively more to less peaks distal from a TSS. **B.** k-means clustering of the data in A. Colour code corresponds to fraction of peaks, from low (0=white) to medium (0.25=yellow) to high (0.5=red).

Green labelled experiments (Figure 1A) have more peaks located downstream of a TSS than upstream. It includes ChIP sequencing experiments for the elongation specific Pol2 phosphorylated on serine 2, as well as ChIP-seq against ZNF274 which is known to bind preferentially to 3' end of zinc finger genes [6]. On the other hand, magenta labelled experiments (Figure 1A) constituted of experiments with higher proportion of peaks located upstream a TSS than downstream. It include RNA polymerase III and its co-factor TFIIIC (and to a lesser extent RPC155 and BRF1). The upstream location of these factors might be due to the presence of unannotated RNA pol III transcribed genes in intragenic regions. In several cases, ChIP against the same factor but in a different condition (either another cell line, another treatment, another antibody, another library preparation protocol or another laboratory) were found in different categories. For example, c-Fos ChIP-seq done by the Yale laboratory in GM12878 cells was found in the first group while those done by Harvard, UCSC or Yale in MCF10-A, HUVEC and HeLa-S3 were in the third group. Understanding whether these situations reflect biological differences or experimental artefacts will need further investigations.

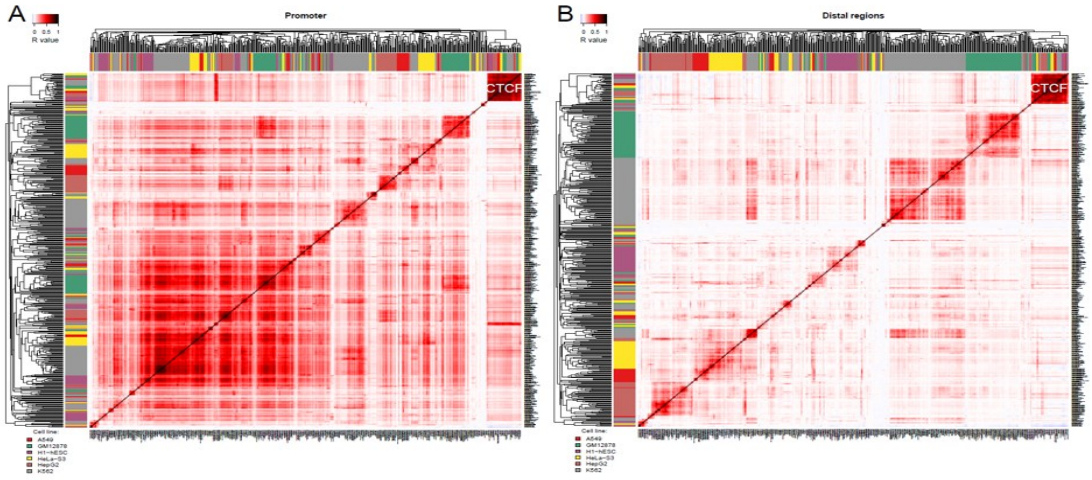


Figure 3: Correlation matrix across all ChIP-seq experiments in the six main ENCODE cell lines. For each pair of experiments, Pearson correlation coefficient was computed between peak lists. Then the correlation matrix was clustered. Side colours correspond to each cell line. **A.** Correlation matrix for peaks close to a TSS (< 1 kb). **B.** Correlation matrix for peaks far from a TSS (> 10 kb).

3.2 Bound regions distal to a TSS are cell line specific

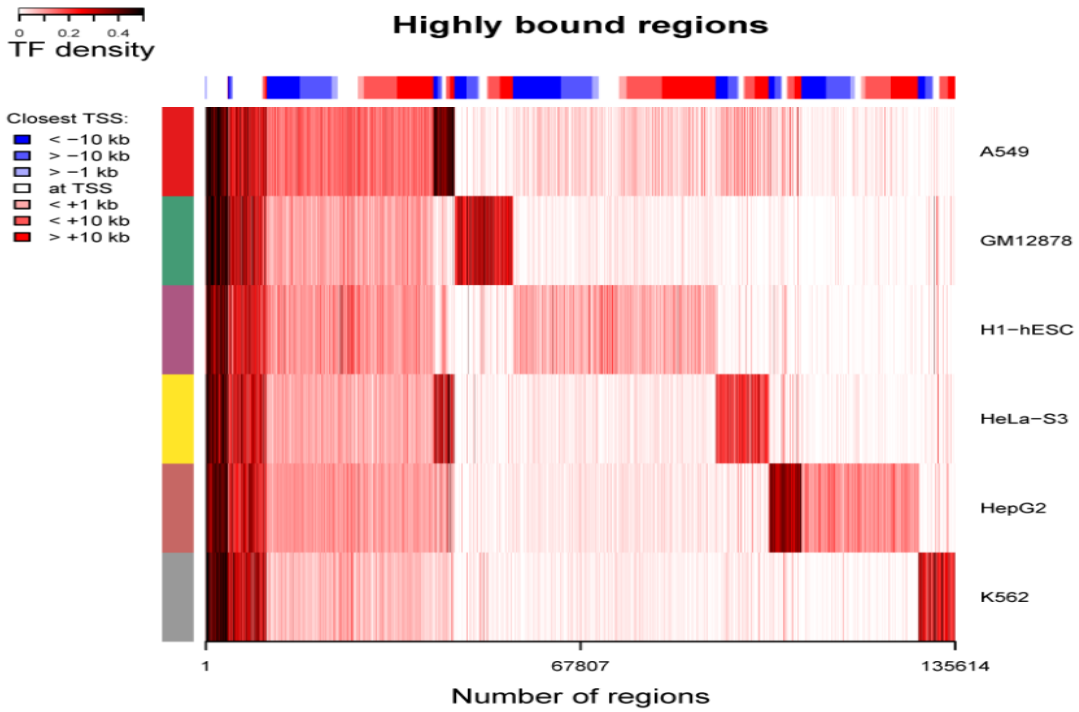


Figure 2: Clusters of genomic regions with high TF density across 6 cell lines. For each non overlapping regions, we computed the proportion of ChIP-seq experiments done in one cell line with a peak in that region. Clusters were identified by k-means clustering. For visibility, a big cluster of low TF density in every cell type is not shown (417,597 regions). Top colors bars represent the fraction of regions localised at a TSS (white), upstream a TSS (blue) or downstream a TSS (red).

In the ENCODE dataset, six cell lines contains more than 35 ChIP-seq experiments (A549, GM12878, H1-hESC, HeLa-S3, HepG2, and K562), with a total of 474 experiments. From the 553,211 regions bound by at least one factor in the 474 experiments, 300,901 where promoter proximal (<1 kb from the nearest TSS), and 144,514 where far from any TSS (> 10 kb). We calculated the Pearson correlation coefficient between each pair of peak lists for the promoter proximal regions (Figure 2A), and for the distal regions (Figure 2B). The hierarchical clustering of the resulting correlation matrix for each of the two sets demonstrated that promoter regions share binding sites of many more factors compared to the distal regions. Moreover, the factors studied in same cell type clustered together more often at the distal regions than at the promoter regions (Figure 2). The promoter landscape represents three main clusters: a CTCF/SMC3/RAD21 cluster, a large multifactor cluster including diverse factors, and the third cluster with very low correlation scores, which includes the aforementioned ZNF274 together with repressive complexes such as EZH2/SUZ12 and SETDB1/KAP1. This cluster also includes c-MYC ChIP-seq experiment in untreated H1-hESC by UTA. Other c-MYC samples do not cluster together with this sample putting its experimental quality into doubt. At distal regions, the CTCF/SMC3/RAD21 cluster remains intact but many other experiments clustered according to their cell line of origin, meaning that TF binding at distal regulatory regions tends to be cell line specific.

rank	1	2	3
A549	HDAC6, P300, ELF1, ETS1	ATF3, BRF1	RNA pol2, CTCFL
GM12878	SAP30, TAF7	STAT5A, BRG1	P300, ETS1
H1-hESC	RAD21, ZNF143	BACH1, MAFK	USF1, USF2
HeLa-S3	EZH2, RNA pol2 4H8, SIN3A, CJUN, CMYC	GTF2B, NR2F2	RNA pol2, RBBP5
HepG2	TAF1, TAF7, TEAD4	SAP30, ATF1, ATF3	PU1, STAT5A
K562	GTF2F1, CTCF	HDAC1, CJUN	E2F6, CTCF

Table 1: Top 3 significant associations between factors in 6 human cell lines. All associations were predicted at very high significance (all P values < 1e-256).

3.3 Combinatorial control of regulators

Mammalian transcription factors are known to work together by binding at the promoter or enhancer regions to activate or repress downstream target genes. To unravel if some regulators are preferentially binding together genome-wide, we built a M*N matrix of all binding events (peaks) in a cell type where M represents the genomic loci bound by at least one factor in that cell type and N being the number of factors studied by ChIP sequencing in that cell type. We build a matrix for each of the six cell types (A549, ES cells, GM12878, HeLa, HepG2 and K562) with N of more than 30. Each cell type has 2N-1 combinations of binding patterns possible. We then evaluated likelihood of frequency of combinatorial patterns to occur by chance by comparing to 100 random datasets generated such that the total number of binding events for each factor was preserved (Table 1). This analysis re-discovered the transcription factors of the same family (e.g. USF1 and USF2 in ES cells) known to bind to overlapping genomic locations due to highly similar sequence motifs or the components of known complexes or well studies interactions (e.g. SUZ12, EZH2). CTCF, RAD21 and ZNF143 form a part of cohesin complex and cluster together in the global clustering considering all peaks. Accordingly, they were enriched across multiple cell types. Interestingly, there are a number of cases of combinatorial control where the two factors do not share most of the binding sites therefore do not cluster together in the global clustering tree (Figure 2A) but are significantly co-occupying a relatively small but statistically significant number of gene loci. For example, GTF2F1 and CTCF co-bind 2,233 loci in K562 cell line (P value < 1e-256) and these loci are not occupied by any other factor of 150 factor studied. Similarly, in ES cells, 957 genomic loci are co-occupied only

by RAD21 and TEAD4 (P value < 2.1e-37). In GM12878, ETS1 co-occupies 1090 binding sites only with EGR1 (P value < 1e-256) and 1249 binding sites only with P300 (P value < 1e-256). This postulates site specific role for these combinatorial interactions shadowed by the global analysis approach.

3.4 Sequence motif preference of factors

To investigate the sequence motif preference of each factor characterized by ChIP sequencing experiment, we identified the enrichment of known motifs in the peak list using HOMER [5]. As expected, for many experiments, the analysis resulted in detecting the sequence motif specific to the factor as the top motif enriched for the factor. For example, a CTCF motif (AYAGTGCCMYCTRGTGGCCA) was top motif of all CTCF ChIP sequencing experiments across 68 cell types. This confirms the quality of the data for downstream characterization. Importantly, CTCF ChIP sequencing in different cell types did not result in enrichment of sequence motifs of cell type specific factors. This demonstrates that CTCF acts mainly as an insulator as well as defining gene regulatory boundaries which are largely independent of cell type. Similarly, the majority of RNA polymerase II experiments across multiple cell types identified ETS as a top enriched motif with only a handful of cases with a cell type motifs enriched such as GATA motif enriched in K562 RNA pol II sample or BZIP motif enriched in HeLa-S3 RNA pol II sample. In ESCs, OCT4-SOX2-TCF-NANOG motif was enriched for the ChIP sequencing of OCT4 and NANOG as expected, moreover this motif was top enriched in P300, BCL11A, HDAC2 and CTBP2 samples as well.

3.5 ChIP sequencing binding overlap with disease susceptibility loci

The ENCODE consortium has demonstrated that a large number of intergenic disease associated gene loci are located in the regulatory regions defined by chromatin modifications and DNase I hypersensitive sites across cell types. To systematically analyse overlap of transcription factors binding loci with Genome Wide Association study (GWAS) high confidence hits,. Overlap of GWAS hits with ChIP-seq peaks was calculated and the significance of overlap was estimated using bonferroni corrected hypergeometric test. Three of the 690 ChIP sequencing samples: BRCA1 bound loci in GM12878 (2% of peaks), NELFE bound loci in K562 (2.1% of peaks), HDAC8 bound loci in K562 (2%) showed statistically significant overlap with GWAS disease associated loci. All three protein have a well-studied role implied in cancer. As expected, the majority BRCA1 bound loci overlapped with the disease loci identified in breast cancer. Interestingly, 37 BRCA1 target loci overlapped with inflammatory bowel disease. BRCA1 also functions as an important mediator of innate immunity and BRCA1 gene therapy reduces systemic inflammatory response [7]. 12 BRCA1 target loci overlapped with genes linked to childhood obesity. In line with this, it has been shown that without BRCA1, muscle cells store excess fat and start to look diabetic [8]. Taken together, the analysis of transcription targets using ChIP sequencing overlapping with disease associated loci has a power to identify novel factors controlling the disease phenotype.

4 Conclusion

As the next generation sequencing tool is becoming readily available to the experimental groups world-wide, the major challenge lies in computational analysis of this large resource.

This data can be analysed in a plethora of ways integrating together with other datasets, to obtain unexplored biological insights as yet. The ENCODE consortium has taken a big initiative to provide a uniformly processed dataset to the scientific community for computational data integration and analysis. This easy accessibility of pre-processed data facilitates generation of novel biological hypotheses from this resource.

In this paper, we demonstrate five ways of analysing this resource integrating it with other data resources. These approaches have a potential to develop new hypotheses about transcriptional control mechanisms. We both reproduced and expand some observations previously made in ENCODE

companion papers such as [9], and we also developed novel analytic approaches. This shows that the in depth analysis of ENCODE data is still far from complete and must be continue. Importantly, all the analysis performed here can be readily transferable to the exploitation of ChIP-Seq datasets in other cellular systems, and thus have the potential to significantly advance our understanding of a wide range of both normal and pathological cellular processes.

Funding

A.J. is a Chancellors Fellow at the University of Edinburgh. This work was supported by the Roslin Institute Strategic Grant funding from the BBSRC.

References

[1] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder, "An integrated encyclopedia of DNA elements in the human genome.," Nature, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[2] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features.," Bioinformatics, vol. 26, no. 6, pp. 841–2, Mar. 2010.

[3] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, "GENCODE: the reference human genome annotation for The ENCODE Project.," Genome Res., vol. 22, no. 9, pp. 1760–74, Sep. 2012.

[4] K. Pötzelberger and H. Strasser, "Data Compression by Unsupervised Classification." Department of Statistics and Mathematics, WU Vienna University of Economics and Business, 11-Jul-1997.

[5] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.," Mol. Cell, vol. 38, no. 4, pp. 576–89, May 2010.

[6] S. Frietze, H. O'Geen, K. R. Blahnik, V. X. Jin, and P. J. Farnham, "ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes.," PLoS One, vol. 5, no. 12, p. e15082, Jan. 2010.

[7] H. Teoh, A. Quan, A. K. Creighton, K. W. Annie Bang, K. K. Singh, P. C. Shukla, N. Gupta, Y. Pan, F. Lovren, H. Leong-Poi, M. Al-Omran, and S. Verma, "BRCA1 gene therapy reduces systemic inflammatory response and multiple organ failure and improves survival in experimental sepsis.," Gene Ther., vol. 20, no. 1, pp. 51–61, Jan. 2013.

[8] K. C. Jackson, E.-K. Gidlund, J. Norrbom, A. P. Valencia, D. M. Thomson, R. A. Schuh, P. D. Neufer, and E. E. Spangenburg, "BRCA1 is a novel regulator of metabolic function in skeletal muscle," J. Lipid Res., vol. 55, no. 4, pp. 668–680, Feb. 2014.

[9] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder, "Architecture of the human regulatory network derived from ENCODE data.," Nature, vol. 489, no. 7414, pp. 91–100, Sep. 2012.

Novel unsupervised learning methods for single cell data visualization and trajectory inference

Leen De Baets[(1,2)], Sofie Van Gassen[(1,2)], Tom Dhaene[(1)], Yvan Saeys[(2,3)]

(1) Internet Based Communication Networks and Services (IBCN), Ghent University - iMinds, Gaston Crommenlaan 8 (Bus 201), B-9050 Gent, Belgium
(2) Data mining and Modelling for Biomedicine (DaMBi), IRC - VIB, Technologiepark 927, B-9052 Gent, Belgium
(3) Department of Respiratory Medicine, Technologiepark - Ghent University Hospital, De Pintelaan 185, B-9000 Ghent, Belgium

Keywords: cell differentiation, trajectory inference, graph modelling, flow cytometry.

Abstract. Single cell analyses allow the study of cell-to-cell variation within mixed cell populations, and recently new computational tools are being developed that allow scientists now to study cell differentiation into greater detail using automated data mining methods. While single cell data (e.g. obtained by flow cytometry) used to be analyzed mainly manually by biological experts, recent advances such as mass cytometry and single cell RNA sequencing are generating large and high-dimensional data that call for automated and unbiased analysis. In this paper we present new methods for the unsupervised analysis of single cell data. We start out by introducing recent visualization techniques and further build on these concepts to define new graph-based methods to infer cell differentiation from single cell cytometry data.

1 Scientific Background

In the bone marrow, stem cells mature into different precursor cells that further develop into different terminal cell types in a coordinated fashion. For hematopoietic stem cells, it is generally assumed that early precursors such as long-term hematopoietic stem cells (LT-HSC) can differentiate into short-term hematopoietic stem cells (ST-HSC) and that these in turn can differentiate into either common myeloid progenitor cells (CMP) or common lymphoid progenitor cells (CLP). Each of these lineages then further differentiates into more mature cell types ending up in a final, mature cell type. To study cell differentiation at the single cell level, traditional flow cytometry methods offer a high-throughput platform that typically measures the presence of 10 up to 20 cellular markers on millions of cells. However, recent advances in cytometry such as mass cytometry, which measures up to 40 markers, and single cell RNA sequencing greatly increase the amount of parameters that can be measured for each single cell.

Traditional analysis of flow cytometry data is done by examining the data using two-dimensional scatter plots. However, as the number of parameters describing each cell has drastically increased, manual analysis of such data is becoming infeasible, and novel computational visualization techniques such as SPADE [1], Visne [2] and FlowSOM [3] have been proposed. As an example, Figure 1 shows the result of the FlowSOM visualization of flow cytometry data regarding mouse hematopoietic stem cells differentiating into common myeloid and lymphoid progenitors. In this visualization, each circle represents a cluster of similar cells, and larger cell types are represented as different branches in a minimal spanning tree representation. While FlowSOM is clearly able to separate known cell types, we can only infer in a limited way information about the underlying cell differentiation process. Reading the figure from top to bottom we are able to see that LT-HSC cells are closely related to ST-HSC cells, and that CLP's branch off on the left

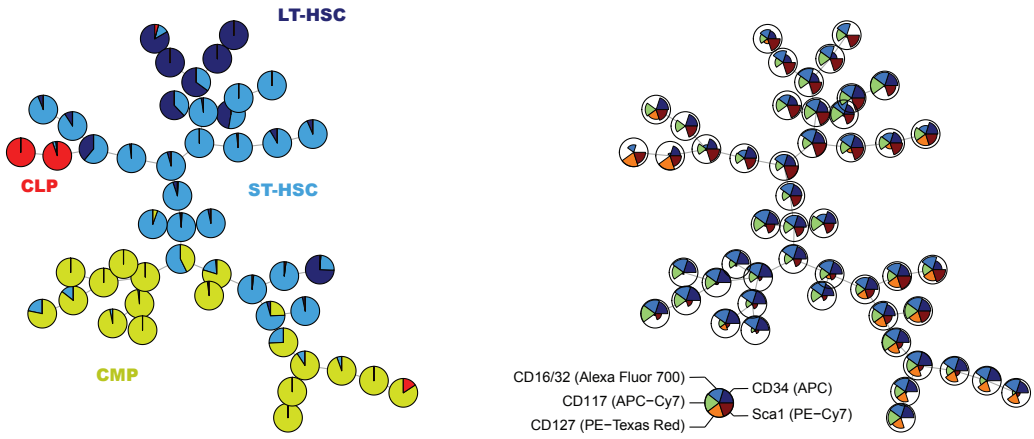


Figure 1: FlowSOM representation of early hematopoietic stem cells. Each node represents a group of similar cells. On the left hand side, a manual annotation of the cells (shown as different colors) is visualized. FlowSOM is clearly able to group known cell types in similar branches of the minimal spanning tree representation. On the right hand side, the mean fluorescence intensities for all measured cell parameters are indicated. Slight variations in marker intensities for a single cell type can be noticed.

hand side, while CMP's branch off on the bottom side. However, no conrete ordering of these cell types regarding developmental order can be derived from such visualizations.

To better understand and model the underlying cell differentiation process, we here propose a new unsupervised learning technique that we refer to as *trajectory inference*. Trajectory inference bears many resemblances to several techniques in different fields, but its specific characteristics call for an extension of the current types of unsupervised learning techniques to model it in an adequate way. To define the problem, imagine a simplified dynamic process underlying hematopoietic stem cell differentiation, where cells start as LT-HSC's (state 1), evolve to ST-HSC's (state 2) and subsequently branch into two separate pathways, finally leading to CMP's (state 3) or CLP's (state 4). Figure 2(a) shows a simplified state diagram representing this developmental process. While transitioning from state 1 to 2, certain aspects of the cells change, e.g. increasing values of markers 1 and 2 (Figure 2(b)), and e.g. while transitioning from state 2 to 3 marker 1 starts to increase more and marker 2 decreases.

As hematopoietic stem cells evolve in a continuous manner with many cells being involved in the process, a single snapshot of the developmental system can be presented in Figure 2(c), assuming for simplicity that cells only have two markers. Ideally, when all intermediate cell states would be known, it would thus be possible to write down the sequence followed from the most immature state (state 1) to the most mature state (state 3 or 4), further referred to as a trajectory. However, as there is still a lot of uncertainty about the developmental trajectories that cells follow in the case of hematopoietic stem cells, but also in general, it would be interesting to infer such trajectories automatically from data represented in Figure 2(c) using computational methods. This will provide novel hypotheses about cell differentiation, possibly revealing new intermediate cell stages and unexpected changes in marker expression.

The inference of trajectories from data like in Figure 2(c) (mixed cell data from a single snapshot), although related to many existing fields in computational modelling, is a relatively new problem with particular challenges. First, the problem is more general than just a clustering or density estimation problem, as the transitions between the

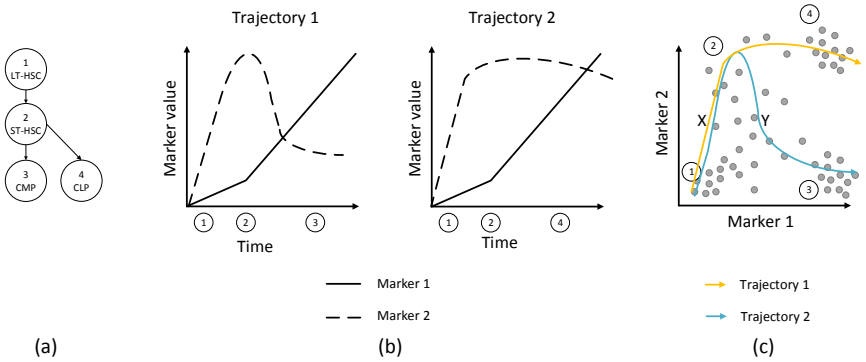


Figure 2: Illustration of the concept of trajectory modelling where (a) shows the states that are traversed, (b) shows the parameter changes when the trajectory is followed, and (c) shows a snapshot of the objects following the two trajectories (yellow/blue lines).

states are smooth and gradual, and the underlying trajectory rather defines a continuous path related to the intrinsic time component of the underlying dynamic process. It should be noted that this 'developmental' time component should be retrieved from the data, which could be deceptive, as illustrated by the cells X and Y in Figure 2(c). These cells, although close in terms of Euclidean distance, are still far away 'developmentally', i.e. when traversing the blue line from X to Y according to the underlying trajectory. Secondly, the problem is more general than state-transition models such as Hidden Markov Models (HMMs) since these assume input that is already ordered along the different states. However, this is exactly what we want our algorithm to present as output, rather than input, since we only have unordered data. Finally, the problem is also more general than general function estimation, as the underlying state transition diagram need not be linear, but may contain branching events, leading to hierarchical or tree-like structures, or more general to any graph-like structure.

2 Materials and Methods

In this work, we present a new computational method to identify trajectories from data given cells representing the most immature and mature cell states. Our method extends the Wanderlust algorithm [4] that is capable of detecting single, linear trajectories where all cells must evolve to the same mature cell type.

In section 2.1, we explain shortly how Wanderlust is capable of constructing a single trajectory in unordered data. In the next section, we formulate our solution to infer branched trajectories from unordered data.

2.1 Inferring a single, linear trajectory

Theoretically, a trajectory can be seen as an ordering of cells, sorted according to their developmental order. Immature cells are thus sorted before mature cells. Cells that are developmentally close will be close in distance, but the opposite is not necessarily true, as was illustrated above in Figure 2(c). The Wanderlust algorithm aims to solve this by representing the data as a k-nearest neighbour (knn) graph such that only developmentally close cells are supposed to be connected. However, due to noise it is still possible that cells that are developmentally far away are connected. To handle this, an ensemble is created consisting of n graphs where in each graph each node contains l edges ($l < k$) that are randomly chosen from the k edges present in the knn graph, resulting in n l-knn graphs.

Wanderlust creates a trajectory by ordering all cells according to their similarity to the starting cell. This cell typically represents the most immature cell, which should be provided as input to the algorithm. The similarity is then defined as the distance

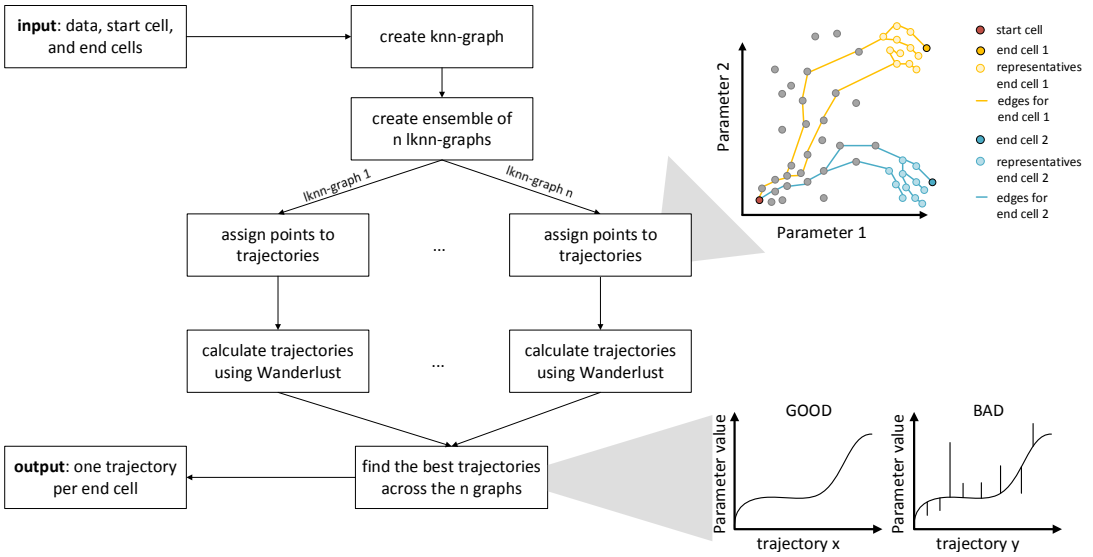


Figure 3: Flow chart of the algorithm to detect branched trajectories. The steps concerning the assignments of points and finding the best trajectories are detailed in the figure insets.

of each cell to the starting cell in the l-knn graph. As this is done for each graph in the ensemble, we have l trajectories. The final one is obtained by averaging each cell's distance across all n l-knn graphs. The most restrictive assumption of this algorithm is that all cells must follow one and the same trajectory. This is a necessary assumption as the cells are ordered with respect to their distance to the starting cell of the trajectory. In the following section, we will formulate a solution that relaxes this constraint, and allows to infer branching trajectories.

2.2 Inferring branched trajectories

Our solution to infer multiple trajectories from unordered data is given in Figure 3. As input, we not only use a starting cell representing the most immature state, but also one end cell for each mature state. Using this information, we first assign each cell in the dataset to one or more trajectories, each trajectory going to one end cell. Once this is done, we can apply the original Wanderlust algorithm and aggregate the results in one final trajectory per mature state.

Assigning cells to a trajectory is done by clustering the data with k-means using a large amount of clusters (overclustering). In these clusters, we find the ones to which the given end cells belong to. This allows us to represent each mature state not only with one end cell, but with all the cells in the cluster: the representatives. In the next step, we determine for each representative the edges of the shortest path to the starting cell, resulting in a collection of edges for each mature state. Then for every cell, we determine the edges of the shortest path to the start cell and compare it to the edge collections resulting in the detection of the most similar collection. This comparison is done by determining the amount of overlap. Knowing the most similar collection, we also know the most similar end cell, and accordingly we also know that this cell should be assigned to the trajectory going from the start cell to this end cell. Using this approach allows for a cell to belong to multiple trajectories as the overlap for different edges collections can be the same. This will for example be the case for the first (common) part in the trajectory of hematopoietic stem cells, where all cells first transition from the LT-HSC to the ST-HSC state.

Due to the usage of the ensemble of graphs, we have for each mature state n possible

trajectories. The next task is to reduce this amount to one per mature state. This is done by taking the best trajectory. To quantify this without using extra information, we add a regularisation procedure by using the fact that changes between the states in the trajectory are gradual. Due to this property, the curves representing the parameter values of the cells along the trajectory (Figure 2(b)) must be smooth. More specific, if these curves contain large jumps (e.g. see Figure 3), there is a high probability that not all cells are included or the ordering is wrong. Concretely, we calculate the difference of these curves with their smoothed version and the trajectory with the smallest difference is the best one. As a result, we have one trajectory per mature state.

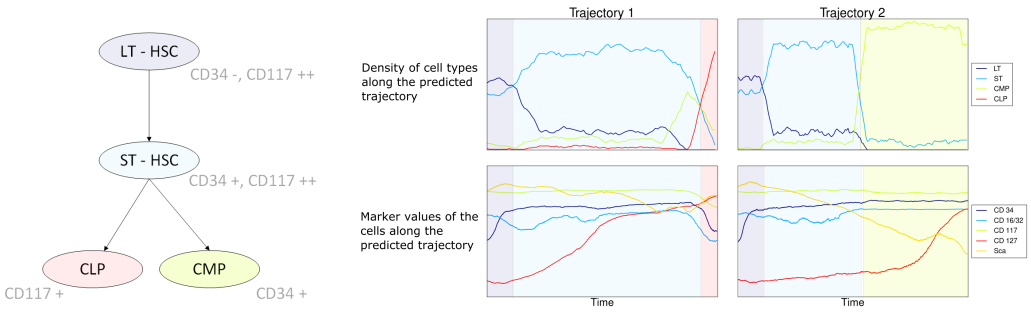
3 Results

We evaluated our algorithm on the use-case of early hematopoietic stem cells, obtained from traditional flow cytometry data concerning 4647 mouse bone marrow cells that evolve from LT-HSC's to ST-HSC's, and then further to CLP's or CMP's. The cells can thus follow two trajectories, both of which are expected to include the LT-HSC and ST-HSC. Our data set includes measurements of five surface markers: CD34, CD$16/32$, CD117, CD127, and Sca-1. Only for a few of these markers it is known which values they have in the different cell states [6], indicated in Figure 4(a). By forming a trajectory from the unordered data, we will evaluate to what extent our algorithm recovers these marker trends automatically, and how and when other changes happen.

To validate our algorithm, we made use of manual cell annotations that were provided by a biological expert. A simple way to visualize the resulting trajectories is to plot the amount of a particular cell type during an interval (e.g. 20 cells) in the trajectory, leading to a density function. This leads to two plots, one for each trajectory, where each plot contains four curves representing the density of the four different cell types present. This can give us an idea of the correctness of the assignment algorithm as it shows the cell states that are traversed by a trajectory. Another way to visualize the resulting trajectories is to plot the curves representing the marker values of the cells along the trajectory as done in Figure 2(b). This results in two plots, one for each trajectory, where each plot contains five curves, one for each marker value. A way to evaluate the result, is to compare these curves with the known theory in Figure 4(a). It is thus required that e.g. in the beginning the curve representing the marker CD34 must be low followed by a part where it is gradually increasing to a high value.

As input, the algorithm takes one starting cell (a LT-HSC cell) and two end cells (a CLP and CMP cell). These are chosen by relying on the theoretical feature values and are indicated for us by the biologists. The result of the algorithm is given in Figure 4(b). On top, we see the density of the different cell states along the calculated trajectories. From this, we can can conclude that the algorithm takes the correct cells for each trajectory because for the trajectory going to the end cell representing CLP's, the cell states LT-HSC, ST-HSC and CLP are traversed consecutively and for the trajectory corresponding to the end cell representing CMP's, the cell states LT-HSC, ST-HSC and CMP are traversed consecutively. This corresponds with the theoretical trajectory in Figure 4(a). When looking at the bottom of Figure 4(b), we can also conclude that the trajectory is correct because in the beginning we see for example an increase in the feature CD34 corresponding to the transition from LT-HSC to ST-HSC as indicated in Figure 4(a). As the cells are assigned to the correct trajectory and they are ordered in the correct way, we can conclude that our algorithm is able to infer these trajectories in an automatic way.

Note that in the top left part of Figure 4(b) the length of the interval where the CLP



(a) (b)

Figure 4: Part (a) shows the theoretical branched trajectory for early hematopoietic stem cells. Part (b) shows the density of the different cell types and the marker values (parameters) of the cells along the trajectory going to CLP's (resp. top and bottom left), and to CMP's (resp. top and bottom right). The X-axis shows all 4647 cells ordered from the most immature to the most mature state.

density is high, is very small in comparison to the length of the interval where the ST-HSC density is high. This has the simple reason that there are just far more ST-HSC's present in the data than CLP's (2192 versus 122). This phenomenon is also present in the bottom left of Figure 4(b) where the feature values for CLP's are only shown in a short interval. A nice extension would thus be to automatically detect the different cell states and rescale the axis.

4 Conclusion

In this paper, we proposed a graph-based approach to infer branched trajectories from mixtures of single cell data. Our approach extends the Wanderlust algorithm [4] that was only able to model a single trajectory. Our new algorithm now allows researchers to model multiple, branched trajectories, and the first results on a case study of early developping hematopoietic stem cells confirms that our algorithm is able to infer existing knowledge about cell differentiation in an automatic way. These results pave the way towards other applications and more complex differentiation scenarios, likely leading to novel biological insights.

Acknowledgments

We would like to thank Lianne van de Laar and Bart Lambrecht for providing the hematopoietic stem cell dataset. Sofie Van Gassen is funded by the Flanders Agency for Innovation by Science and Technology (IWT).

References

[1] P. Qiu, E.F. Simonds, S.C. Bendall, K.D. Gibbs Jr, R.V. Bruggner, M.D. Linderman, K. Sachs, G.P. Nolan and S.K. Plevritis. "Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE". *Nature biotechnology*, vol.29, no.10, pp. 886–891, 2011.

[2] E.D. Amir, K.L. Davis, M.D. Tadmor, E.F. Simonds, J.H. Levine, S.C. Bendall, D.K. Shenfeld, S. Krishnaswamy, G.P. Nolan and D. Pe'er. "viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia". *Nature biotechnology*, vol.36, no.6, pp. 545-552, 2013.

[3] S. Van Gassen, B. Callebaut, M.J. Van Helden, B.N. Lambrecht, P. Demeester, T. Dhaene, and Y. Saeys. "FlowSOM: Using selforganizing maps for visualization and interpretation of cytometry data". *Cytometry Part A*, 2015.

[4] Bendall, Sean C., et al. "Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development." *Cell* 157.3 (2014): 714-725.

[5] Trapnell, Cole, et al. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells." *Nature biotechnology* (2014).

[6] Seita, J., and Weissman, I. L. (2010). Hematopoietic stem cell: selfrenewal versus differentiation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(6), 640-653.

Phosphoproteomics: A critical view through the bioinformatics lens

Panayotis Vlastaridis[1], Stephen G. Oliver[2], Yves Van de Peer[3], Grigoris D. Amoutzias[1]*

1. Department of Biochemistry and Biotechnology, University of Thessaly, Greece
2. Department of Biochemistry, University of Cambridge, UK.
3. Department of Plant Systems Biology, VIB/UGent, Belgium

Correspondence to: amoutzias@bio.uth.gr

Keywords: Phosphoproteomics, phosphorylation, bioinformatics, data integration

Abstract

The advent of high-throughput (HTP) phosphoproteomics in the last decade has revolutionized the field, where hundreds or even thousands of phosphorylation sites (p-sites) are detected in a single experiment. This success is attributed to a combination of very sensitive Mass Spectrometry instruments, better phosphopeptide enrichment techniques and bioinformatics software that are capable of detecting peptides and localizing p-sites. These new technologies open up a whole level of gene regulation to be studied, with great potential for therapeutics and synthetic biology. Nevertheless, many challenges remain to be resolved, concerning the biases and noise of these omic technologies, the biological noise that is present as well as the incompleteness of the current datasets. Despite all the above problems, the currently published datasets seem to be a good sample of a complete phosphoproteome that are capable of revealing its major properties.

The biological significance of protein phosphorylation

To understand a biological system, it is not enough to know which molecules are expressed in various conditions/states. Recent advances in high-throughput proteomics and phosphoproteomics have highlighted the importance of knowing whether the expressed proteins have their molecular functions turned on/off via post-translational modifications. Phosphorylation is the most abundant reversible post-translational modification (PTM) (Krüger et al., 2006), that effectively functions as a digital switch or as a rheostat, regulating one or more functions in a protein, such as enzyme activity, subcellular localization, complex formation and degradation, among others. These effects are mediated via allosteric or orthosteric effects, such as conformational changes, regulation of order/disorder transitions, affinity change on molecular interaction surfaces (Nishi et al., 2014). It is also a key component of signal transduction. More than one switches of this kind may be present in a protein and they may be independent of each other or there may be interdependencies among them or even with other types of switches (Cohen, 2000). Previous and recent findings (Amoutzias et al., 2012; Cohen, 2002; Sadowski et al., 2013) estimate that 1/3 – 2/3 of the proteins in a eukaryotic genome are expected to be phosphorylated, whereas a protein may have only one or even up to tens of p-sites. Therefore, the combinatorics behind this process as well as the potential for complexity at the molecular level are enormous.

Mutation of only one site of phosphorylation in a key protein may have dramatic effects not only for the function of that specific protein, but also for the pathways that it is involved in, or even for the phenotype of the organism (Amoutzias et al., 2006; Papadopoulou et al., 2004). For example a point mutation in cdc28 (S42->A) results in decrease of cell size, whereas mutation of a another p-site may even be lethal, within a certain genetic context, or it may rescue the lethal effect of another point mutation (Zhang et al., 2005).

Abnormal protein phosphorylation is involved in many diseases, such as cancer, diabetes, autoimmune, cardiovascular, neurodegenerative diseases, among others (Van Eyk, 2011; Gaestel et al., 2009; Iwai et al., 2010; Lim, 2005; Tan et al., 2009; Xia et al., 2008). New generations of drugs in cancer and other diseases target this PTM, whereas there is intense interest in measuring serum or blood phosphoproteomes, for improved diagnostics (Khadjavi et al., 2011). Furthermore, many Bacteria disrupt the host immune system by interfering at the phosphorylation networks of the host (Jers et al., 2008), whereas many viruses rely on host kinases to phosphorylate and regulate their proteins (Schwartz and Church, 2010).

Therefore, phosphorylation appears as an extremely attractive area of research not only for understanding organismal complexity or how the cell is regulated, but also for therapeutics and even for synthetic biology. It holds the promise of manipulating molecular pathways and phenotypes, by modifying a small number of phosphorylation sites, via a few point mutations.

The advent of high-throughput (HTP) phosphoproteomics in the last decade has revolutionized the field, where hundreds or even thousands of phosphorylation sites (p-sites) are detected in a single experiment. This success is attributed to a combination of very sensitive Mass Spectrometry instruments, better phosphopeptide enrichment techniques and bioinformatics software that are capable of detecting peptides and localizing p-sites (Doll and Burlingame, 2015; Engholm-Keller and Larsen, 2013; Olsen and Mann, 2013). Nevertheless, many challenges remain to be resolved.

The challenges of phosphoproteomics

Biological noise and technical problems

A major challenge relates to the noise and quality of the generated phosphoproteomic datasets. As with any new HTP technology, data are afflicted by experimental biases and noise. The various phosphopeptide enrichment techniques capture a slice of the complete phosphoproteome and they also introduce biases (Bodenmiller et al., 2007). (Lienhard, 2008) has raised the possibility that, due to the high sensitivity of these MS instruments, biologically noisy p-sites are being detected. ‘Biological noise’, in this case, represents phosphorylation events occurring in degenerate motifs by non-cognate kinases; frequent (but of low abundance) off-target phosphorylations. Also, during the cell-lysation process, kinases and scaffold proteins may encounter target proteins from different cellular compartments, that would not meet under normal conditions. More concern is raised by the observed low occupancy (~10%) of the majority of phosphosites for a given condition (Olsen et al., 2010). In addition, less than 20% of p-sites identified in a single phosphoproteomic experiment are up/down-regulated when a perturbation occurs (Soufi et al., 2009). (Landry et al., 2009) exploited evolutionary information to estimate that up

to 65% of p-sites in HTP experiments could be non-functional, indicating that biological noise may actually be a significant problem. Very recently and in accordance with (Landry et al., 2009; Lienhard, 2008), it was demonstrated that within a compendium of 12 HTP phosphoproteomic experiments from yeast, more than half of non-redundant p-sites were identified only once, further highlighting the problem of potential false positive or non-functional p-sites in HTP datasets (Amoutzias et al., 2012).

Another concern relates to the stringency of the criteria and algorithms used to identify phosphopeptides and to correctly localize p-sites within a phosphopeptide. Some databases, bioinformatics analyses or even prediction tools extract phosphorylation sites from supplementary material of publications without applying very stringent criteria, they mostly rely on the criteria set by each publication, which are not uniform. The general drive to publish phosphoproteomic datasets with as many p-sites as possible means that not very stringent filtering criteria are applied in some of the original publications. Nevertheless, in the last few years, this problem has been ameliorated, as more software have appeared that try to detect phosphopeptides and also localize the phosphosite, by either estimating the phosphosite correct localization probability or the Search engine difference scores (Lee et al., 2015). In parallel, more recent studies have started to adopt stringent criteria, with a cutoff of 99% probability of correct peptide identification and 99% probability of correct phosphorylation site localization.

The bioinformatics analysis of 12 phosphoproteomic yeast datasets revealed that the phosphoprotein and phosphosite overlap between two experiments from two different research groups in very similar conditions (alpha-factor treated yeast cells) was 31% and 11% respectively, whereas the overlap between two experiments of one research group in two different phases of the yeast cell-cycle were 54% and 28% respectively (Amoutzias et al., 2012; Gruhler et al., 2005; Holt et al., 2009; Li et al., 2007). In the same direction, a 2010 study by the Proteome Informatics Research group from ABRF showed that for the same phosphoproteomic dataset, the average agreement of identified phosphoproteins and phosphosites by any two different software was ~57% and ~38% respectively (Lee et al., 2015). Clearly, the detection of phosphoproteins and phosphopeptides is still a very much protocol dependent issue.

The coming of even more phosphoproteomic datasets for a given species, in combination with more sensitive instruments, better localization software and comparative phosphoproteomics will help filter out noisy p-sites.

Incompleteness of the datasets.

For the best studied unicellular eukaryote that harbors only ~6.000 proteins, the baker's yeast, a plethora of HTP phosphoproteomic experiments (performed under a reasonably wide range of conditions) has probably revealed the majority of proteins (~3800) that are regulated at some stage by phosphorylation (Amoutzias et al., 2012; Sadowski et al., 2013). Another analysis estimated that high-throughput phosphoproteomic studies have revealed about 80-90% of the unicellular yeast S. cerevisiae phosphoproteins (Beltrao et al., 2009). Yet, we are far away from identifying the majority of p-sites in that relatively simple organism. In an updated compendium of yeast phosphosites, (Sadowski et al., 2013) found that 45% of low-throughput identified psites were also identified by at least two independent high-throughput experiments. There are also many reports in literature where a well known psite was not detectable by high-throughput technologies. In the

unicellular yeast, more than 70% of its whole proteome is detectable by MS/MS technology in a single experiment (de Godoy et al., 2008; Wu et al., 2011). Clearly, for a multicellular organism such as *Homo sapiens*, with a much more complex proteome and more transient expression patterns, the identification of its entire phosphoproteome, estimated in the hundreds of thousands of p-sites, is not to be expected in the very near future. Nevertheless, comparative phosphoproteomics from several closely related species will help us estimate how much more is missing.

The problems of evolutionary analyses

Another disturbing finding that complicates especially the evolutionary analyses of p-sites is that the precise positioning of p-sites is not always required for proper regulation (Landry et al., 2014; Moses et al., 2007). The implication is that multiple alignments of orthologous proteins are not sufficient to fully determine the conservation of a p-site in another organism. It may be the case that the phosphorylated amino acid is not conserved in another organism, but an equivalent p-site has emerged in the vicinity. Therefore, it is not enough to have the phosphoproteome of one reference organism and the multiple alignments of orthologous proteins.

The problems of predicting phosphorylation sites

The flood of HTP phosphoproteomic data has stimulated research in the field of predicting p-sites from amino acid sequence alone, or in combination with structural and other types of information (Iakoucheva et al., 2004; Ingrell et al., 2007; Mok et al., 2010; Schwartz and Church, 2010). More than 40 prediction methods have been published on this computational problem, applying artificial neural networks, support vector machines, decision trees, genetic algorithms or position specific scoring matrices, whereas a plethora of databases also exists (see two extensive reviews on this subject, Xue et al., 2010; Trost and Kusalik, 2011). There are still ongoing discussions on what is the optimal size of the sequence region around the p-site that contains enough information, without decreasing signal/noise ratio and still remaining computationally tractable for the machine learning algorithms to analyze. A crucial issue is the training datasets used for these algorithms. Abundant and high quality p-sites, as well as very good negative datasets are needed for successful implementation. Nevertheless, especially the negative datasets are very difficult to obtain, since a large fraction of the phosphoproteome of an organism remains unknown. Also, gold-standard reference datasets (both positive and negative) are needed by the community to evaluate any new algorithm/tool and compare it to existing ones. So far, the datasets used to train such algorithms suffer from noisy p-sites, poor filtering of technical and biological noise, while they do not account for the very recent finding that kinases may mistakenly phosphorylate a serine that is very close to the cognate site (Amoutzias et al., 2012; Landry et al., 2009; Lienhard, 2008).

Biological properties of the phosphoproteome

Despite all the above problems that exist, many studies have tried to understand the properties of the best studied phosphoproteome of a model organism, *S. cerevisiae*, either from only one or from a compendium of filtered datasets. In yeast, phosphorylation occurs most frequently in serines (81%), then in threonines (17%) whereas Tyrosines are very rarely phosphorylated (2%) (Amoutzias et al., 2012) probably due to the lack of

Tyrosine specific kinases and the presence of Threonine/Tyrosine dual specificity kinases (Manning et al., 2002; Li et al., 2007). About 90% of p-sites are identified within intrinsically disordered regions, whereas between 12-17% of psites are identified either within or in the vicinity (10 amino acids) of a conserved and characterized domain (Amoutzias et al., 2012; Iakoucheva et al., 2004). Most of the identified phosphoproteins have a small number of psites, whereas there exists a very small number of proteins with many psites. Phosphoproteins are more ancient and are under tighter regulatory control, with shorter protein half-lives, more ubiquitination, more genetic interactions and more protein-protein interactions (Amoutzias et al., 2012; Chi et al., 2007; Yachie et al., 2011). In addition, the number of phosphorylation sites in a protein has an impact on the evolution and survival of duplicated genes, especially after whole genome duplications (Amoutzias et al., 2010).

Interestingly, there is no strong correlation between the number of kinases targeting a phosphoprotein and the number of psites in that protein. A kinase may phosphorylate more than one psites in the same protein or a psite may be phosphorylated by more than one closely related kinases (Amoutzias et al., 2012; Mok et al., 2010) . In addition, p-sites tend to cluster together and this is not due to false detection/localization of psites (Amoutzias et al., 2012; Moses et al., 2007; Schweiger and Linial, 2010).

Despite the noise that is present in the current phosphoproteomic datasets and their acknowledged incompleteness, the conclusions of computational analyses done so far with limited datasets are not necessarily invalid. As demonstrated in (Amoutzias et al., 2012b), several conclusions remain robust, even when creating large and high-quality compendiums of p-sites. Nevertheless, efficient filtering of noise will substantially increase confidence and resolution in the results and computational analyses and will allow new discoveries.

Acknowledgements

This work is supported and implemented under the "ARISTEIA II" Action of the "OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING" and is co-funded by the European Social Fund (ESF) and National Resources (code 4288 to G.D.A). GDA acknowledges additional support by research grants from the Postgraduate Programme ‘Applications of Molecular Biology-Genetics, Diagnostic Biomarkers’, code 3817 of the University of Thessaly, School of Health Sciences, Department of Biochemistry & Biotechnology.

Bibliography

Amoutzias, G.D., Bornberg-Bauer, E., Oliver, S.G., and Robertson, D.L. (2006). Reduction/oxidation-phosphorylation control of DNA binding in the bZIP dimerization network. BMC Genomics *7*, 107.

Amoutzias, G.D., He, Y., Gordon, J., Mossialos, D., Oliver, S.G., and Van de Peer, Y. (2010). Posttranslational regulation impacts the fate of duplicated genes. Proc. Natl. Acad. Sci. U. S. A. *107*, 2967–2971.

Amoutzias, G.D., He, Y., Lilley, K.S., Van de Peer, Y., and Oliver, S.G. (2012a). Evaluation and properties of the budding yeast phosphoproteome. Mol. Cell. Proteomics MCP *11*, M111.009555.

Beltrao, P., Trinidad, J.C., Fiedler, D., Roguev, A., Lim, W.A., Shokat, K.M., Burlingame, A.L., and Krogan, N.J. (2009). Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. PLoS Biol. *7*, e1000134.

Bodenmiller, B., Mueller, L.N., Mueller, M., Domon, B., and Aebersold, R. (2007). Reproducible isolation of distinct, overlapping segments of the phosphoproteome. Nat. Methods *4*, 231–237.

Chi, A., Huttenhower, C., Geer, L.Y., Coon, J.J., Syka, J.E.P., Bai, D.L., Shabanowitz, J., Burke, D.J., Troyanskaya, O.G., and Hunt, D.F. (2007). Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. *104*, 2193–2198.

Cohen, P. (2000). The regulation of protein function by multisite phosphorylation--a 25 year update. Trends Biochem. Sci. *25*, 596–601.

Cohen, P. (2002). The origins of protein phosphorylation. Nat. Cell Biol. *4*, E127–E130.

Doll, S., and Burlingame, A.L. (2015). Mass spectrometry-based detection and assignment of protein posttranslational modifications. ACS Chem. Biol. *10*, 63–71.

Engholm-Keller, K., and Larsen, M.R. (2013). Technologies and challenges in large-scale phosphoproteomics. Proteomics *13*, 910–931.

Van Eyk, J.E. (2011). Overview: the maturing of proteomics in cardiovascular research. Circ. Res. *108*, 490–498.

Gaestel, M., Kotlyarov, A., and Kracht, M. (2009). Targeting innate immunity protein kinase signalling in inflammation. Nat. Rev. Drug Discov. *8*, 480–499.

De Godoy, L.M.F., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251–1254.

Gruhler, A., Olsen, J.V., Mohammed, S., Mortensen, P., Færgeman, N.J., Mann, M., and Jensen, O.N. (2005a). Quantitative Phosphoproteomics Applied to the Yeast Pheromone Signaling Pathway. Mol. Cell. Proteomics *4*, 310–327.

Holt, L.J., Tuch, B.B., Villén, J., Johnson, A.D., Gygi, S.P., and Morgan, D.O. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. Science *325*, 1682–1686.

Huber, A., Bodenmiller, B., Uotila, A., Stahl, M., Wanka, S., Gerrits, B., Aebersold, R., and Loewith, R. (2009). Characterization of the Rapamycin-Sensitive Phosphoproteome Reveals That Sch9 Is a Central Coordinator of Protein Synthesis. Genes Dev. *23*, 1929–1943.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004a). The Importance of Intrinsic Disorder for Protein Phosphorylation. Nucleic Acids Res. *32*, 1037–1049.

Ingrell, C.R., Miller, M.L., Jensen, O.N., and Blom, N. (2007). NetPhosYeast: Prediction of Protein Phosphorylation Sites in Yeast. Bioinformatics *23*, 895–897.

Iwai, L.K., Benoist, C., Mathis, D., and White, F.M. (2010). Quantitative phosphoproteomic analysis of T cell receptor signaling in diabetes prone and resistant mice. J. Proteome Res. *9*, 3135–3145.

Jers, C., Soufi, B., Grangeasse, C., Deutscher, J., and Mijakovic, I. (2008). Phosphoproteomics in bacteria: towards a systemic understanding of bacterial phosphorylation networks. Expert Rev. Proteomics *5*, 619–627.

Khadjavi, A., Barbero, G., Destefanis, P., Mandili, G., Giribaldi, G., Mannu, F., Pantaleo, A., Ceruti, C., Bosio, A., Rolle, L., et al. (2011). Evidence of Abnormal Tyrosine Phosphorylated Proteins in the Urine of Patients With Bladder Cancer: The Road Toward a New Diagnostic Tool? J. Urol. *185*, 1922–1929.

Krüger, R., Kübler, D., Pallissé, R., Burkovski, A., and Lehmann, W.D. (2006). Protein and Proteome Phosphorylation Stoichiometry Analysis by Element Mass Spectrometry. Anal Chem *78*, 1987–1994.

Landry, C.R., Levy, E.D., and Michnick, S.W. (2009). Weak functional constraints on phosphoproteomes. Trends Genet. TIG *25*, 193–197.

Landry, C.R., Freschi, L., Zarin, T., and Moses, A.M. (2014). Turnover of protein phosphorylation evolving under stabilizing selection. Front. Genet. *5*, 245.

Lee, D.C.H., Jones, A.R., and Hubbard, S.J. (2015). Computational phosphoproteomics: from identification to localization. Proteomics *15*, 950–963.

Li, X., Gerber, S.A., Rudner, A.D., Beausoleil, S.A., Haas, W., Villén, J., Elias, J.E., and Gygi, S.P. (2007). Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. J. Proteome Res. *6*, 1190–1197.

Lienhard, G.E. (2008). Non-functional phosphorylations? Trends Biochem. Sci. *33*, 351–352.

Lim, Y.P. (2005). Mining the Tumor Phosphoproteome for Cancer Markers. Clin. Cancer Res. *11*, 3163–3169.

Manning, G., Plowman, G.D., Hunter, T., and Sudarsanam, S. (2002). Evolution of protein kinase signaling from yeast to man. Trends Biochem. Sci. *27*, 514–520.

Metodiev, M., and Alldridge, L. (2008). Phosphoproteomics: A possible route to novel biomarkers of breast cancer. Proteomics Clin. Appl. *2*, 181–194.

Mok, J., Kim, P.M., Lam, H.Y.K., Piccirillo, S., Zhou, X., Jeschke, G.R., Sheridan, D.L., Parker, S.A., Desai, V., Jwa, M., et al. (2010a). Deciphering Protein Kinase Specificity Through Large-Scale Analysis of Yeast Phosphorylation Site Motifs. Sci Signal *3*, ra12.

Moses, A.M., Hériché, J.-K., and Durbin, R. (2007). Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. Genome Biol. *8*, R23.

Nishi, H., Shaytan, A., and Panchenko, A.R. (2014). Physicochemical mechanisms of protein regulation by phosphorylation. Front. Genet. *5*, 270.

Olsen, J.V., and Mann, M. (2013). Status of large-scale analysis of post-translational modifications by mass spectrometry. Mol. Cell. Proteomics MCP *12*, 3444–3452.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A., et al. (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci. Signal. *3*, ra3.

Papadopoulou, N., Chen, J., Randeva, H.S., Levine, M.A., Hillhouse, E.W., and Grammatopoulos, D.K. (2004a). Protein kinase A-induced negative regulation of the corticotropin-releasing hormone R1alpha receptor-extracellularly regulated kinase signal transduction pathway: the critical role of Ser301 for signaling switch and selectivity. Mol. Endocrinol. Baltim. Md *18*, 624–639.

Sadowski, I., Breitkreutz, B.-J., Stark, C., Su, T.-C., Dahabieh, M., Raithatha, S., Bernhard, W., Oughtred, R., Dolinski, K., Barreto, K., et al. (2013). The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update. Database J. Biol. Databases Curation *2013*, bat026.

Schwartz, D., and Church, G.M. (2010). Collection and motif-based prediction of phosphorylation sites in human viruses. Sci. Signal. *3*, rs2.

Schweiger, R., and Linial, M. (2010). Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. Biol. Direct *5*, 6.

Soufi, B., Kelstrup, C.D., Stoehr, G., Fröhlich, F., Walther, T.C., and Olsen, J.V. (2009). Global analysis of the yeast osmotic stress response by quantitative proteomics. Mol. Biosyst. *5*, 1337–1346.

Tan, C.S.H., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M.O., Jørgensen, C., Bader, G.D., Aebersold, R., Pawson, T., and Linding, R. (2009). Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. Sci. Signal. *2*, ra39.

Trost, B., and Kusalik, A. (2011). Computational Prediction of Eukaryotic Phosphorylation Sites. Bioinformatics *27*, 2927–2935.

Wu, R., Dephoure, N., Haas, W., Huttlin, E.L., Zhai, B., Sowa, M.E., and Gygi, S.P. (2011). Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. Mol. Cell. Proteomics MCP *10*, M111.009654.

Xia, Q., Cheng, D., Duong, D.M., Gearing, M., Lah, J.J., Levey, A.I., and Peng, J. (2008). Phosphoproteomic analysis of human brain by calcium phosphate precipitation and mass spectrometry. J. Proteome Res. *7*, 2845–2851.

Xue, Y., Gao, X., Cao, J., Liu, Z., Jin, C., Wen, L., Yao, X., and Ren, J. (2010). A summary of computational resources for protein phosphorylation. Curr. Protein Pept. Sci. *11*, 485–496.

Yachie, N., Saito, R., Sugiyama, N., Tomita, M., and Ishihama, Y. (2011). Integrative features of the yeast phosphoproteome and protein-protein interaction map. PLoS Comput. Biol. *7*, e1001064.

Zhang, K., Lin, W., Latham, J.A., Riefler, G.M., Schumacher, J.M., Chan, C., Tatchell, K., Hawke, D.H., Kobayashi, R., and Dent, S.Y.R. (2005). The Set1 methyltransferase opposes Ipl1 aurora kinase functions in chromosome segregation. Cell *122*, 723–734.

Meta-analysis of human cell cycle-associated transcripts using published data

Bruno Giotti[1] and Tom C. Freeman[1]

[1]Systems Immunology Group, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, Midlothian, UK EH25 9RG

Keywords: microarray, cell cycle, systems biology, network analysis, BioLayout *Express*^{3D}

Abstract

Data from thousands of gene expression studies are now freely available from public data repositories. A useful approach to investigate this multidimensional data is by network analysis. BioLayout *Express*^{3D} is a tool designed to transform small to large unstructured datasets into networks, and to cluster these graphs to identify groups of genes with a similar expression profile and therefore function. The transcriptional network underpinning the cell cycle has been studied extensively in a number of species using microarray technology. Previous analyses of the human cell cycle gene signature have been performed on different human cell lines identifying hundreds of transcripts which demonstrate phase-specific expression. However, comparison of the gene-sets from the different cell lines studies found that there was little overlap in the genes identified. Here we describe the application of a network analysis approach to further investigate this observation by the reanalysis of published human cell cycle data.

Background

Gene expression microarrays have been widely employed for the past decade and more to analyze genome-wide expression patterns for both basic research and clinical studies. Data derived from these experiments have been shared on public databases such as GEO (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). Analysis of such multidimensional data becomes challenging to interpret with standard statistical methods, especially in time-course experiments where pairwise comparison between groups may produce multiple overlapping lists of differentially expressed genes. BioLayout *Express*^{3D} [1] is a tool designed to transform data into network graphs. Graphs are generated by first calculating a (Pearson) correlation matrix that compares the expression profile of every gene to every other gene. Relationships greater than a user-defined threshold are then used to define edges, with nodes representing a given transcript. Network graphs are laid out and rendered in either 2 or 3-dimensional space, and the MCL [2] cluster algorithm is used to identify cliques of highly connected nodes, representing genes with similar expression profiles, which may be involved in a common biological pathway or process.

The transcriptional network underpinning the cell cycle has been studied extensively in a number of species using microarray technology. Previous analyses of the human cell cycle gene signature have been performed on four different human cell lines [3][4][5][6][7]. For these experiments cell populations need to be synchronized which can be achieved in a variety of ways: double thymidine block which stops cells in S-phase, thymidine-nocodazole block which stops cells in M-phase, and by starvation which induces quiescence (G_0) in cells that are then

released by adding back the serum. Each of these studies demonstrated hundreds of transcripts to be expressed in a phase-specific manner (referred to as oscillation when multiple rounds of division were observed), an intrinsic feature of a "cell cycle gene". However, comparisons of the gene lists reported from four studies [3][4][6][7] concluded that there is little similarity in the cell cycle transcriptome across the cell lines employed [6][7]. Observations we have made suggested that this conclusion may not be correct so we decided to investigate further. Here we describe the application of the network analysis approach to investigate this question by the reanalysis of published cell cycle data.

Material and methods

Raw data derived from four cell cycle studies was downloaded from GEO (Acc. numbers: GSE52100, GSE26922) and ArrayExpress (Acc. numbers: E-MTAB-454, E-TABM-263). With the exception of Grant *et al.* data, which was only provided as a processed 'series matrix' file, quality control and RMA normalisation of CEL files was performed using *oligo* and *affy* packages [8]. Multiple probesets mapping to the same HGNC symbol were averaged to one value for each dataset. To adjust for different average intensity between studies ComBat [9], a batch correcting algorithm, was applied. A graph of the resulting dataset was generated with BioLayout *Express*^{3D} with a Pearson correlation threshold of $r = 0.60$ and an MCL inflation value (which controls the granularity of clustering) set to 1.4 and pre-inflation set to 1.6. Enrichment analysis was performed with DAVID web tool [10] using Functional Annotation Clustering and selecting a reference Gene Ontology (GO) Biological Process term for each of the top significant clusters.

Results

To investigate the conservation of the cell cycle signature we downloaded raw data (see Methods) from four microarray gene expression studies [4][5][6][7] derived from the analysis of four different human cell lines: NHDF (primary fibroblasts), HeLa (cervical cancer cells), HaCat (immortal keratinocytes), and U2OS (osteosarcoma cells). Probesets were mapped to HGNC symbols which were then used as a reference for binding measurements across studies. The batch correction algorithm, ComBat [9], was used to normalize expression values across the four datasets. A graph of the data was rendered within BioLayout *Express*^{3D} at a correlation value of $r = 0.60$. The graph was comprised of 708 nodes connected by 7,546 edges. The MCL algorithm was used to cluster the data into groups of genes that share a similar expression pattern. Four major clusters of genes were identified (Fig. 1A). Cluster 1 comprised of 222 transcripts, many of which were well known S-phase-related cell cycle genes including: cyclins D and E (*CCND3*, *CCNE1* and *CCNE2*), *CDC25A*, *CDC6*, various polymerases (*POLA1*, *POLA2*, *POD1*, *POLD3*, *POLE*, *POLE2*), *PCNA* and other protein complexes necessary to DNA synthesis e.g. members of the GINSs, MCMs, and RFCs protein families. Cluster 2 accounted for 142 genes of which several G_2 and mitotic regulators such as *BUB1*, *BIRC5*, cyclins A and B (*CCNA2*, *CCNB1*, *CCNB2*), *CDC25B* and *CDC25C*, various kinetochore proteins (*CENPA*, *CENPE*, *CENPF*, *CENPI*) etc. These clusters including 364 genes showed consistent peaks of expression at each cell cycle event across the four studies with Cluster 1 being up-regulated consistently before Cluster 2 (Fig 1B, green and red profiles, arrows). GO enrichment analysis performed with DAVID (see methods) and further confirmed Clusters 1 and 2 to represent S-phase and M-phase-associated gene expression, respectively (Fig. 1C, top histograms). Cluster 3 was comprised of 254 genes that also showed an oscillatory expression similar to that of the genes in

Cluster 2 (Fig .1B, blue profile). Significantly enriched GO terms in the cluster included pathways involved in metabolism and ribosome biogenesis (Fig. 1B, bottom left histogram); these pathways are likely associated with G_1-phase of the cell cycle. Lastly Cluster 4 accounted for 90 transcripts and appeared to have peaks of expression at the very start of each experiment but without peaking at subsequent cell cycle events (Fig. 1B, yellow profile). GO enrichment showed significant enrichment for pro-apoptotic related terms (Fig. 1C, bottom right histogram). It is possible that this cluster reflected a stress-induced response due to synchronisation methods rather than cell cycle-related biology.

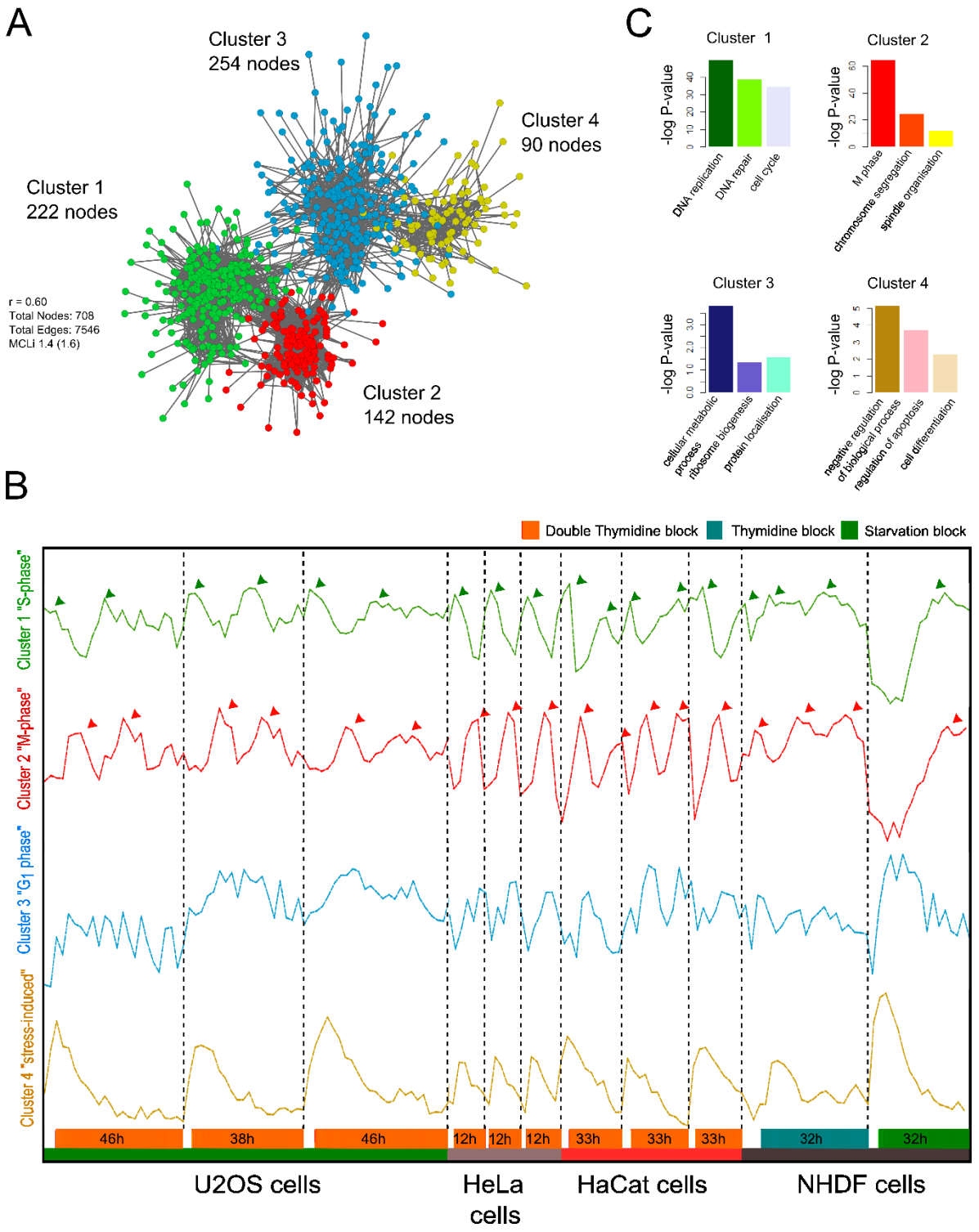


Figure 1. *Meta-analysis of the human cell cycle-associated transcriptional network.* **(A)** Graph generated with Biolayout *Express*^{3D} showing four color-coded clusters. The graph represents the merged datasets derived from four independent studies. **(B)** GO enrichment performed with DAVID of transcripts included in each cluster. **(C)** Average expression profiles of the transcripts included in each cluster. Dashed lines mark different time-course experiments of length specified on the bottom of the chart, time points running left to right for each study. Different synchronization methods are color-coded. Green arrows indicate S-phase, red arrows M-phase.

Conclusion

By querying public databases, merging data from independent studies and analyzing them with BioLayout *Express*^{3D} we were able to identify a core set of cell cycle-regulated genes whose expression is conserved across four human cell lines. Of the 708 transcripts separated into four different clusters three of them, Cluster 1-3, showed consistent peaks of expression per cell cycle event across each cell line. Of these, two clusters were confirmed to be cell cycle-related by Gene Ontology (GO) enrichment: Cluster 1 was enriched in S-phase-related terms and Cluster 2 was enriched in M-phase-related terms. Expression profiles of the two clusters also neatly separated as Cluster 1 was up-regulated consistently before Cluster 2 (Fig. 1B). Cluster 3 was enriched with metabolism and ribosome synthesis-related GO terms (Fig. 1C, Cluster 3 histogram) which are pathways likely to be up-regulated during G_1-phase. However its expression profile was not clearly separated from that of Cluster 2 except for the experiment with cells synchronised by serum-starvation technique. Perhaps G_1-phase related expression become more distinguishable in cells entering cell cycle from G_0 (Fig. 1B, 'starvation block' experiment). Cluster 4 instead could reflect a stress-related gene expression due to synchronisation methods as peaks of expression were seen only at early time-points in each experiment and did not show up-regulation on following cell cycle events. Overall this work suggests that conservation of the cell cycle signature is more prominent than previously observed, even across transformed (HeLa and U2OS) and non-transformed cells (HaCat and NHDF). Indeed Diaz et al. found only 125 cell cycle genes conserved across studies on HeLa [3], NHDF [4] and HaCaT [6] while Grant et al. suggested an even smaller core cell cycle signature of 96 cell cycle genes when data from U2OS cells was added [7]. The large discrepancies between studies previously reported could be explained as gene lists were directly compared without re-analysis of the primary data. Re-analysis of the data is indeed crucial as the original studies used different statistical cut-offs, filtering methods etc. to detect cell cycle genes. Thus, the large discrepancies found between previous studies are likely due to variance in the original analysis methods, rather than true biological differences in the cell cycle transcriptome across cell types.

Reference

[1] T. C. Freeman, L. Goldovsky, M. Brosch, S. van Dongen, P. Mazière, R. J. Grocock, S. Freilich, J. Thornton, and A. J. Enright, 'Construction, Visualisation, and Clustering of Transcription Networks from Microarray Expression Data', *PLoS Comput Biol*, vol. 3, no. 10, p. e206, Oct. 2007.

[2] A. J. Enright, S. V. Dongen, and C. A. Ouzounis, 'An efficient algorithm for large-scale detection of protein families', *Nucleic Acids Res.*, vol. 30, no. 7, pp. 1575–1584, Jan. 2002.

[3] M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, and D. Botstein, 'Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors', *Mol. Biol. Cell*, vol. 13, no. 6, pp. 1977–2000, Jan. 2002.

[4] Z. Bar-Joseph, Z. Siegfried, M. Brandeis, B. Brors, Y. Lu, R. Eils, B. D. Dynlacht, and I. Simon, 'Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells', *Proc. Natl. Acad. Sci.*, vol. 105, no. 3, p. 955, 2008.
[5] S. Sadasivam, S. Duan, and J. A. DeCaprio, 'The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression', *Genes Dev.*, vol. 26, no. 5, pp. 474–489, Jan. 2012.
[6] J. Peña-Diaz, S. A. Hegre, E. Anderssen, P. A. Aas, R. Mjelle, G. D. Gilfillan, R. Lyle, F. Drabløs, H. E. Krokan, and P. Sætrom, 'Transcription profiling during the cell cycle shows that a subset of Polycomb-targeted genes is upregulated during DNA replication', *Nucleic Acids Res.*, vol. 41, no. 5, pp. 2846–2856, Jan. 2013.
[7] G. D. Grant, L. Brooks, X. Zhang, J. M. Mahoney, V. Martyanov, T. A. Wood, G. Sherlock, C. Cheng, and M. L. Whitfield, 'Identification of cell cycle–regulated genes periodically expressed in U2OS cells and their regulation by FOXM1 and E2F transcription factors', *Mol. Biol. Cell*, vol. 24, no. 23, pp. 3634–3650, Jan. 2013.
[8] B. S. Carvalho and R. A. Irizarry, 'A framework for oligonucleotide microarray preprocessing', *Bioinformatics*, vol. 26, no. 19, pp. 2363–2367, Jan. 2010.
[9] W. E. Johnson, C. Li, and A. Rabinovic, 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007.
[10] D. W. Huang, B. T. Sherman, and R. A. Lempicki, 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, Dec. 2008.

D-Optimal Designs: Differences in HIV risk profiles between Gen X black women and entire population of black women attending antenatal clinics in South Africa

Wilbert Sibanda
School of Information Technology
North-West University, Vaal Triangle campus
Van Eck Blvd, Vanderbijlpark, South Africa, 1900

Abstract— **A D-optimal design was used to compare the effects of demographic characteristics on the risk of HIV infection amongst Gen X black women and the entire population of black women using ten year annual antenatal HIV seroprevalence data collected in antenatal clinics across the nine provinces of South Africa during the period 2001 to 2010. The data was fitted in MODDE 10.1™ using partial least squares (PLS). The results of the study showed that an increase in the age of the pregnant woman resulted in a corresponding increase in the risk of HIV infection amongst the two populations studied. However, an improvement in educational level of the entire population of black women resulted in a decrease in HIV risk. It was also observed that a pregnant woman's transition from syphilis negative to positive resulted in a significant increase in risk of HIV infection.**

Keywords—D-optimal, Partial Least Squares, HIV risk, Gen X, South Africa

1. Scientific Background

a. Introduction

The National Department of Health of South Africa uses the annual antenatal HIV seroprevalence survey to monitor the changes in HIV prevalence rates across the nine provinces of the country. This HIV surveillance technique was started in 1990 and is the most elaborate in Africa south of the Sahara. The annual antenatal HIV survey is the only existing national surveillance activity for determining HIV prevalence in South Africa and is therefore a vitally important tool to track the geographic and spatial trends of the epidemic [1]. The demographic characteristics captured for each pregnant woman are the pregnant woman's age (herein called mothage), male sexual partner's age (herein fathage), race, pregnant woman's educational level, gravidity (number of pregnancies), parity (number of children born), name of clinic, HIV and syphilis infection. This study attempts to use a D-optimal design methodology to study HIV risk profiles between Gen X black women and the entire population of black women attending antenatal clinics in South Africa during the period 2001 to 2010.

b. D-Optimal Design

A D-optimal design is a computer generated design, which consists of the best subset of experiments selected from a candidate set. The candidate set is the pool of theoretically possible and practically conceivable experiments. In order to select the best design, the computer evaluates a selection of experimental runs according to a given criterion. The criterion is that the selected design should maximize the determinant of the matrix X'X for a given regression model. The maximization criterion explains the derivation of the letter 'D' in Optimal from D in determinant. The search for the best subset of experiments is carried out using an automatic search algorithm in MODDE 10.1™. The best coefficients are derived according to least squares model in (1).

$$b = (X'X)^{-1}X'Y \quad (1)$$

The three terms used to calculate confidence intervals of coefficients are the inverse of X'X, the residual standard deviation of the model and student's t parameter. It can then be concluded that the smallest $(X'X)^{-1}$ or the largest X'X is beneficial for the precision of the regression coefficients.

The D-optimal selection of experiments may be evaluated by means of several criteria, namely condition number and G-efficiency. The condition number is a measure of sphericity and symmetry of a D-optimal design. It is the ratio of the largest and smallest singular values of the X-matrix. It can also be considered as the ratio of the largest and smallest design diagonals. For an orthogonal design, the condition number is 1, and the higher the number the less orthogonality. The aim is to make the designs as orthogonal as possible.

The second evaluation criterion is the G-efficiency which compares the efficiency, or performance of a D-optimal design to that of a fractional factorial design. G-efficiency is computed as in (2)

$$G_{eff} = 100*p/n*d \quad (2)$$

Where *p* is the number of model terms, *n* is the number of runs in the design, and *d* is the maximum relative prediction variance across the candidate set.

The upper limit of the G_{eff} is 100% which implies that the fractional factorial design was returned by the D-optimal search. A G_{eff} above 60-70% is recommended.

2. Materials and Methods

a. Experimental Data

This study used national annual South African HIV seroprevalence data. The entire dataset for the period between 2001 and 2010 contained 247 843 pregnant women that attended antenatal clinics across the nine provinces of the country. However, only 190 686 individuals with complete records were included in the study.

b. Variables

The variables used in the study were pregnant woman's age (mothage), male sexual partner's age (fathage), gravidity, parity, pregnant woman's educational level, pregnant woman's syphilis and HIV status. The educational level of the pregnant women was classified as no formal education, primary education, secondary and tertiary education. The syphilis and HIV status was binary coded, with 1 representing positive status, and 0 representing a negative status.

c. Experimental Design

In this study, the aim was to use a D-optimal design to study the effects of demographic characteristics on influencing the risk of acquiring HIV infection amongst pregnant women in South Africa. The D-optimal designs are always an option regardless of the type of model to be fitted, such as first order, first order and interactions, quadratic and cubic or the objective specified for experiment such as screening or response surface model. The two advantages of using D-optimal designs over standard classical designs include standard factorial or fractional factorial designs require too many runs for the amount of resources or time allowed for the experiment and classical designs present constrained design spaces, such as design spaces that are not feasible or impossible to run. Partial least squares (PLS) projections to latent structures were used in conjunction with a D-optimal experimental design. PLS regression (PLSR) is a generalization of multiple linear regression (MLR) [2]. Unlike MLR, PLSR can analyze data with strongly correlated, noisy and numerous x-variables. In handling numerous and collinear x and y variables, PLSR enables the investigation of more complex problems in a more realistic way. The goodness-of-fit of the model is given by R^2 and Q^2 (cross-validated R^2). In addition to the above goodness-of-fit measures, the PLSR model has additional diagnostics such as loadings, regression coefficients and variable importance (VIP) measure.

d. PLS Model Interpretation-Loadings

The loading plots display the correlation between demographic characteristics and HIV risk. This is an efficient method for interpreting the PLS model. In model interpretation one considers the distance to the plot origin. The further away from the plot origin a demographic characteristic lies, the stronger the model impact that particular demographic characteristic has. The sign of the PLS loading indicates the correlation amongst the variables.

e. PLS Model Interpretation-Variable Importance (VIP)

The variable importance plot (VIP) represents the contribution of each demographic characteristic in fitting the PLS model for both demographic characteristics and HIV risk. VIP provides a summary of the contribution a demographic characteristic makes to the D-optimal model [2]. VIP is therefore a weighted summary of all loadings across the response. If a demographic characteristic has a relatively small coefficient (in absolute value) and a small value of VIP less than 0.08, it is a prime candidate for deletion from the D-optimal design model [3].

f. Multivariate data analysis

The annual antenatal HIV and syphilis prevalence data was analyzed using MODDE 10.1^{TM} software (Umetrics, Umea, Sweden). As stated above, MODDE 10.1^{TM} was also used to generate the D-optimal experimental design. For all PLS models, the variables were centered and scaled to unit variance. Factors that did not improve the model according to cross-validation were removed before interpretation [4]. All 95% confidence intervals were calculated using MODDE 10.1^{TM}.

g. Selection of factor levels

TABLE I. FACTORS (GEN X)

Factor	Level 1		Level 2		Level 3		Level 4		Level 5	
	Value	*Code*	*Value*	*Code*	*Value*	*Code*	*Value*	*Code*	*Value*	*Code*
Mothage (years)	20-27	-1	28-33	-0.5	34-38	0.5	39-45	1		
Fathage (years)	15-22	-1	23-32	-0.5	33-37	0	38-43	0.5	44-65	1
Gravidity	0-1	-1	2	-0.5	3	0	4	0.5	5-9	1
Parity	0	-1	1	-0.5	2	0	3	0.5	4-8	1
Education	0	-1			1	0			2-3	1
Syphilis	0	-1							1	1

TABLE II. FACTORS (ENTIRE BLACK WOMEN POPULATION IN THE SURVEY)

Factor	Level 1		Level 2		Level 3		Level 4		Level 5	
	Value	*Code*	*Value*	*Code*	*Value*	*Code*	*Value*	*Code*	*Value*	*Code*
Mothage (years)	13-19	-1	20-27	-0.5	28-33	0	34-38	0.5	39-45	1
Fathage (years)	15-22	-1	23-32	-0.5	33-37	0	38-43	0.5	44-65	1
Gravidity	0-1	-1	2	-0.5	3	0	4	0.5	5-9	1
Parity	0	-1	1	-0.5	2	0	3	0.5	4-8	1
Education	0	-1			1	0			2-3	1
Syphilis	0	-1							1	1

3. Results

a. Evaluation of Raw Data using Replicate Plot

The replicate plots look good and show that all experiments lie between the maximum and minimum with no outliers, to be excluded from the experiment. The center-points have limited spread and the values are in the middle of the response range as shown in Figs. 1 and 2.

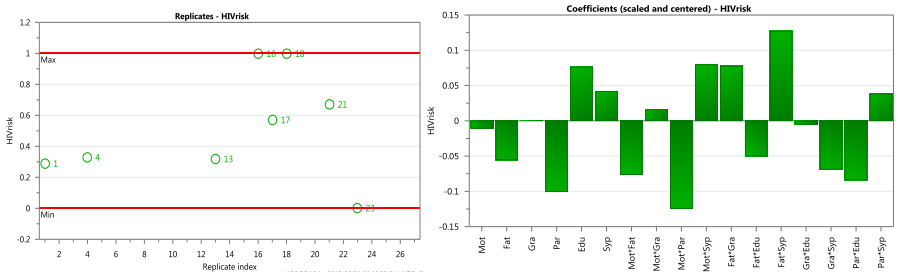


Fig. 1.Model Diagnostics for Gen X black women

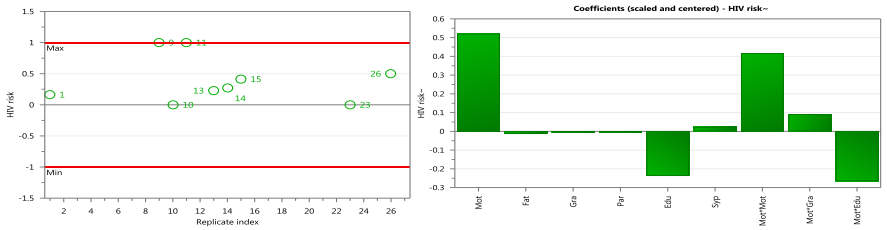


Fig. 2.Model Diagnostics for entire black women population

b. Regression Analysis

i. Summary of Fit

The summary of fit plot derived from Gen X black women indicates a very good model with high R^2 and Q^2 values of 1.00 and 0.82 respectively. An equally good summary of fit plot was obtained for the entire population of black women with R^2 and Q^2 values of 0.99 and 0.75 respectively as shown in Table 3.

TABLE III. SUMMARY OF FIT STATISTICS

Gen X Black women		Entire black women population in the study	
R^2	**Q^2**	**R^2**	**Q^2**
1.00	0.82	0.99	0.75

The data was fitted in MODDE 10.1™ using partial least squares (PLS) fitting. R^2 gives goodness-of-fit of the model, while Q^2 shows goodness-of-prediction of the model and estimates the future prediction precision of the model. In general a Q^2 value greater than 0.1 means a significant model, while a Q^2 value greater than 0.5 suggests a good model. Appropriate model tuning through the removal of non-significant model parameters or selecting the correct transformation results in higher summary statistics. The Q^2 value is considered to be the best and most sensitive model diagnostic tool [5].

c. Model Interpretation

i. Coefficient Plot

The effects of demographic characteristics on HIV risk were represented by coefficient plots in Figs. 1 and 2. The sign of the coefficient determines whether the demographic characteristic should be high or low, the size of the bar indicates the importance of that demographic characteristic in the model. Amongst Gen X an increase in age of the pregnant woman resulted in a decrease in HIV risk. A similar observation was made for the age of the male partner and parity of the pregnant woman. In other words, older women were less at risk of acquiring an HIV infection. This also meant that as the age of the male sexual partners increased, the risk of acquiring HIV infection decreased amongst women attending antenatal clinics in South Africa. An

increase in parity was observed to result in a decrease in risk of HIV infection. However, parity cannot be separated from the age of the pregnant woman. Older woman are likely to have more children, but we have already observed that older women have reduced levels of HIV risk. In addition amongst Gen X women, it was observed that syphilis increased the risk of HIV infection. This is to be expected as syphilis is a sexually transmitted disease and is contracted through unsafe sexual practice. A surprising finding was that an increase in educational level of the pregnant women increased the risk of acquiring HIV infection. This result should not be taken to mean that education does not help in curbing the spread of HIV within populations. However, the above result could be understood in the sense that Gen X women are generally older women amongst antenatal clinic attendees. Based on the history of South Africa, where during apartheid years black people had limited schooling opportunities, an improvement in the level of education is bound to be confined to younger women who are more sexually active. Amongst the entire population of black women the most significant positive effect on HIV risk was found to be the pregnant woman's age. In other words, HIV risk amongst the entire black women population of antenatal clinic attendees increases with an increase in the age of the pregnant woman. Syphilis infection was also found to increase the risk of HIV infection amongst pregnant women. However, amongst the entire population of black women attending antenatal clinics, the woman's educational level was found to reduce the risk of HIV infection as expected. This means older women with low education are especially at risk.

ii. Regression model

The regression coefficients represent the importance of each demographic characteristic in the prediction of the HIV risk.

TABLE IV. HIV RISK REGRESSION MODEL

HIV risk	**Entire black women population**	**Gen X black women**
Constant	1.20	0.50
Mothage	0.52	-0.01
Fathage	-0.01	-0.06
Gravidity	0.01	0.001
Parity	0.006	-0.101
Education	-0.242	0.077
Syphilis	0.024	0.042
Mothage*Mothage	0.42	
Mothage*Gravidity	0.091	0.016
Mothage*Education	-0.265	
Mothage*Fathage		-0.077
Mothage*Parity		-0.125
Mothage*Syphilis		0.080
Fathage*Gravidity		0.078
Fathage*Education		-0.050
Fathage*Syphilis		0.127
Gravidity*Education		-0.001
Gravidity*Syphilis		-0.069
Parity*Education		-0.084
Parity*Syphilis		0.039

Amongst the entire population of black women attending antenatal clinics, the estimated parameters indicate that the pregnant woman's age (mothage) has the greatest positive effect on the risk of HIV infection

at an alpha level of 0.05. The second most influential demographic characteristic is the educational level of the pregnant woman. Therefore the ranked order of the effect of demographic characteristics on HIV risk is as shown in (2);

Mothage>Education>Syphilis>Fathage (2)

An improvement in the educational level of the pregnant woman decreases the risk of acquiring HIV infection. Syphilis infection was found to increase the risk of HIV infection. The interaction of pregnant woman's age and her educational level (mothage*education) decreased the risk of HIV infection. The age of the male sexual partner (fathage) had minimal effect on HIV risk and its removal from the regression equation based on VIP measure increased the Q^2 value to 0.94. Individually, gravidity and parity had very small effect on HIV risk, however, the latter two demographic characteristics were found to have very significant effects on the D-optimal model, based on VIP, because the two variables are highly correlated. It can be concluded that the two variables have the same effect. Even though gravidity had a negligible effect on HIV risk, its interaction with woman's age (mothage*gravidity) had elevated the risk of HIV infection. On that basis, gravidity was not removed from the HIV risk predictive model. The square of the pregnant woman's age (mothage*mothage) was retained in the model, because it provided a high Q^2 value of 0.84. Removal of mothage*mothage interaction from the HIV predictive model reduced the Q^2 value to 0.667. This is an important finding as it shows that the effect of the pregnant woman's age on HIV risk is not linear but quadratic. In contrast, amongst Gen X women the ranked order of the effects of demographic characteristics on the risk of HIV infection was found to be as shown in equation 3.

Parity>Education>Fathage>Syphilis>Mothage (3)

iii. Factor Effects Plots

Factor effects plots display the predicted values of HIV risk, when a selected demographic characteristic varies from its low to high level, all other demographics held constant. It is important to note that a factor plot shows a response back transformed to its original units.

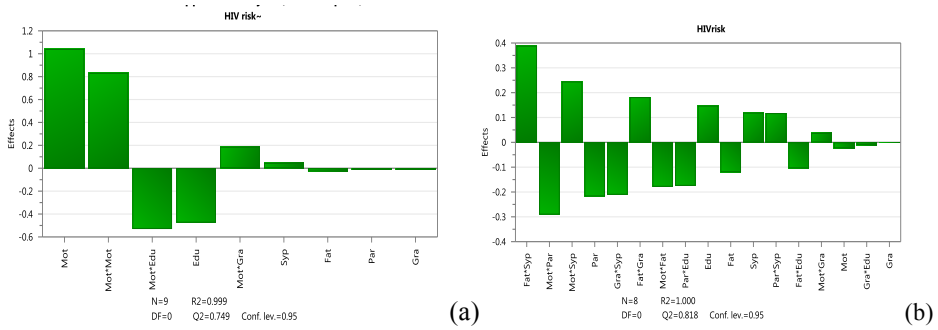


Fig. 3.Factor effects plots for (a) entire black woman population and (b) Gen X black women

The factor effects plot (Fig. 3a) shows that amongst the entire population of black women attending antenatal clinics in South Africa, HIV risk decreases with an improvement in the pregnant woman's educational level from no formal education to tertiary education. As the age of the pregnant women increased from 13 to 45 years, the risk of HIV infection also increased. However, it was observed that a decrease in the age of pregnant women amongst Gen X women resulted in a corresponding decrease in risk of HIV infection. The transition from syphilis negative to positive resulted in an increase in risk of HIV infection. The study also showed that amongst the two women populations, a decrease in the age of the male sexual partner resulted in a corresponding decrease in risk of HIV infection. The latter observation was more marked amongst Gen X black women compared to the entire population of black women. Decreases in gravidity and parity were found to result in a corresponding decrease in risk of HIV infection amongst both women populations. Once more the latter observations were significantly marked amongst Gen X black women as shown in Fig. 3a.

iv. Interaction Effects in 2^2 case

The regression coefficient plot tends to be cluttered and hard to interpret. Fortunately, there is another tool to facilitate model interpretation. This is the interaction plot. Main effects and two-factor interactions have different impacts on the appearance of a semi-empirical model. The main effects make the surface shape and two-factor interactions cause it to twist. In addition, it is possible to create interaction plots specifically exploring the nature of interactions. These plots can be thought of as representing edges of a response surface plot as shown in Fig. 4.

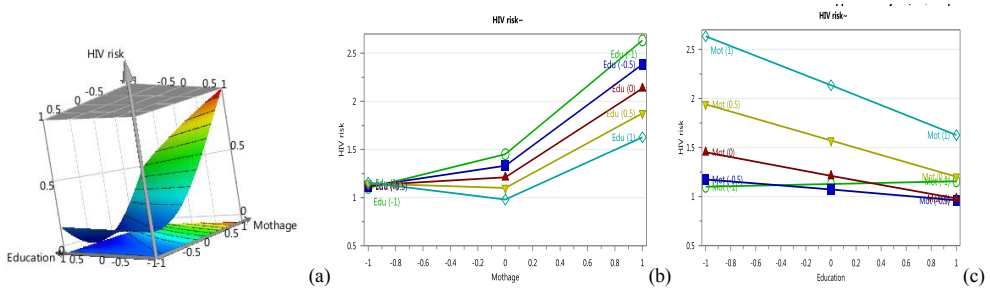


Fig. 4. Interaction plot of mothage and educational level amongst the entire population of black women attending antenatal clinics.

The interaction plot in Fig. 4 confirms that an increase in the woman's age results in an increase in risk of HIV infection. However, the increase in the influence of the woman's age on HIV risk is greater if the woman has no formal education. The increase in HIV risk was gradual until the ages of 28 to 33 years (coded 0). After the age of 33 years, the increase in HIV risk for women with no formal education was remarkably steep. Women with tertiary education demonstrated a decrease in risk of HIV infection as the age increased to between 28 to 33 years (coded 0). Thereafter, a gradual increase in HIV risk was observed. It can be concluded that the effect of the pregnant woman's age on HIV risk is dependent on her educational level as shown in Fig. 3b. Fig. 4c indicates that an increase in the pregnant woman's age increases her risk of HIV infection. However, the influence of educational level on HIV risk is dependent on the woman's age. Older women between the ages of 39 to 45 (coded 1) were associated with the highest level of HIV risk. However, HIV risk amongst this age-group decreased as the educational level improved. In general HIV risk was observed to decrease with a decrease in pregnant woman's age. It can be concluded that the effect of the pregnant woman's educational level on the risk of HIV infection is dependent on her age.

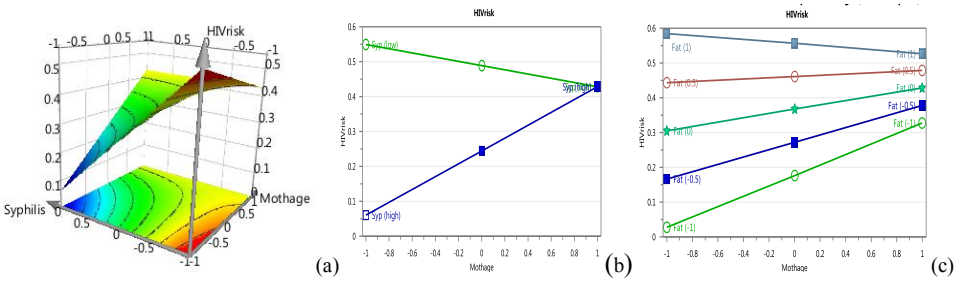


Fig. 5. Interaction Plot of (a) mothage and syphilis and (b) mothage and fathage for Gen X black women.

Fig. 5 shows that amongst syphilis negative (uninfected) women, an increase in the woman's age is associated with a decrease in the risk of HIV infection. However, amongst syphilis infected women an increase in the woman's age is associated with an increase in HIV risk. The most plausible reason for the above observation would be that individuals infected with any sexually transmitted infection (STI) are likely to be risk taking individuals. STIs can only be acquired by indulgence in unprotected sexual practices. This therefore explains the observed increase in HIV risk amongst individuals infected with STIs. The interaction between the pregnant woman's age (mothage) and the male sexual partner's age (fathage) demonstrated an interesting phenomenon. Amongst young women less than 19 years old (coded -1), the highest HIV risk was

associated with male sexual partners older than 44 years old. The lowest HIV risk was observed for women sexually involved with males less than 22 years old. In other words, the younger the women and the younger the male sexual partner, means reduced risk of HIV infection. The HIV risk is observed to increase as the age of the male sexual partner increases. A similar observation was made for all age-groups of women. Also of interest is the fact that as the age of the female increases, the risk of acquiring of acquiring an HIV infection steeply increases for women sexually involved with young males below the age of 22 years (coded -1). The steepness of HIV risk decreases gradually as the age of the male sexual partner increases to the age of 43 years old (coded 0.5). However, for women whose sexual partners were older than 44 years (coded 1) it was observed that an increase in the female's age resulted in a gradual decrease in HIV risk. In other words, even though sexual involvement with males older than 44 years resulted in high risk of HIV infection, it was observed that the risk of HIV infection decreases gradually as the age of the female partner increased.

v. PLS Model Interpretation-Loadings

The loading plots in Fig. 6 display the correlation between the demographic characteristics and risk of HIV infection.

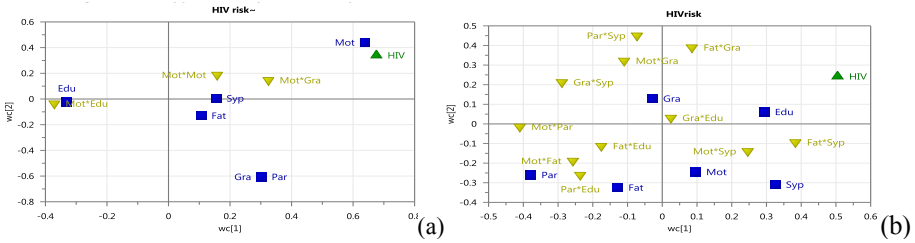


Fig. 6.PLS model loadings for (a) the entire population of black women and (b) Gen X black women attending antenatal clinics.

The proximity of the female' age (mothage) to the response (HIV risk) in Fig. 6(a), illustrates that amongst the entire population of black African women, the age of the pregnant woman is the most positively influential variable on the risk of HIV infection. In other words, as the age of the woman increases, so does the risk of HIV infection. Educational level has the most negative effect on HIV risk, implying that an improvement in educational level results in a corresponding decrease in risk of HIV infection. However, amongst Gen X women, the PLS loadings plot (Fig. 6(b)), shows that parity has the most influential negative effect on HIV risk. In other words, as the number of children increases, a corresponding decrease in HIV risk is observed. Syphilis was observed to exhibit the most positive influence on HIV risk. This meant that a change in syphilis status from uninfected to infected, resulted in a significant increase in the risk of HIV infection.

vi. PLS Model Interpretation-Variable Importance (VIP)

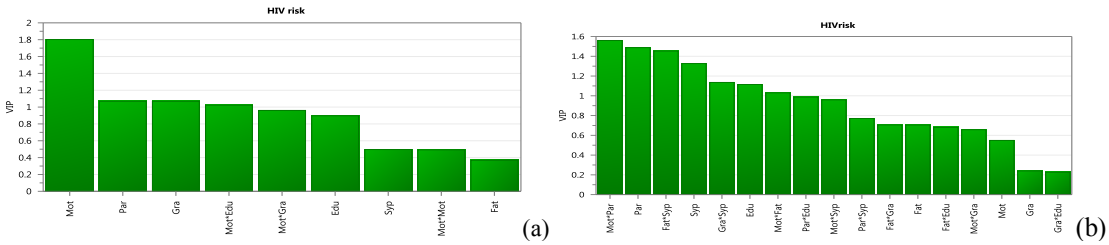


Fig. 7.PLS model VIP for (a) the entire black women population and (b) Gen X black women population

The study once more illustrated that amongst the entire black women population attending antenatal clinics in South Africa, the age of the woman was pivotal in the prediction of the risk of HIV infection. It was also demonstrated that the pregnant woman's age had a significant effect on the D-optimal model, with a VIP value greater than 0.8. However, gravidity and parity had little contribution to HIV risk, but significant effect

on the model with VIP estimate of 1.1. Therefore, gravidity and parity were retained in the D-optimal design model. Syphilis had a small but significant positive effect on HIV risk, though the VIP estimate less than 0.8 as shown in Fig. 7 (a). This makes syphilis a good candidate for exclusion from the predictive model.

The male sexual partner's age (fathage) had a very small and negative effect on the risk of HIV infection amongst the entire population of pregnant black women. Furthermore the age of the male sexual partner had an equally very small VIP less than 0.8, making the variable another candidate for exclusion from the model. The two-factor interactions of mothage*education and mothage*gravidity had significant effects on HIV risk as shown by high VIP values greater than 0.8, as shown in Fig. 7(a). The latter two-factor interactions deserved to be retained in the model. The educational level amongst the entire population of black women had the greatest negative effect on HIV risk. In other words as the educational level of the pregnant woman improves, the HIV risk is observed to decrease. The most significant two-factor interaction is the pregnant woman's age and her educational level (mothage*education). However, amongst Gen X black women the main effects parity, syphilis and educational level of the pregnant women were found to be both important for the prediction of HIV risk as well as significant effect on the D-optimal design based on the VIP estimates, greater than 0.8. The pregnant woman's age (VIP 0.55), male sexual partner's age (VIP 0.71) and gravidity (VIP 0.23), were found to have little importance as terms in the D-optimal model both with respect to the response HIV risk and other X variables, based on the VIP values. However, the elevated VIP values for the interaction of the above factors with each other and other variables made them paramount to the final D-optimal model and thus were not removed from the model. For example, the two-factor interactions of mothage*parity, fathage*syphilis and gravidity*syphilis ranked as the most important interaction for the D-optimal design. Removal of the fathage, mothage and gravidity would have distorted the hierarchical structure of the model.

d. HIV Risk Prediction Plots

The prediction plots display the predicted values of HIV risk for the low, center and high values of the demographic characteristics.

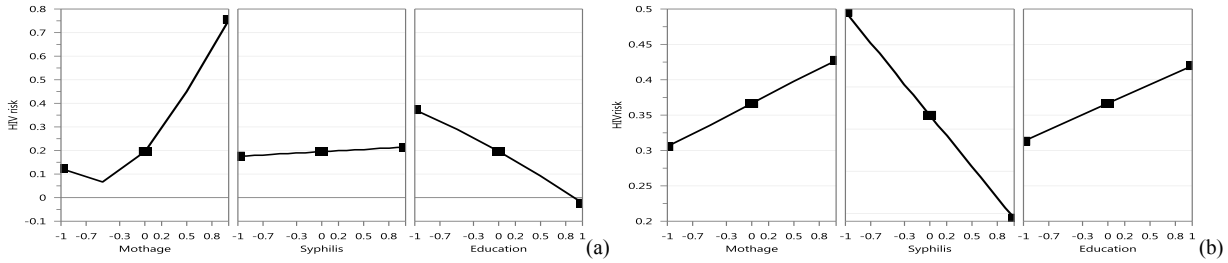


Fig. 8. HIV risk prediction plots for (a) the entire black women population and (b) Gen X black women attending antenatal clinics in South Africa

Amongst the entire population of black women, it was observed that changes in the pregnant woman's age (mothage) had a significant effect on predicted HIV risk, as shown in Fig. 8a. Between, the ages of 20 and 27 years (coded 0.5), an increase in the pregnant woman's age resulted in a 5% decrease in HIV prevalence rate. However, from the age of 28 to 45 years, an increase in age of pregnant woman resulted in a 68% increase in HIV prevalence rate of 68%. The highest HIV risk was predicted at about 75% between the ages of 39 and 45 years (coded 1). This once more illustrates that older women are more susceptible to HIV infection than their younger counterparts. Fig. 8a, also shows that women uninfected with syphilis had the lowest HIV prevalence rates at 18%. However, the risk of HIV infection increased to 21% amongst syphilis infected women. This therefore shows that the transition from syphilis negative to syphilis positive resulted in a 3% increase in risk of HIV infection. It was also observed that amongst the entire population of black women, an improvement in the level of education of the pregnant woman resulted in a decrease in risk of HIV infection. Women with no formal education had the highest risk of HIV infection at 37%. However women with a secondary and tertiary education displayed the lowest risk of HIV infection of about 0%. Therefore, the

improvement in education from no formal education to a secondary or tertiary education resulted in a 37% decrease in risk of HIV infection. Similarly, amongst Gen X black women, an increase in the pregnant woman's age was also observed to result in an increase in the risk of HIV infection of about 11%. However, the transition from syphilis negative to syphilis positive was found to result in a decrease in risk of HIV infection. In other words, women uninfected with syphilis had the highest risk of HIV infection of about 48%. Women infected with syphilis had the lowest HIV infection rate of about 24%. The latter scenario indicates a decrease of 24% in risk of HIV infection, when an individual changes her syphilis status from negative to positive. This is an interesting observation. However, in order to understand this result one requires to appreciate that Gen X are generally the older women born between the years 1961 to 1981. Therefore studying the entire women population means including both Generation X and Y, where the latter represents the younger generation born between the years 1982 to 2002. It could be concluded that amongst Gen X black women, syphilis infection causes individuals to be wary of HIV infection, thereby engaging in safe sexual practices, hence the observed decrease in risk of HIV infection. Furthermore, it has to be noted that amongst the entire black women population the increase in risk of HIV infection following syphilis infection was marginal at only 3%.

It was observed in Fig. 8a, that amongst the entire black women population, an improvement in educational level from no formal education to tertiary education resulted in a decrease in risk of HIV infection. However, amongst the older Gen X women alone, an improvement in education resulted in an increase in risk of HIV infection. That means studying both older and younger women together, an improvement in education resulted in a decrease in risk of HIV infection. Amongst, older Gen X women alone, an improvement in level of education resulted in a corresponding increase in risk of HIV infection. This could mean that, education amongst the older Gen X women does not assist in reducing the spread of HIV. Perhaps the latter behavior could be related to the availability of antiretroviral medicines that help individuals to live with HIV infection, resulting in an increase in risk taking behavior.

4. Conclusion

The D-optimal design was successfully used to study the effects of demographic characteristics on the risk of HIV infection amongst pregnant women attending antenatal clinics in South Africa. The study compared HIV risk profiles between Gen X black women and the entire population of black women attending antenatal clinics in South Africa. The observed differences were remarkable. However, it is important to note that Gen X women are a sub-population of the entire black women population. Therefore, studying the entire population of black women means including the effects of Gen X women. Perhaps, to better understand the observed differences in HIV risk profiles, the follow-up study should look at two distinct sub-populations of black women that make up the entire population namely Gen X and Gen Y.

5. References

[1] National Department of Health, "Protocol for implementing national antenatal HIV and syphilis prevalence survey, South Africa". 2010.

[2] S. Wold, "PLS for multivariate linear modelling QSAR chemometric methods in modular design methods and principles in medicinal chemistry".

[3] Umetrics, Multivariate analysis 3-day course, Winchester, MA.

[4] S. Wold, "Cross-validatory estimation of the number of components in factor and principal components model", Technometrics, 20, pp397-405,1978.

[5] L.Eriksson, Design of Experiments: Principles and Applications, 3rd Ed, Umetrics AB, Sweden.

Special session on

Regularization methods for genomic data analysis

Organisers

Dr. Franck Picard (CNRS, Univ. Lyon)
Prof. Vivian Viallon (Univ. Lyon)

Fast tree inference with weighted fusion penalties

Julien Chiquet[(1)]

(1) CNRS UMR 8071 and University d' Évry
France

Keywords:

Abstract. Given a data set with many features observed in a large number of conditions, it is desirable to fuse and aggregate conditions which are similar to ease the interpretation and extract the main characteristics of the data. This paper presents a multidimensional fusion penalty framework to address this question when the number of conditions is large. If the fusion penalty is encoded by an l_q-norm, we prove for uniform weights that the path of solutions is a tree which is suitable for interpretability. For the l_1 and l_∞-norms, the path is piecewise linear and we derive a homotopy algorithm to recover exactly the whole tree structure. For weighted l_1-fusion penalties, we demonstrate that distance-decreasing weights lead to balanced tree structures. For a subclass of these weights that we call "exponentially adaptive", we derive an O(nlog(n))homotopy algorithm and we prove an asymptotic oracle property. This guarantees that we recover the underlying structure of the data efficiently both from a statistical and a computational point of view. We provide a fast implementation of the homotopy algorithm for the single feature case, as well as an efficient embedded cross-validation procedure that takes advantage of the tree structure of the path of solutions. Our proposal outperforms its competing procedures on simulations both in terms of timings and prediction accuracy. As an example we consider phenotypic data: given one or several traits, we reconstruct a balanced tree structure and assess its agreement with the known taxonomy.

Special session on

Large-Scale and HPC data analysis in bioinformatics: intelligent methods for computational, systems and synthetic biology

Organisers

Dr. Andrea Bracciali (Dept. of Computing Science and Mathematics, University of Stirling, UK)

Dr. Ivan Merelli (Institute for Biomedical Technologies – Italian National Research Council, Italy)

Dr. Mario Guarracino (High Performance Computing and Networking Institute – Italian National Research Council, Italy)

This special session has been sponsored by the Italian Flagship initiative InterOmics (PB05).

nuChaRt: embedding High Performance Computing in R for augmented DNA Exploration.

Fabio Tordini[(1)], Ivan Merelli[(2)], Pietro Liò[(3)], Marco Aldinucci[(1)], Luciano Milanesi[(2)]

(1) Computer Science Dep., University of Turin
Corso Svizzera 185, 10149 Torino, Italy.
tordini@di.unito.it, aldinuc@di.unito.it

(2) Institute for Biomedical Technologies - Italian National Research Council
via F.lli Cervi 93, 20090 Segrate (Mi), Italy
ivan.merelli@itb.cnr.it, luciano.milanesi@itb.cnr.it

(3) Computer Laboratory, University of Cambridge
Trinity Lane, Cambridge CB2 1TN, UK.
pietro.lio@cl.cam.ac.uk

Keywords: Next-Generation Sequencing, Neighbourhood graph, High-Performance Computing, Multi-omics features, Systems Biology

Abstract Recent advances in molecular biology and bioinformatics techniques brought to an explosion of the information about the spatial organisation of the DNA in the nucleus. High-throughput chromosome conformation capture techniques provide a genome-wide capture of chromatin contacts at unprecedented scales, which permit to identify physical interactions between genetic elements located throughout the human genome. These important studies are hampered by the lack of biologists-friendly software. In this work we present *nuChaRt*, an R package that wraps NuChart-II, an efficient and highly optimized C++ tool for the exploration of Hi-C data. By rising the level of abstraction, *nuChaRt* proposes a high-performance pipeline that allows users to orchestrate analysis and visualisation of multi-omics data, making optimal use of the computing capabilities offered by modern multi-core architectures, combined with the versatile and well known R environment for statistical analysis and data visualisation.

1 Scientific Background

A huge amount of information is daily produced in the molecular biology laboratories all around the world, but the representation and interpretation of this data in an effective way is a complex and challenging task. Specifically, sequencing results from expression profiles, methylation patterns and chromatin domains are difficult to describe in a systemic view. Also, an increasing number of experiments highlight the importance of co-localization and co-expression of genes: there is an undeniable need for a software that permits the integration and the interpretation of multi-omics features on a nuclear map, capable of representing the effective disposition of genes in the three-dimensional (3D) space.

Over the last decade, a series of molecular genomic approaches have been developed to study the spatial organisation of chromosomes at increasing resolution and throughput. These methods are all based on Chromosome Conformation Capture (3C) and allow the determination of the interaction frequency of any pair of loci in the genome [2], which is related to their physical proximity (probably in the range of 10-100nm). The 3D conformation of chromosomes is involved in compartmentalizing the nucleus while bringing widely separated functional elements into close spatial proximity. Furthermore, understanding chromosome spatial organisation is crucial for many cellular processes related to gene expression regulation, including DNA accessibility, epigenetics patterns and chromosome translocations.

Among 3C-based solutions, Hi-C is a method that exploits Next-Generation Sequencing techniques to investigate genomic loci that physically interact in the nucleus [7]. Hi-C gives information about coupled DNA fragments that are cross-linked together (during the formaldehyde fixation step of the experimental protocol) due to spatial proximity, providing data about the chromosomal arrangement in the 3D space of the nucleus. The output of a Hi-C process is a list of pairs of locations along all chromosomes, which can be represented as a square matrix where each element (i, j) of the matrix indicates the sum of read pairs matching in positions i and j. This matrix-based representation, called *contact map*, is reliable while looking at the interactions between two chromosomes, but becomes unsuitable to describe long-range chromatin interactions or to model a contact based metric of gene distances. A graph-based representation of genomic data offers a more comprehensive characterization of the chromatin conformation, which can be very useful to create a representation on which other omics data can be mapped, in order to characterize different spatially-associated domains.

In previous works we proposed *NuChart-II* as a highly optimised, C++ version of an early prototype software designed to integrate information about genes positions and paired-ends reads resulting from Hi-C experiments, in order to describe the chromosome spatial organisation using a gene-centric, graph-based approach [8, 9]. In that work we have investigated the possibility of introducing network concepts to represent the behaviour of genomic actors: a network (or graph) has a high level of expressiveness, since nodes represent the actors of a process while edges identify relationships among the actors. Structural properties of a network can reveal significant information on how the actors of the represented process interact, while parallel algorithms can be employed to operate over a network. The graph-based approach has been proved to be a valuable way for the interpretation of genomic information by mean of complex, dynamical structures that organize items in an integrated way.

The novel C++ software has been designed using high-level parallel programming patterns that facilitate the implementation of the algorithms employed over the graph: this choice permits to boost performances while conducting genome-wide analysis of the DNA. In order to stick with the original idea of providing a complete suite of tools for the analysis of Hi-C data, here we propose *nuChaRt*, an R/Rcpp package that combines the high efficiency offered by the parallelized C++ implementation of NuChart-II, with the versatile and well known R environment: by embedding NuChart-II into an R package we obtain a high performance pipeline that facilitate the orchestration of genomic data analysis.

2 Materials and Methods

NuChart-II was rebuilt on top of *FastFlow*, a C++ header-only library that provides high-level parallel programming patterns and exposes the `ParallelFor` skeleton to easily deal with loop parallelism [1]. The coupled usage of C++ with advanced techniques of parallel computing (such as lock-free algorithms and memory-affinity) strengthens genomic research, because it makes possible to process much faster, much more data: informative results can be achieved to an unprecedented degree [4].

However, parallel programming in C++ is not widely used in bioinformatics, because it requires highly specialised skills and does not fully support the rapid development of new interactive pipelines. Conversely, the modularity of R and the huge amount of already existing statistical packages facilitates the integration of exploratory data analysis and permits to easily move through the steps of model development, from data analysis to implementation and visualisation.

In order to improve the usability of the software while preserving the high performances achievable with NuChart-II, we opted to combine it with the R environment, developing a package that can fulfil the needs of a fast and usable tool for Hi-C data

interpretation. We used *Rcpp* to bridge C++ and R: Rcpp comes as an R add-on package that provides a consistent API for accessing, extending or modifying R objects at the C++ level [5]. It can be used to accelerate computing by replacing an R function with its C++ equivalent and facilitates data interchange from R to C++ and vice-versa, through the templated functions `Rcpp::as<>()` and `Rcpp::wrap()`, respectively.

In our context, we have dealt with four C++ objects that abstract the leading actors of the software: `SamData`, `Gene`, `Fragment` and `Edge`. These objects contain much of the information that is needed to build a topographical map of the DNA from Hi-C data. Using Rcpp's functions we can convert these classes into a *S Expression Pointer* (called `SEXP`), that can be handled on the R side to construct `Lists` or `DataFrames`, which are essential object types in R and are used by almost all modelling functions. For instance, the `SamData` class has private fields that describe the reads, thus containing the chromosomes' names and a starting coordinate for each chromosome of the paired-ends read, plus the genomic sequence. In order to exchange a `SamData` object between C++ and R we have specialised the templated functions above: a `std::vector<SamData>` is thus treated by R as a list of `Lists`, while a list of `Lists` in R (or a `DataFrame`) is managed in C++ by casting the `SEXP` object to a `Rcpp::List` (or a `Rcpp::DataFrame`) object, and by subsequently filling each field of the `SamData` class with the value contained in the respective field of the `List`.

Algorithm 1 – *Example of `as` and `wrap` usage*

```
template<> SEXP wrap(const SamData &s) {
        List ret = Rcpp::List :: create ( Rcpp::Named("Id")      = s.getId(),
                                          Rcpp::Named("Chr1")    = s.getChr1(),
                                          Rcpp::Named("Start1")  = s. getStart1 (),
                                          Rcpp::Named("Chr2")    = s.getChr2(),
                                          Rcpp::Named("Start2")  = s. getStart2 (),
                                          Rcpp::Named("Seq")     = s.getSeq()
                                        );
        return Rcpp::wrap(ret);
}

template<> SamData as( SEXP s ) {
        List samL = Rcpp::as<Rcpp::List>(s);
        SamData sam;

        sam.setId    ( Rcpp::as<long>( samL["Id"])    );
        sam.setChr1  ( Rcpp::as<string>(samL["Chr1"]) );
        sam.setStart1( Rcpp::as<long>( samL["Start1"]) );
        sam.setChr2  ( Rcpp::as<string>(samL["Chr2"]) );
        sam.setStart2( Rcpp::as<long>( samL["Start2"]) );
        sam.setSeq   ( Rcpp::as<string>(samL["Seq"])  );

        return sam;
}
```

Embedding NuChart-II in R creates an application that can be used either to conduct step-by-step analysis of genomic data, or as a high-performance workflow that takes heterogeneous datasets in input, processes the data and produces a graph-based representation of the chromosomal information provided, supported by a rich set of default descriptive statistics derived from the topology of the graph. Thanks to the mechanism showed in Algorithm 1, C++ objects holding the output of the computation are available within the R environment, ready to be used as source for advanced statistical analysis.

2.1 *nuChaRt*

The work on the *nuChaRt* package started from a thorough study of the software, from which we identified five main phases that compose the application:

1) data retrieval and parsing;
2) neighbourhood graph construction;
3) weighing of the edges as a result of the normalisation step;
4) statistical analysis;
5) output and visualisation.

Phase *1)* is responsible for data collection and early data processing. Datasets are provided as static csv-like files but can also be downloaded from on-line repositories. The information contained therein is parsed and processed, in order to build the data structures needed to perform the computations: unneeded fields are dropped and elements are ordered in a consistent way, while a unique identifier for each element of a collection is generated, when needed. Problems may arise if these operations are performed on the R side, as it may lead to memory overflows with big size files ($> 2GB$, as it is the case with SAM files) due to the way R objects are constructed and stored in memory. For this reason parsing of data whose size exceeds $2GB$ is kept on the C++ side. No matter where these operations are executed, objects can be moved from C++ side to R side and vice-versa, as explained above.

Phases *2)* and *3)* constitute by far the most onerous parts of the application, in terms of execution time. Both of them are suitable for being revisited in the context of loop parallelism, since their kernels can be run concurrently on multiple processors with no data dependencies involved.

Neighbourhood Graph Construction A graph is a collection of vertices V connected by edges E that model a relationship among vertices. In our context, vertices represent genes. Two genes $g_1, g_2 \in V$ are connected if there exists a paired-ends Hi-C read encompassing both of them. We define this paired-ends Hi-C read as a *connection*, meaning a spatial relationship between two genes. If a connection between two genes g_1, g_2 exists, then exists an edge $e = (g_1, g_2) \in E$. When constructing the graph, any arbitrary subset of Hi-C reads can be processed independently from each other. This means it can be parallelised in a seamless way by wrapping the business logic within a `ParallelFor` skeleton, whose semantic amounts to execute in parallel the instructions inside the loop.

Edges Weighing A normalisation process is needed to remove systematic biases arising from sequencing and mapping. The weight of each edge of the graph is the result of the normalisation phase and provides a likelihood of the actual physical proximity for the adjacent genes. Inspired by the work of Hu et al. [6], for each edge a contact map Y is constructed directly modelling the read count data at a resolution level of 1 Mb. Each point $Y_{i,j}$ of the contact map denotes the intensity of the interaction between positions i and j. Using local genomic features that describe the chromosome (fragment length, GC-content and mappability) we can set up a linear model with Poisson regression, by which we estimate the maximum likelihood of the model parameters. This likelihood is then expressed as the weight of the edge that qualifies the reliability of the *gene–gene* contact. Again, any arbitrary subset of the edges can be processed independently from each other, making also this part of the application suitable for a parallelisation with a `ParallelFor` skeleton.

Phase *4)* encompasses essential features that the package ought to provide, in order to fulfil the requirements of a useful tool for genomic data interpretation. With a graph-based representation we can apply network analysis over the resulting graph: topological measures capture graph's structure for nodes and edges and highlight the "importance" of the actors. For instance, centrality metrics describe the interactions

that (may) occur among local entities. Ranking of nodes by topological features (such as degree distribution) can help to prioritize targets of further studies or lead to a more local, in-deep analysis of specific chromosome locations. Here studies of functional similarity can suggest new testable hypotheses [10].

Finally, visualisation is crucial for a tool that aims at facilitating a better interpretation of genomic data. NuChart-II supplies both tabular output and graphical visualization. Concerning the latter, *iGraph* and *GraphViz* are used as plotting engines, but while these tools perform nicely with small-to-medium sized graphs, they cannot provide useful representation of huge graphs. On the R side there are several graphic libraries – *MuxViz* or *networkD3* just to cite a few – that facilitate the interactive visualization and exploration of complex networks. With *nuChaRt* we can seamlessly exploit these libraries to create navigable and interactive maps of the chromosome.

3 Results

The novel package benefits of the combined use of a High-Performance Computing foundation, provided by the C++ engine, and the flexibility of R. *nuChaRt* maintains excellent performance and scalability as showed in NuChart-II [9, 4], being able to perform genome-wide analysis of Hi-C data with reduced memory footprint and exploiting loop parallelism in the graph construction phase and in the normalisation phase. Moreover, within the R environment it is possible to execute a step-by-step analysis in a seamless, unconstrained, way: the graph creation would be the first step, after which is already possible to create a graph and apply initial statistical analysis to study the topological features of the network. Results are also globally available in form of `DataTables`, and can be easily queried and inspected. The normalisation follows, where edges (connections) are evaluated and weighted: the existence of each identified Hi-C connection is assessed and graded. Again, the graph can be plotted and results can be visualised and browsed. Eventually, one can draw from the huge R's libraries basket the one that suits her need, and conduct advanced analysis over the resulting data. For instance, we tested the *ergm* package that permit to understand the processes of network structure emergence and tie formation.

Case Study The study of the interactions of the actor genes with the environment is of critical importance for understanding the entire system. By using the modelling functions of the package we can statistically characterize the distribution of the edges in relation to the characteristics of the nodes that represent mapped multi-omics features. In order to show the possibilities of *nuChaRt* in terms of statistical inference on the graph we performed the analysis of the clusters of genes Kruppel-Associated Box (KRAB) and Human Leukocyte Antigen (HLA) in the context of four Dixon experiments [3], to verify the correlation of the edges distribution in relation to some genomic features (hypersensitive sites, CTCF binding sites, isochores, RSSs). The correlation between the presence of CTCF binding sites and edges was predictable since Linking Gene Regulatory Elements are demanded to keep different regions of the genome close to each other, but it is very interesting to quantify this association. On the other hand, regions with isochores seem less involved in long-range interactions, which can be quite surprising considering that these portions of the genome are considered gene-rich. The correlation between cryptic RSS sites and edges is more pronounced in the HLA cluster in comparison to the KRAB cluster, probably due to a more consistent presence of this kind of sequences in genes related to the immune system. Finally, the correlation between hypersensitive sites (super sensitivity to cleavage by DNase) and edges, although positive, is poor, probably because the accessibility of these regions are impaired by a large number of long-range interactions.

4 Conclusion

Coupling the broad R modelling capabilities and the high performances achievable with the parallel C++ implementation of NuChart-II is clearly a winning combination and a solid response to the demand for software tools that help scientists in drawing more concrete biological knowledge. The graph-based approach fosters a tight coupling of topological observations to biological knowledge, which is likely to bring remarkable biological insights to the whole research community. From a computational point of view, the ever-increasing amount of information generated by novel bioinformatics techniques require proper solutions that permit the full exploitation of the computing power offered by modern computing systems, together with advanced tools for an efficient analysis and interpretation of genomic data. These tasks require high skills, but we believe that *nuChaRt* can be a valuable mean to support researchers in pursuing these objectives.

Acknowledgments

This work has been partially supported by the EC-FP7 STREP project "REPARA" (no. 609666), the Italian Ministry of Education and Research Flagship (PB05) "InterOmics", and the EC-FP7 innovation project "MIMOMICS".

References

[1] Danelutto M, Torquati M, "Loop parallelism: a new skeleton perspective on data parallel patterns", in *Proc. of Intl. Euromicro PDP 2014: Parallel Distributed and network-based Processing*, 2014.

[2] Dekker J, Rippe K, Dekker M, Kleckner N, "Capturing chromosome conformation", *Science*, vol. 295, pp. 1306-1311, 2002.

[3] Dixon J R, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al., "Topological domains in mammalian genomes identified by analysis of chromatin interactions", *Nature*, vol. 485, pp. 376-380, 2012.

[4] Drocco M, Misale C, Pezzi GP, Tordini F, Aldinucci M, "Memory-Optimised Parallel Processing of Hi-C Data", in *Proc. of Intl. Euromicro PDP 2015: Parallel Distributed and network-based Processing*, pp. 1-8, 2015.

[5] Eddelbuettel D and Francois R, "Rcpp: Seamless R and C++ Integration", *Journal of Statistical Software*, vol. 40, pp. 1-18, 2011.

[6] Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS, "HiCNorm: removing biases in Hi-C data via Poisson regression", *Bioinformatics*, vol. 28, pp. 3131-3133, 2012.

[7] Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al., "Comprehensive mapping of long-range interactions reveals folding principles of the human genome", *Science*, vol. 326, pp. 289-293, 2009.

[8] Merelli I, Tordini F, Drocco M, Aldinucci M, Lió P, Milanesi L, "Integrating Multi-omic features exploiting Chromosome Conformation Capture data", in *Frontiers in Genetics*, vol. 6, 2015.

[9] Tordini F, Drocco M, Merelli I, Milanesi L, Lió, Aldinucci M, "NuChart-II: a graph-based approach for the analysis and interpretation of Hi-C data", in *Proc. of the 11th Intl. meeting on Computational Intelligence methods for Bioinformatics and Biostatistics* (CIBB 2014), Cambridge, UK, 2015.

[10] Winterbach W, Van Mieghem P, Reinders M, Wang H, de Ridder D, "Topology of molecular interaction networks", *BMC Systems Biology*, vol. 7, 2013.

A NOVEL TECHNIQUE FOR REDUCTION OF FALSE POSITIVES IN PREDICTED GENE REGULATORY NETWORKS

Abhinandan Khan [(1)], Rajat Kumar Pal [(1)], Goutam Saha [(2)]

(1) University of Calcutta, Department of Computer Science and Engineering, Acharya Prafulla Chandra Roy Siksha Prangan, JD – 2, Sector – III, Saltlake City, Kolkata – 700 098, West Bengal, India. {khan.abhinandan, pal.rajatk}@gmail.com

(2) North–Eastern Hill University, Department of Information Technology Umshing, Mawkynroh, Shillong – 793 022, Meghalaya, India. dr.goutamsaha@gmail.com

Keywords: bat algorithm, gene regulatory network, particle swarm optimisation, recurrent neural network, time series microarray data.

Abstract. In this paper, we have proposed a novel method for the reduction of the number of false positives in gene regulatory networks, constructed from time series microarray genetic expression datasets. We have implemented a hybrid statistical / swarm intelligence technique for the purpose of reverse engineering genetic networks from temporal expression data. The theory of combination has been used to reduce the search space of network topologies effectively. Recurrent neural networks have been employed to obtain the underlying dynamics of the expression data accurately. Two swarm intelligence techniques, namely, Particle Swarm Optimisation and a Bat Algorithm inspired variant of the same, have been used to train the corresponding model parameters. Subsequently, we have identified and used their common portions to construct a final network where the incorrect predictions have been filtered out. We have done preliminary investigations on experimental (*in vivo*) data sets of the real world network for SOS DNA repair in *Escherichia coli*. Experimental results are quite encouraging, and they suggest that the proposed methodology is capable of reducing the number of false positives, thus, increasing the overall accuracy and the biological plausibility of the predicted genetic network.

1. Scientific Background

To fully comprehend the critical cellular activities of living beings [1], it is imperative that we learn the exact nature of the genetic relationships using the knowledge of genetic expression patterns. Investigations on gene regulatory networks (GRNs), thus, have enticed the research fraternity considerably. Hence, the development of suitable methodologies to completely understand the causal relationships between genes has ensued. A GRN represents the complex, genetic inter-regulatory relationships.

The simultaneous measurement of the genetic expression levels of several thousand genes has been made possible by the innovations of DNA microarray technology. However, the microarray data contains unwanted, experimental noise. Additionally, the number of genes investigated is two to three orders of magnitude higher compared to the number of time points.

The problem is known as the *curse of dimensionality* and it severely undermines the potential of any applied methodology for the construction of GRNs from temporal microarray datasets. Several research endeavours have been made to solve this problem but researchers have achieved only partial success in this regard.

Various methodologies for reverse engineering of GRNs from time series expression data such as Boolean Networks [2], Recurrent Neural Networks (RNN) [3], S-systems [4], etc. have been investigated. A review of the various methods used for reverse engineering of GRNs from time series microarray data is given in [5]. The training of the corresponding model parameters is an optimisation problem. Thus, metaheuristic techniques like particle swarm optimisation (PSO) [6] are quite popular among researchers worldwide. In this investigation, apart from PSO, a bat algorithm (BA) [7] inspired variant of PSO has been introduced and implemented. Results were evaluated based on their combination.

Although metaheuristics are extensively used for model parameter training purposes, the number of parameters increases in a quadratic manner with respect to the number of genes in a GRN. Thus, for $N \sim 10^2$ or 10^3 numbers of genes, optimisation becomes computationally implausible. Researchers have proposed to solve this particular problem by decomposing the global optimisation problem (model parameter optimisation of all genes in a GRN) into several local optimisation problems (model parameter optimisation of a single gene) [8]. Extensive investigations on GRNs reveal that a GRN contains only a handful of regulators [9], i.e. GRNs are sparse in nature. This information points towards the possibility of some form of topological constraints being applied on the predicted GRNs. It thus, becomes feasible to decouple the architectural and the dynamical features of this reverse engineering problem. This can be realised by decoupling the discrete network architecture search space from the continuous model parameter search space [8]. The continuous search supervises the discrete search.

2. Materials and Methods

2.1. *Materials*

We have implemented the proposed approach on the *in vivo* time series microarray datasets of the SOS DNA repair network of *E. coli* [10]. This DNA repair mechanism involves only 8 genes as were studied by Ronen et al. [10]. The original network comprising of these 8 genes is shown in Fig. 1, and involves a total of 9 interactions. These datasets are mostly used as a benchmark for the comparison of the results obtained from different computational methodologies for reconstruction of GRNs. Ronen et al. [10] performed four such experiments, producing four microarray datasets. Each dataset comprises of the expression values of the eight genes shown for 50 time points at an interval of 6 minutes.

2.2. *Methods*

In this work, we propose to implement a statistical framework hybridised with Particle Swarm Optimisation (PSO) [6], and another one, hybridised with a Bat Algorithm [7] inspired Particle Swarm Optimisation (BA-PSO) algorithm for the construction of GRNs. The theory of

combination has been introduced to reduce the search space of network topologies. Considering each sub-problem, we prefix the maximum number of allowable regulators for a particular gene to $m = N/2$, where N is the number of genes in a GRN. Thus, we achieve a reduction of the search space from a maximum dimension of $2^N - 1$ to ${}^{N}C_{m}$. Furthermore, it increases the likelihood of inferring biologically plausible networks with a lesser chance of false predictions. The RNN technique is employed for the purpose of modelling the underlying information of the dynamics of the temporal genetic expression data [3].

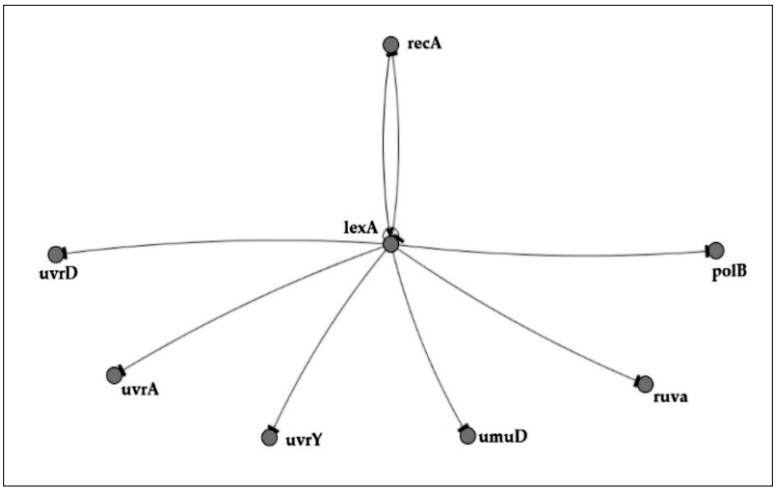


Fig. 1. The original structure of the SOS DNA repair transcriptional network of *E. coli*. Arrowheads represent activation; T-heads represent repression. (courtesy: (http://wws.weizmann.ac.il/mcb/UriAlon/sites/mcb.UriAlon)

The novelty of this work lies in the final network construction strategy based on the results of the two formalisms above. We compare, rather, superimpose the two GRNs constructed. Subsequently, we use only the edges, common to both the structures, to assemble the final inferred GRN. This filters out the false positives while keeping the true positives intact. This can be explained based on the fact that GRNs are sparse. Thus, false positives can be scattered throughout the entire search space and the possibility of identifying the same false positive, by different metaheuristic techniques, is low. As a result, the incorrect predictions vary in their positions in the inferred networks for different methodologies and get filtered out. On the other hand, the true positives are unwavering in their positions and thus, survive the filtering process.

The traditional RNN formalism as described in [3] is implemented in this work. Traditional PSO is used for the first implementation of the proposed methodology, described by equations (1) and (2).

$$v'_i = w \otimes v_i + r_1 c_1 \otimes \left(p_i^b - p_i\right) + r_2 c_2 \otimes (g^b - p_i) \quad (1)$$

$$p'_i = p_i + v'_i \quad (2)$$

Here, v' and v denote the particle velocities for the next and the current generations, respectively;

x' and x are the particle positions in the next and the current generations, respectively;

p^b is the best solution achieved by a particle;

g^b is the best solution achieved by the swarm;

w is the inertia weight term that controls the efficient balance between exploration and exploitation undertaken by a particle;

r_1, r_2 are random numbers in the range [0, 1]; and

c_1, c_2 are taken as 2.

In the BA-PSO, the update of w is inspired by the frequency update of BA. Here, we uniformly draw a random value of w from $[w_{min}, w_{max}]$. We assign $w_{min} = 0$ and $w_{max} = 1$ in this work. This, to a certain extent, counterbalances the problem of getting trapped at local minima thus, getting an upper hand over one the very few problems plaguing PSO.

Another modification introduced into the proposed BA-PSO algorithm, is the initialisation of the particle velocity to zero. Inspired by virtual bats, this might help in preventing particles from acquiring an initial unguided velocity that may sidetrack it from a potential optimal solution in the search space.

Last but not the least, for this investigation, GRNs need to be represented computationally, and that is most easily achieved via directed graphs. A directed graph, $G = (V, E)$ can represent a GRN, where V is the set of all genes (nodes or vertices) and E is the set of all interactions between the elements of V (edges). E contains an edge, $e_{i,j}$ if and only if a causal relationship is present between the vertices (genes), i and j. This structure can be represented with the help of an adjacency matrix, $G = [g_{i,j}]_{N\times N}$ where N = the number of nodes (genes) in the graph. The element $g_{i,j}$ has a value 0 or 1 depending on the absence or presence of any regulatory interaction from gene j to gene i, respectively.

3. Experimental Results

In this work, we employ a collaborative learning scheme due to the stochastic character of the methodologies involved (as a consequence the GRNs vary in their structures for each independent experiment). Subsequently, we assign a plausibility score, $ps_{i,j}$, for each edge $e_{i,j}$, for its inclusion in the final predicted network, as:

$$ps_{i,j} = \frac{1}{L}\sum_{1}^{L} g_{i,j} \tag{3}$$

In the above equation, $g_{i,j} \in$ G and $ps_{i,j} \in [0, 1]$, L = numbers of independent experiments constructed (inferred GRNs) corresponding to each methodology. After the evaluation of $ps_{i,j}$ for all i and j, we store the final inferred GRN as a matrix, $G_F = [g^f_{i,j}]_{N\times N}$. The value of $g^f_{i,j}$ can be either 0 or 1 and we adopt the following technique to assign a value to $g^f_{i,j}$:

$$g^f_{i,j} = \begin{cases} 1, & if\ ps_{i,j} \geq \alpha \\ 0, & otherwise \end{cases} \tag{4}$$

The parameter, α is a threshold of $ps_{i,j}$, i.e. the plausibility score. It regulates the inclusion of a particular edge in the final GRN. In order to evaluate the accuracy of the implemented methodology, we compare this final GRN, G_F with G_O, the original GRN. The experimental results are validated in this manner. The following metrics help in the quantitative comparison of the proposed methodology with those in contemporary literature. An inferred edge is categorised into four types:

- True Positive (TP): if $g^o_{i,j} = 1$ and $g^f_{i,j} = 1$, True Negative (TN): if $g^o_{i,j} = 0$ and $g^f_{i,j} = 0$.

- False Positive (FP): if $g^o_{i,j} = 0$ and $g^f_{i,j} = 1$, False Negative (FN): if $g^o_{i,j} = 1$ and $g^f_{i,j} = 0$.

$$True\ Positive\ Rate\ (TPR)/Sensitivity/Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Positive\ Predictive\ Value\ (PPV)/Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

$$F - Score\ (F) = \frac{2TP}{2TP + FP + FN} \tag{8}$$

Here, in the SOS DNA repair mechanism of *E. coli, L* = 20 independent experiments are conducted with a swarm population of n = ${}^8C_4$, and a maximum number of 5000 iterations. The statistical properties of the inferred GRNs for a plausibility score threshold, α = 0.85 are shown in Table 1. The obtained prediction error, (MSE) is ~10^{-2}.

Table 1: Experimental Results.

Dataset	Technique	TP	FP	Sensitivity	Precision	Accuracy	F-Score	Graph Edges
1	eDSF [8]	3	10	0.33	0.23	0.75	0.27	13
	PSO	5	9	0.56	0.36	0.80	0.43	14
	BAPSO	7	9	0.78	0.44	0.83	0.56	16
	Proposed	**7**	**9**	**0.78**	**0.44**	**0.83**	**0.56**	**16**
2	eDSF [8]	8	5	0.89	0.62	0.91	0.73	13
	PSO	4	10	0.44	0.29	0.77	0.35	14
	BAPSO	7	15	0.78	0.32	0.73	0.45	22
	Proposed	**7**	**12**	**0.78**	**0.37**	**0.78**	**0.50**	**19**
3	eDSF [8]	3	10	0.33	0.23	0.75	0.27	13
	PSO	4	9	0.44	0.31	0.78	0.36	13
	BAPSO	4	9	0.44	0.31	0.78	0.36	13
	Proposed	**7**	**9**	**0.78**	**0.44**	**0.83**	**0.56**	**16**
4	eDSF [8]	0	9	0.00	0.00	0.72	0.00	9
	PSO	3	8	0.33	0.27	0.78	0.30	11
	BAPSO	4	12	0.44	0.25	0.73	0.32	16
	Proposed	**4**	**0**	**0.44**	**1.00**	**0.92**	**0.62**	**4**

Table 1 shows the comparison of the inferred network structures with those inferred in [8]. The proposed framework is consistent with respect to the number of correct (true positive) and incorrect (false positive) predictions for each dataset unlike the unevenness in the results of [8].

Moreover, the novel scheme of constructing the final inferred topology has succeeded in the *reduction of false positives*, as seen in Table 1. In general, BA-PSO identifies a greater number of edges compared to PSO. As a result, there are more true positives identified, but at the same time, more false positives creep into the final inferred topology as well. Constructing a GRN with only those edges that are common to both the methodologies, helps in filtering out the noisy information i.e. false positives from the GRNs at the same time preserving the increased number of true predictions. The results show the extent of such filtering: in the first

three datasets, the false positives reduce but for the final dataset, the false positives vanish altogether.

We could only present the truncated statistical results of only one experiment due to the shortage of space. However, we have performed more experiments with different datasets, and the results are similarly encouraging.

4. Conclusion

In this work, we have explored the construction of GRNs from temporal expression datasets where a decoupled methodology based on the ideas of combining swarm intelligence algorithms with RNN has been implemented. Results show that this methodology is capable of achieving a reduction in the number of false positives without sacrificing true positives. In the present context, the identification and reduction of the number of false predictions require conscious efforts that may have been slightly neglected in the endeavour of predicting more and more true positives.

References

[1] Geoffrey McLachlan, Kim-Anh Do, and Christophe Ambroise. “Analysing microarray gene expression data.” Vol. 422. John Wiley & Sons, 2005.

[2] Stuart A. Kauffman. “Metabolic stability and epigenesis in randomly constructed genetic nets.” Journal of Theoretical Biology 22, no. 3 (1969): 437-467.

[3] Jiri Vohradsky. “Neural model of the genetic network.” Journal of Biological Chemistry 276, no. 39 (2001): 36168-36173.

[4] Eberhard O Voit. Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists. Cambridge University Press, 2000.

[5] Hendrik Hache, Hans Lehrach, and Ralf Herwig. "Reverse engineering of gene regulatory networks: a comparative study." *EURASIP Journal on Bioinformatics and Systems Biology* 2009 (2009): 8.

[6] Russ C. Eberhart and James Kennedy. “A new optimizer using particle swarm theory.” In Proceedings of the Sixth International Symposium on Micro Machine and Human Science, vol. 1, pp. 39-43. 1995.

[7] Xin-She Yang. “A new metaheuristic bat-inspired algorithm.” In Nature Inspired Cooperative Strategies for Optimisation (NICSO 2010), pp. 65-74. Springer Berlin Heidelberg, 2010.

[8] Kyriakos Kentzoglanakis and Matthew Poole. “A swarm intelligence framework for reconstructing gene networks: Searching for biologically plausible architectures.” Computational Biology and Bioinformatics, IEEE/ACM Transactions on 9, no. 2 (2012): 358-371.

[9] Eugene P. van Someren, L. F. A. Wessels, Eric Backer, and M. J. T. Reinders. “Genetic network modelling.” Pharmacogenomics 3, no. 4 (2002): 507-525.

[10] Michal Ronen, Revital Rosenberg, Boris I. Shraiman, and Uri Alon. “Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics.” Proceedings of the National Academy of Sciences 99, no. 16 (2002): 10555-10560.

DETERMINISTIC SIMULATIONS OF LARGE-SCALE MODELS OF CELLULAR PROCESSES ACCELERATED ON GRAPHICS PROCESSING UNITS

Andrea Tangherloni[(1)], Paolo Cazzaniga[(2,4)], Marco S. Nobile [(1,4)], Daniela Besozzi[(3,4)], Giancarlo Mauri[(1,4)]

(1) Università degli Studi di Milano-Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Viale Sarca 336, 20126 Milano, Italy
a.tangherloni@campus.unimib.it; nobile/mauri@disco.unimib.it

(2) Università degli Studi di Bergamo
Dipartimento di Scienze Umane e Sociali
Piazzale S. Agostino 2, 24129 Bergamo, Italy
paolo.cazzaniga@unibg.it

(3) Università degli Studi di Milano
Dipartimento di Informatica
Via Comelico 39, 20135 Milano, Italy
besozzi@di.unimi.it

(4) SYSBIO Centre of Systems Biology
Piazza della Scienza 2, 20126 Milano, Italy

Keywords: GPU computing, deterministic simulation, Runge-Kutta, biochemical reaction network, large-scale model.

Abstract. The computational investigation of mathematical models of cellular processes represents an essential methodology in Systems Biology, complementary to conventional experimental biology, to understand the emerging behavior of biological systems. The simulation of the dynamics of these models, usually required for computational studies such as parameter estimation or sensitivity analysis, can become burdensome if the corresponding biochemical reaction network is characterized by hundreds or thousands of different molecular species and reactions. In this work, we introduce a novel GPU-powered fine-grain deterministic simulator of large-scale models of biochemical reaction networks, and test its computational performances on a set of randomly generated synthetic models of increasing size. We show that our parallel simulator, running on a GPU Nvidia GeForce GTX Titan Z, outperforms the sequential version, running on a CPU Intel i7-4790K 4.00GHz, achieving up to $7.8\times$ speed-up.

1 Scientific Background

Thanks to the availability of efficient high performance computing architectures, *in silico* investigation of biological systems can nowadays be strongly supported by effective computational strategies, which allow to achieve an accurate understanding of the behavior and the emerging properties of biochemical reaction networks (BRNs).

Given a stochastic or deterministic model of a BRN, typical computational methods used in Systems Biology—e.g., parameter estimation or identifiability, parameter sweep analysis, sensitivity analysis, reverse engineering [2]—usually require the execution of large numbers of simulations, possibly resulting in prohibitive running times on Central

Processing Units (CPUs). In addition, when the BRNs are characterized by a large number of different molecular species or reactions, these computational tasks can become excessively burdensome.

To overcome these limitations, Graphics Processing Units (GPUs) can be adopted as an alternative approach to classic parallel architectures (e.g., computer clusters) for the parallelization and the speed-up of the simulation of large-scale models of BRNs. Indeed, GPUs are cheaper than classic CPU cluster architectures, as they allow the execution of demanding computational analyses also on a standard desktop computer, and they also provide an effective mean to sustainable computing by drastically reducing the power consumption.

Anyway, the development of methods that fully exploit the GPU's peculiar architecture is a challenging task, since specific programming skills and a complete redesign of the algorithms are necessary. In order to make GPU-powered simulators available to the Systems Biology community, different GPU-based versions of the most efficient numerical integration methods have been introduced so far. Specifically, both *coarse-grain* and *fine-grain* parallel implementation have been developed. On the one hand, coarse-grain parallelization allows to simultaneously run a massive number of simulations of the same model, each one characterized by a different parameterization; on the other hand, fine-grain parallelization permits to distribute all the calculations required by a single simulation on the cores of the GPU.

An example of coarse-grain parallelization of deterministic simulations of BRNs was presented by Ackermann *et al.* [1]. In this work, a SBML model of a biochemical system is automatically converted into CUDA code, compiled and linked with the parallel simulator and the host code, producing a final executable file ready to simulate the dynamics. The performances of this method were assessed with a NVIDIA GeForce 9800 GX2, which shown a speed-up between $28\times$ to $63\times$ with respect to a CPU Xeon 2.66 GHz. Following a similar approach, Nobile *et al.* presented a parallel simulator named cupSODA [9], which allows to automatically generate the system of ODEs and the corresponding Jacobian matrix from a reaction-based mechanistic model of a biological system, described according to the mass-action kinetics. Differently from Ackermann's solution based on Euler's method, cupSODA relies on the LSODA numeric integration algorithm, achieving a higher accuracy in the simulation of the dynamics, and accelerating the computation in case of systems characterized by stiffness. cupSODA allows a $86\times$ speed-up with respect to COPASI, used as a reference CPU-based LSODA simulator. Another strategy to accelerate deterministic simulations by means of LSODA was presented by Zhou *et al.* [10]. This simulator, named cuda-sim, performs just in time compilation by converting a SBML model into CUDA code, and allows to achieve a $47\times$ speed-up with respect to a CPU implementation written in Python.

Differently from the solutions previously presented in literature, in this work we introduce a novel GPU-based simulator, named LASSIE (LArge-Scale SImulator), specifically designed to accelerate *fine-grain* deterministic simulations of *large-scale* mechanistic models of BRNs.

2 Materials and Methods

In this work we assume that a mechanistic model of a BRN is specified by giving the set of the N different molecular species $\{S_1, \ldots, S_N\}$ that are involved, either as reactants or products, in a set of M biochemical reactions $\{R_1, \ldots, R_M\}$. The amount (concentration) of species S_i is denoted by X_i, for $i = 1, \ldots, N$, while the kinetic constant associated with reaction R_j is denoted by k_j, for $j = 1, \ldots, M$.

The system of ordinary differential equations (ODEs)—which describes how the concentration of each species occurring in the BRN varies in time—can be easily derived by assuming the law of mass action. This basic biochemical kinetic law states

that, in a dilute solution, the rate of an elementary reaction (i.e., a reaction with a single mechanistic step) is proportional to the product of the concentration of its reactants raised to the power of the corresponding stoichiometric coefficient.

As an example, consider the BRN described by 3 reactions among 4 species:

$$S_1 + S_2 \underset{k_2}{\overset{k_1}{\rightleftarrows}} S_3 \overset{k_3}{\rightarrow} S_4, \tag{1}$$

and let us derive the ODE for species S_3, that is,

$$\frac{dX_3}{dt} = k_1 X_1 X_2 - k_2 X_3 - k_3 X_3. \tag{2}$$

Eq. 2 is given by the sum of the rates of production and degradation of S_3, determined by multiplying the concentrations (raised to the power of 1, in this case) of all species involved in Eq. 1 together with the corresponding kinetic constants. Note that, in this ODE, the right-hand side consists of 3 terms, corresponding to the number of reactions where S_3 is involved in. By repeating the same procedure for all species in the BRN, we can derive the whole system of ODEs and proceed with its numerical integration by using some ODE solver.

Briefly, LASSIE implements the following workflow:

1. given as input a mechanistic model of a BRN, it automatically derives the system of ODEs, as described above;
2. it determines an ordering of the various ODEs;
3. it solves the systems of ODEs by exploiting the Runge-Kutta (RK4) method.

The novel feature of this tool consists in the ODE-ordering step. The rationale behind this step is twofold. On the one hand, if the BRN consists in a number of species and a number of reactions that are in the order of hundreds of thousands, than also the corresponding system of ODEs will be very large (measured both in the number of ODEs and in the number of terms per ODE). In this case, standard numerical integration methods on CPU might require a long running time, therefore demanding alternative parallel solutions, as general-purpose GPU computing. On the other hand, to implement fast numerical integration methods on GPUs, one should also take into account their peculiar architecture.

To better explain this issue, it is worth considering that GPUs are based on a SIMD (Same Instruction Multiple Data) programming model, combined with multi-threading: all threads execute the same code, accessing different memory areas. In LASSIE, this feature is used to launch multiple threads, so that each thread solves the ODE related to a specific molecular species, by using the RK4 method. All threads do not run simultaneously, but are organized in blocks, which are distributed on the available streaming multiprocessors (SMs). SMs organize the threads within a block into sub-groups of 32 threads, named warps, which are executed in a lockstep. However, GPUs allow the temporary divergence of the execution flow of warps, so that a part of the threads can execute different portions of code. When this situation occurs (e.g., due to an IF/THEN/ELSE statement), a part of the threads get stalled, waiting for reconvergence. Although this mechanism provides the developer with the freedom of temporarily abandoning the SIMD paradigm, it can potentially lead to the complete serialization of a warp execution, strongly impairing the performance. For this reason, conditional branches should be avoided as much as possible. Since all ODEs can have a different

structure, they consider different variables (i.e., concentrations of the molecular species) which might have different exponents, all threads can, in principle, take a divergent thread and lead to serialization.

In order to mitigate this problem, before the numerical integration takes place, LASSIE reorganizes the ODEs according to the structure of their expressions. Specifically, LASSIE sorts the ODEs by their length, that is, according to the number of terms that appear in the right-hand side of each ODE. The goal of this heuristic is to reduce the serialization of warps, increasing the overall performances.

3 Results

The computational performance of LASSIE was evaluated by simulating a set of synthetic models of BRNs of increasing size, generated with the methodology used in [8]. These BRN models are characterized by a number of species N and of reactions M equal to $(64 \times 64), (128 \times 128), (256 \times 256), (512 \times 512), (704 \times 704), (960 \times 960), (1216 \times 1216), (1472 \times 1472), (1728 \times 1728), (1984 \times 1984), (2240 \times 2240), (2496 \times 2496), (2752 \times 2752), (3008 \times 3008), (5760 \times 5760), (10000 \times 10000)$. For each model, the kinetic constants were randomly sampled with uniform distribution in $(0, 1)$.

Figure 1 shows the running time necessary to carry out a deterministic simulation of the various synthetic models by using a CPU Intel i7-4790K 4.00GHz, a dual-GPU Nvidia GeForce GTX Titan Z equipped with $2 \times 2880 = 5760$ CUDA cores and 12GB of memory, and a Nvidia Tesla K20c equipped with 2496 CUDA cores and 5GB of memory. As shown in the plot, the running time of the CPU linearly increases with the size of the models. On the contrary, both GPU video cards are characterized by an almost constant running time, irrespective of the size of the models when the number of ODEs is fewer than the available cores, as shown by the tailored test with 10000 chemical species. In particular, the GeForce GTX Titan Z allows to obtain the best results—thanks to its higher number of CUDA cores with respect to the Tesla K20c—with a speed-up of $7.8\times$ with respect to the CPU. This is due to the fact that ODEs are outnumbered by the available cores, so that calculations can be completely executed in a parallel fashion, without any stalling due to blocks scheduling. It is worth noting that the break-even between the CPU and the GeForce GTX Titan Z is observed when the number of ODEs to be integrated is around 1200.

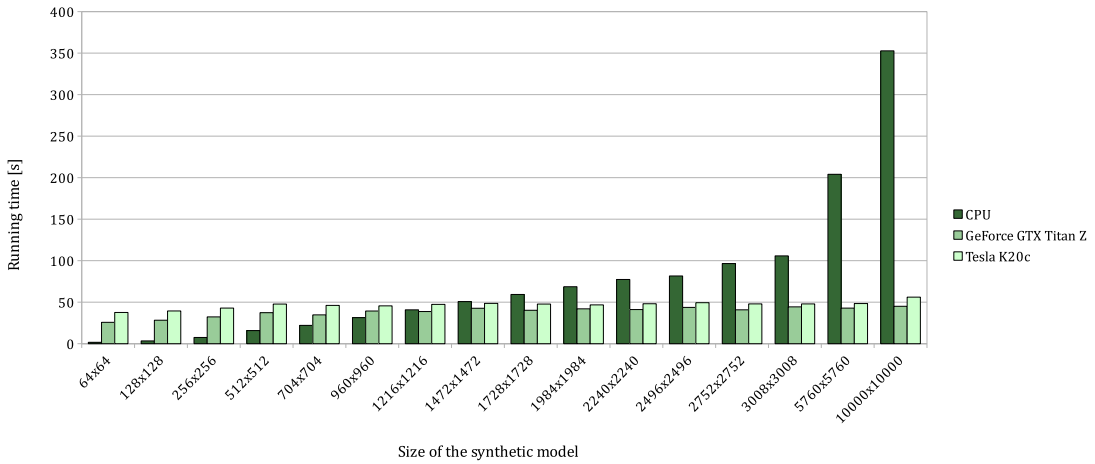


Figure 1: Comparison of the running times achieved by LASSIE on CPU Intel i7-4790K 4.00GHz, Nvidia GeForce GTX Titan Z and Nvidia Tesla K20c for the simulation of synthetic models of increasing size, having a number of species and of reactions $N \times M$ as specified on the x-axis.

4 Conclusion

We compared the performances of CPU and GPU to execute fine-grain deterministic simulations of the dynamics of large-scale BRNs of synthetic models. Our results highlight that our GPU-powered tool outperforms the CPU to simulate mechanistic models consisting in more than 1200 ODEs. However, the speed-up could be smaller by using a multi-threaded CPU simulator: we plan to exploit OpenMP, in the future, for a fair comparison between the architectures. We also plan to replace the ODE-ordering step with a more sophisticated clustering technique, that we expect will further reduce the serialization of threads execution [6]. To be more precise, the ODEs clustering organizes ODEs with similar structures within the same warp, thus avoiding as much as possible the conditional branching, which impairs the performance of the simulator. Second, we plan to optimize the GPU code in order to fully exploit the GPU memory hierarchy and reduce the accesses to the global memory that strongly affect the performance. Third, we will implement on the GPU the Runge-Kutta-Fehlberg algorithm [7], an alternative numerical integration method characterized by adaptive step-size, which allows better performances and simulation quality with respect to classic RK4.

For a thorough assessment of the computational performance of LASSIE, we plan to apply it to simulate large-scale models of real biological systems, such as the ErbB model presented in [3], which consists in more than 1000 ODEs. In particular, we believe that LASSIE will be especially useful for advanced computational analysis of rule-based models [4]. In addition, considering the ongoing research in developing whole-cell models at a detailed level of molecular description [5], we hope that our tool will provide a parallel solution to simulate real models of ever increasing size.

Finally, we plan to include LASSIE within the GPU-based toolbox that we are currently developing. We expect that this innovative, efficient and usable toolbox will obtain a beneficial adoption by the Systems Biology community, and help researchers to easily predict the still unknown behaviors of biological systems and to design more focused experiments in a very reduced time with respect to standard CPU analyses.

References

[1] J. Ackermann, P. Baecher, T. Franzel, M. Goesele, K. Hamacher. “Massively-parallel simulation of biochemical systems”. Proceedings of Massively Parallel Computational Biology on GPUs, Jahrestagung der Gesellschaft für Informatik e.V, pp. 739–750, 2009.

[2] B.B. Aldridge, J.M. Burke, D.A. Lauffenburger, P.K. Sorger. “Physicochemical modelling of cell signalling pathways”. Nature Cell Biology, vol. 8, no.11, pp.1195-1203, 2006.

[3] W.W. Chen, B. Schoeberl, P.J. Jasper, M. Niepel, U.B. Nielsen, D.A. Lauffenburger, P.K. Sorger. “Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data”. Molecular Systems Biology, vol. 5:239, 2009.

[4] L.A. Chylek, L.A. Harris, C.-S. Tung, J.R. Faeder, C.F. Lopez, W.S. Hlavacek. “Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems”. WIREs Systems Biology and Medicine, vol. 6, no.1, pp.13–36, 2014.

[5] J.R. Karr, J.C. Sanghvi, D.N. Macklin, M.V. Gutschow, J.M. Jacobs, B. Bolival Jr., N. Assad-Garcia, J.I. Glass, M.W. Covert. “A whole-cell computational model predicts phenotype from genotype”. Cell, vol. 150, no.2, pp.389–401, 2012.

[6] S. Lei, A.Y. Allidina, K. Malinowski. “Clustering technique for rearranging ODE systems”. In: M. G. Singh, A. Y. Allidina, B. K. Daniels eds. “Parallel Processing Techniques for Simulation” pp. 31–43, Springer US, 1986.

[7] J.H. Mathews, K.K. Fink. “Numerical Methods Using Matlab”, 4th Edition, Prentice-Hall Inc., Upper Saddle River, New Jersey, USA, 2004.

[8] M.S. Nobile, P. Cazzaniga, D. Besozzi, D. Pescini, G. Mauri. “cuTauLeaping: A GPU-powered tau-leaping stochastic simulator for massive parallel analyses of biological systems”. PLoS ONE, vol. 9, no.3:e91963, 2014.

[9] M.S. Nobile, D. Besozzi, P. Cazzaniga, G. Mauri. “GPU-accelerated simulations of mass-action kinetics models with cupSODA”. Journal of Supercomputing, vol. 69, no.1, pp.17–24, 2014.

[10] Y. Zhou, J. Liepe, X. Sheng, M.P.H. Stumpf, C. Barnes. “GPU accelerated biochemical network simulation”. Bioinformatics, vol. 27, pp. 874-876, no. 6, 2011.

An Hadoop-based algorithm for clustering Protein Structures

Giacomo Paschina[(1)], Luca Roverelli[(1)], Daniele D'Agostino[(1)], Federica Chiappori[(2)], Ivan Merelli[(2)]

(1) Institute of Applied Mathematics and Information Technologies, Italian National Council of Research

(2) Institute of Biomedical Technologies, Italian National Council of Research

Keywords: Hadoop, Clustering protein structures, Molecular Dynamics, Data parallel.

Abstract. While analysing large datasets of protein structures, derived from NMR experiments or molecular dynamics simulations, a good approach is to cluster structures in order to infer common and driving characteristics shared by different structural conformations. A common problem of the software that implement protein clustering is the scalability of the performance, in particular concerning the data load into memory. In this work we show how it is possible to improve the parallel performance of the GROMOS clustering algorithm by using Hadoop. The preliminary results show the validity of this approach, providing an hint for future development in this field.

1 Scientific Background

Many protein-structure prediction tools and protein-folding simulation software generate a large ensemble of candidate structures using different starting conditions. In particular, molecular dynamics simulations produce trajectories of atomic positions, velocities, and energies as a function of time and provide a sampling of the accessible conformational of a given macromolecule. As simulations on the 100ns - 1μs time scale are becoming routinely, with sampled configurations stored on the picosecond time scale, the resulting trajectories contain large amounts of data.

Data-mining techniques, like clustering, provide a valuable tool to make sense of the information in these trajectories. In particular, by clustering these structures, the overall consistency and accuracy of the final predictions can be increased, since conformations more frequently assumed during a trajectory, are representative of the structure. This explains why clustering is a widely used technique in structural biology.

¿From the computational point of view this process is very time consuming, since there are dozens or even hundreds of thousand structures to compare. Many algorithms have been implemented in parallel in order to overcome this problem, such as MAX_CLUST [1] and FAST_PROTEIN_CLUSTER [2]. The latter is based on an efficient GPU-accelerated solution, although not specifically designed to analyse molecular dynamics trajectories.

Nonetheless, analysing the scalability of these software, a clear bottleneck emerges, that is the large number of I/O operations necessary for data acquisition and possible use of virtual memory when processing these large structures. This is particularly true for bunches of PDB structures, which are text files, while trajectories are usually stored in binary files. It is therefore clear that a large amount of computational time is spent in reading the structures for the clustering operation.

Our idea to solve this problem is to exploit Hadoop and MapReduce in order to parallelize the clustering on partitions of the original dataset. To implement this idea we started from the GROMOS [3] algorithm for clustering protein structures, which is provided as part of the GROMACS [4] package. Although GROMOS is not parallel in its released implementation, it is widely used to cluster conformation from molecular dynamics simulations. The algorithm is fast, but it suffers, as introduced, of a slow loading

of the structures in memory. In the following we present a preliminary implementation of a solution based on the Hadoop-MapReduce architecture and the preliminary results.

The remainder of this work is structured as follows. In Section 2 the clustering problem is presented along with a real-life test case. In Section 3 the implementation is sketched. Section 4 presents the results achieved in terms of performance, while in Section 5 some conclusions are drawn.

2 Material and Methods

Clustering involves partitioning models into sets of similar structures. The input of these algorithms is a distance matrix in which are reported, for each sampled structural conformation of the molecular dynamics trajectory, the RMSD distances, time frame after time frame, of the C_α atoms that compose the backbone of the protein. The GROMOS algorithm relies on the nearest neighbour algorithm, which is a non-parametric method used for classification and regression [5]. In particular, the GROMOS implementation of this algorithm is an iterative process:

- the neighbours of each data point are defined according to a cut-off distance c;
- the point with largest neighbourhood defines the "best" cluster, corresponding to a stable conformation of the protein;
- all the points belonging to the cluster are removed ;
- the algorithm is iterated until all data have been assigned to a cluster and removed.

In the context of this work, the input dataset for testing the clustering algorithm consisted of a molecular dynamics (MD) trajectory of the human protein Hsp70 in complex with ADP ligand. Hsp70 is a molecular chaperone, which prevents the incorrect folding of other proteins, or it is involved in protein translocation though the mitochondrial membrane [6]. This protein belongs to a large protein family, which is present in different organisms, from bacteria to human. Hsp70 displays two active conformations the closed bound to ADP and the open bound to ATP. Due to the absence of crystallized human structure, simulated protein was build with homology modelling based on bacterial DnaK in ATP conformation [7]. The analysis consists in evaluating the effect of the nucleotide exchange (ATP to ADP) on protein conformation (open to closed). To identify the more stable conformation, that is the conformation assumed more frequently during the trajectory, cluster analysis is a useful tool.

In detail, the complex has in about 600 atoms. To evaluate the protein conformation two independent trajectories of 25ns and 100ns of MD simulation of the solvated complex were obtained with Gromacs 4.0 [8]. Both trajectories were skipped every 50 frames, and we obtained two trajectories of 500 and 2000 conformations. The two files have a size of, respectively, 200 and 800 MBs, while the original one has a size of about 40 GB. Cutoff was set at 0.5 nm, in agreement with the number of atoms to be clustered and to the protein conformation type. This is also the default value for GROMOS: higher values result in less cluster, composed buy a larger number of points and vice versa for smaller values.

At the end of clustering analysis we obtain 30 clusters distributed as shown in Figure 1. The fist three clusters can be considered representative for this simulation. As shown in Figure 2 the central structure of the first cluster obtained is an intermediate conformation between the open and the closed one, showing that molecular dynamics simulation is a useful tool to evaluate a conformational change and clustering allow to identify a representative conformation.

The implemented MapReduce algorithm consists of the following stages of both parallel and sequential processing.

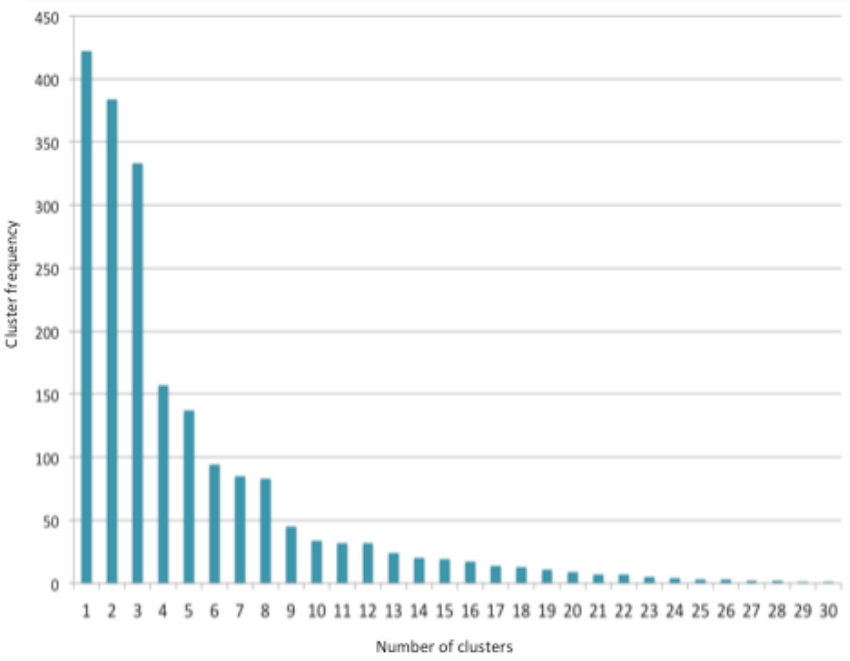


Figure 1: Histogram of cluster population distribution in the Hsp70 example with 2000 conformations.

Data Structure and preprocessing Figure 3 shows the data structure contained in the trajectories file. Each row in the file contains the atom id, the atom type, the residue name, the chain to which it belongs, the residue number, the x, y, z coordinates, the occupancy, and the B-factor. The idea is to recover in parallel the trajectories related to the atoms involved in the computation and re-create a data file with only the data of interest, in order to avoid transfers of large amounts of data through the network between the nodes in the Hadoop cluster during runtime. The trajectory file is split into several parts and each of them is given to a Map task that retrieves data relying on the type of atom of interest (step A). At the end of the elaboration, each Map task sends the data to a Reducer task which, after waiting the termination of all Map task, writes on HDFS the reduced trajectories file containing only the data of interest.

Retrieval of the coordinates from trajectory files Subsequently, every Map task reads portions of the new trajectories file and associates to each atom the x, y, z coordinates assumed at a given time instant (step B). At the end of the Map tasks, a Reducer collects all data and creates a new single file containing the coordinates of all atoms for each frame. The data structure used by Map Task to temporary store the coordinates consists of an id that identifies the atom and a list in which each node is formed by an object that wraps the coordinates taken by the atom at a given time instant. In this way we get a list of nodes in which each node represents a time frame with inside the coordinates taken from the atom at that instant.

Creation of distances matrix Once we have created the file containing all the coordinates of all atoms for each frame as described in step B, new Map tasks are performed to calculate the Euclidean distance of the positions taken by each atom at each time instant (step C). After completing these tasks, a Reducer collects data by summing the various distances of all atoms at each time frame to obtain the variation of the centre of gravity of the entire molecule with respect to all the others frames .

Clusterization The distances matrix resulting from step C is processed by several Map tasks: the matrix is divided in different parts and each part is elaborated by a single Map task that produces partial frame clusters on the basis of a cutoff established

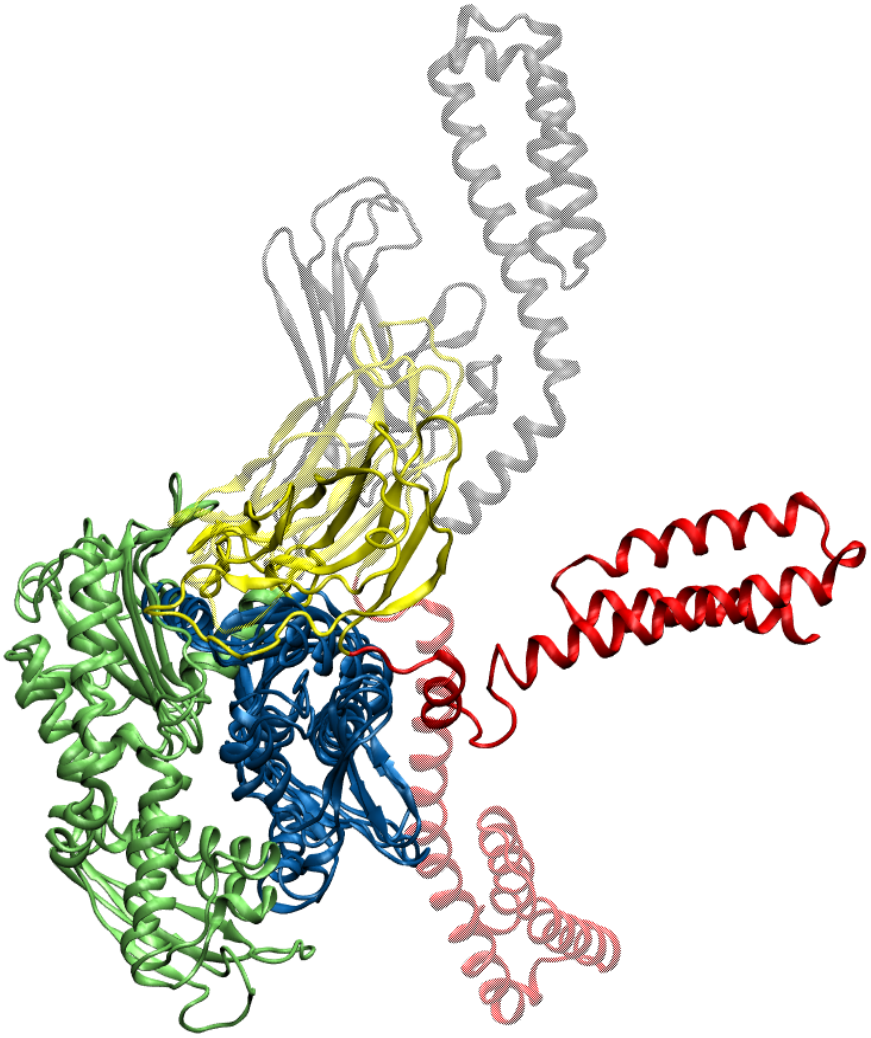


Figure 2: Central structure of the first cluster. Hsp70 sub-domains are shown in different colours for the open conformations, in solid colours the central structure of the first cluster, in transparent colours the starting conformation; in transparent grey the closed conformations.

a priori (in our case it is set to 0.5 nm, as said before). A Reducer task manages the data and defines the final frame clusters, selecting the biggest one, and removing the frame associated with the largest cluster from the other ones (step D). The elaboration is repeated until clusters consume themselves. In this case, the processing is the union of a parallel part (the elaboration of parts of matrix by Map tasks) and a serial part (the reorganization of clusters and the cycle until the end of the clusters by the Reducer task).

3 Results

The computer cluster named *imatihpc* has been used to experiment our Hadoop based solution and it is composed by three computational nodes and a front-end node, each of those presents the following hardware configuration: two 6-core Intel Xeon E5645 CPUs, 32 GB of RAM, 2 TB of SATA hard disk. Nodes are linked together with a Gigabit network connection. The Hadoop cluster follows the physical configuration of the *imatihpc* cluster on which it is implemented: the front-end node acts as the master while the 3 nodes used for the computation act as a slave, with an HDFS filesystem of 3 TB.

Performances are shown in Table I. They are based on the average on 10 executions. The number of parallel processes in Hadoop is determined by the framework. Typically,

```
ATOM     33  MG    MG A   2      46.670  61.020  48.290  1.00  0.00
ATOM     34  N    LYS B   3      64.740  68.010  60.640  1.00  0.00
ATOM     35  H1   LYS B   3      65.620  67.730  61.030  1.00  0.00
ATOM     36  H2   LYS B   3      64.650  68.990  60.800  1.00  0.00
ATOM     37  H3   LYS B   3      64.030  67.520  61.150  1.00  0.00
ATOM     38  CA   LYS B   3      64.660  67.690  59.210  1.00  0.00
ATOM     39  CB   LYS B   3      64.890  66.200  58.960  1.00  0.00
ATOM     40  CG   LYS B   3      66.330  65.800  58.620  1.00  0.00
ATOM     41  CD   LYS B   3      67.320  66.190  59.720  1.00  0.00
ATOM     42  CE   LYS B   3      68.760  65.730  59.520  1.00  0.00
ATOM     43  NZ   LYS B   3      68.940  64.300  59.820  1.00  0.00
ATOM     44  HZ1  LYS B   3      69.910  64.060  59.810  1.00  0.00
ATOM     45  HZ2  LYS B   3      68.590  64.100  60.740  1.00  0.00
ATOM     46  HZ3  LYS B   3      68.480  63.720  59.150  1.00  0.00
ATOM     47  C    LYS B   3      63.320  68.090  58.610  1.00  0.00
ATOM     48  O    LYS B   3      62.340  68.210  59.350  1.00  0.00
ATOM     49  N    ILE B   4      63.250  68.320  57.300  1.00  0.00
ATOM     50  H    ILE B   4      64.060  68.320  56.710  1.00  0.00
ATOM     51  CA   ILE B   4      62.000  68.690  56.610  1.00  0.00
ATOM     52  CB   ILE B   4      62.260  69.480  55.330  1.00  0.00
```

Figure 3: The data structure within trajectories file.

Table 1: Elapsed times (Sec) of the sequential and MapReduce versions of the Gromos algorithm subdivided in the steps described in Section IV.

	500			2000		
	A	B	C+D	A	B	C+D
Sequential	3.40	1.2	54.75	19.38	5.06	2184.10
MapReduce	1.70	0.6	41.88	1.8	0.49	312.50

a MapTask is created for each map split, and there is a map split for each input file or, for files larger than the data block size of HDFS, with 64MB as default value, there is a MapTask for each data block. This means that in the first test (500 conformations) 2 MapTask are launched, and 13 in the second test (2000 conformations). The number of ReduceTasks is normally set to 1, in order to create the result at a global level and to produce results also when one or more nodes of the cluster fail.

The first step of the algorithm (A - Data Structure and pre-processing) achieves good results because a good overlapping between the map and reduce tasks. Similar considerations hold true for the second step (B - Retrival of the coordinates from trajectories file). In both cases the speedup in the first case is around 2, in the second around 10. The last two steps (C - Creation of distances matrix, D - Clusterization) instead suffer of the framework overheads and of the bottleneck due to the reduce task, presenting therefore a speedup of 1.3 and 7 respectively.

4 Conclusions

Structure clustering is very important in the analysis of Molecular Dynamics trajectories, since it allow to identify stable conformations. The number of structures to analyse in this kind of analysis is very high because the increasing computational power allow to generate very large simulations, which poses important scalability problems to clustering algorithm. Here we present an Hadoop based solution to accelerate the loading of data into memory, which is the real bottleneck of this algorithm. The presented results show interesting performance figures, which support the use of Hadoop when analysing very large data files.

Acknowledgments

This paper has been supported by the Italian Ministry of Education and Research (MIUR) through the Flagship (PB05) InterOmics, HIRMA (RBAP11YS7 K), and the European MIMOMICS projects.

References

[1] MaxCluster - A tool for Protein Structure Comparison and Clustering, http://www.sbg.bio.ic.ac.uk/ maxcluster

[2] L.H. Hung, R. Samudrala, "fast_protein_cluster: parallel and optimized clustering of large-scale protein modeling data", *Bioinformatics*, 15;30(12):1774-6, 2014.

[3] X. Daura, K. Gademann, B. Jaun, D. Seebach, W.F. van Gunsteren, A.E. Mark, "Peptide Folding: When Simulation Meets Experiment", *Angewandte Chemie International Edition*, 38(1-2):236240, 1999.

[4] H.J.C. Berendsen, D. van der Spoel, R. van Drunen, "GROMACS: A message-passing parallel molecular dynamics implementation" *Comp Phys Comm.*, 91:4356, 1995.

[5] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression", *The American Statistician*, 46 (3): 175185, 1995.

[6] M.P. Mayer, B. Bukau, "Hsp70 chaperones: cellular functions and molecular mechanism", *Cell Mol Life Sci.*, 62(6):670-84, 2005.

[7] R. Kityk , J. Kopp, I. Sinning, M.P. Mayer, Structure and dynamics of the ATP-bound open conformation of Hsp70 chaperones. *Mol Cell.* 28;48(6):863-74, 2012.

[8] D. Van Der Spoel, E. Lindahl, B. Hess,G. Groenhof, A.E. Mark, H.J.C. Berendsen, "GROMACS: Fast, flexible, and free", *J. Comput. Chem.*, 26:17011718, 2005.

[9] I. Merelli, H. Prez-Snchez, S. Gesing, D. DAgostino, "Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives", *BioMed Research International*, 134023, 2014.

[10] D. Eadline, "Is Hadoop the New HPC?", http://www.admin-magazine.com/HPC/Articles/Is-Hadoop-the-New-HPC

CiBB2015

Special session on

Multi Omic metabolic models and statistical Bioinformatics of a daptations and biological associations

Organisers

Dr. Claudio Angione, Computer Laboratory, University of Cambridge.
Dr. Pietro Liò, Computer Laboratory, University of Cambridge
Dr. Sandra Pucciarelli, University of Camerino.
Dr. Barbara Simionati , BMR Genomics.

This Session has been sponsored by MSCA-RISE MeTABLE Project

MARIE CURIE ACTIONS

METABLE
METAGENOME BIOCATALISTS EVOLUTION

Measuring adaptation to extreme environments with a multi-omic approach

Claudio Angione[(1)], Sandra Pucciarelli[(2)], Barbara Simionati[(3)], Pietro Lió[(1)]

(1) Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, CB3 0FD Cambridge, UK

(2) School of Biosciences and Veterinary Medicine, University of Camerino
via Gentile III da Varano, 62032, Camerino (MC), Italy

(2) BMR Genomics
Via Redipuglia 21A, 35131 Padova, Italy

Keywords: multi-omic, adaptation, metabolism.

Abstract. During the past decade, more than 4000 organisms have been subjects of genome-sequencing projects. This has enabled the retrieval of information such as evolutionary relationships among all living organisms, and the understanding of complex phenomena, including evolution, adaptation, and ecology. Bioinformatics tools have allowed us to perform genome annotation, cross-comparison, and to understand the metabolic potential of living organisms. In the last few years, research in bioinformatics started to evolve from the analysis of genomic sequences and structural biology problems, to the analysis of complex post-genomic interaction networks. The main outcome of this effort has been the creation of: (i) databases and ontologies of protein class and cell components; (ii) repositories of models of cell processes, through the definition of common exchange formats such as the Systems Biology Markup Language (SBML); (iii) simulation tools; (iv) Multi-scale modeling (micro-macro biology). However, the following questions remain open: what is the real functioning scheme of the cell? What is the metabolic influence of changes in gene expression? How to understand the metabolic interactions among living organisms in a ecosystem? How to understand evolutionary mechanisms of adaptation? Here, we describe a pipeline that summarizes the methodologies recently proposed to address these questions.

Introduction

This paper presents a representation of the Bioinformatics workout for molecular adaptation. Our design is divided into distinct methods/software blocks (Fig. 1), further analyzed in the following sections. The sequence of functional blocks leads to the identification of pathways, genes and proteins involved in adaptation. Each block in the proposed pipeline contains distinct methodologies which could be implemented in one or more existing software tools (developed by the authors or by other groups). For sake of clarity, each block is numbered and described below in the paper.

The strengths of our design are the following: (i) use and calibration of multi omic information; (ii) use of pathway information; (iii) machine learning, bioinformatics and multi-objective optimization are integrated in a powerful and novel inferential engine.

1 – Sampling

One of the most important parts in studying environmental adaptation is the collection of data needed to perform multi-omic analysis. It is important to carefully plan the number and location for up-taking environmental samples that will be collected, in order to avoid a number of gaps for future analysis. The analysis of metagenomes from samples collected in extreme environments, such as volcanoes, glaciers, or deep ocean waters (Fig. 1, panel 1) represents a valuable resource to study the molecular

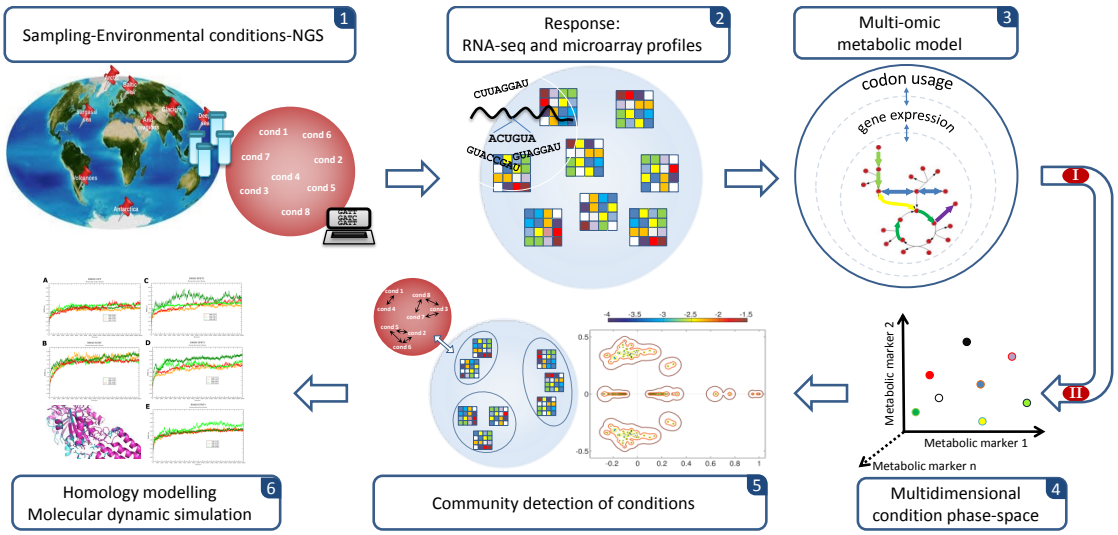


Figure 1: Sampling in extreme ecosystems and Bioinformatics methodological applications. B). The response to environmental conditions (Arctic and Antarctic regions, glaciers, deep ocean seawater, volcanoes and arid areas) is sampled for different associated species (1), and measured through expression profiling (2). To evaluate the environmental conditions and detect their community structure, a multi-omic model (3) can be applied to the species' metabolism, taking into account gene expression and codon usage. The model is able to associate each environmental condition with a single point in a multidimensional condition phase-space (4). Statistical estimators defined on the multi-omic model can be used to investigate the pathway basis of the relationships between species in the association (5). Finally, homology modeling and molecular dynamic simulation can be applied to calculate structure flexibility and binding affinity of molecules of interest at different temperatures (6).

mechanisms underlying environmental adaptation. According to the extreme environment under consideration, different molecular, physiological and phenotypic strategies can be unraveled by applying multi-omic approaches. For example, a comprehensive survey of the distribution of bacteria from 213 samples, generated from 60 stations along the horizontal and vertical salinity gradients of the Baltic Sea, represented the first detailed taxonomic study of an indigenous brackish water microbiome composed by a diverse combination of freshwater and marine clades that appears to have adapted to the brackish conditions [Herlemann et al., 2011]. Furthermore, by applying "whole-genome shotgun sequencing" to microbial populations collected en masse from seawater samples collected from the Sargasso Sea near Bermuda, it was possible to discover 148 previously unknown bacterial phylotypes and to identify over 1.2 million previously unknown genes, suggesting substantial oceanic microbial diversity [Venter et al., 2004].

2-3 – Response and multi-omic models

Several computational algorithms have been developed to analyze gene expression profiles. The main goals are detecting dependencies among genes over different conditions and unraveling gene expression programs controlled by the dynamic interactions of hundreds of transcriptional regulators [Faith et al., 2007]. By combining network inference algorithms and experimental data derived from 445 Escherichia coli microarrays, Faith et al. identified 1,079 regulatory interactions, 741 of which were new regulators of amino acid biosynthesis, flagella biosynthesis, osmotic stress response, antibiotic resistance, and iron regulation. This approach contributed to the understanding on how organisms can adapt to changing environments. Furthermore, [Angione et al., 2015] recently proposed a hybrid method combining multi-omic flux-balance analysis (FBA) and Bayesian inference, with the aim of investigating the cellular activities of a bacterium from the transcriptomic, fluxomic and pathway standpoints under different environmental conditions. The authors integrate an augmented FBA model of E. coli and a Bayesian factor model to regard pathways as latent factors between environmental

conditions and reaction rates. Then, they determine the degree of metabolic pathway responsiveness and detect pathway cross-correlations. They also infer pathway activation profiles as a response to a set of environmental conditions. Finally, they use time series of gene expression profiles combined with their hybrid model in order to investigate how metabolic pathway responsiveness vary over time.

4-5 – Community detection of environmental conditions

In two research works, Taffi and colleagues proposed a computational framework that integrates bioaccumulation information at the ecosystem level with genome-scale metabolic models of degrading bacteria [Taffi et al., 2014, Taffi et al., 2015]. The authors applied their methods to the case study of the polychlorinated biphenyls (PCBs) bioremediation in the Adriatic food web. Remarkably, they were able to discover species acting as key players in transferring pollutants in contaminated food web. In particular, the role of the bacterial strain *Pseudomonas putida* KT2440, known to be able to degrade organic compound, in the reduction of PBCs in the trophic network, was assessed in different scenarios. Interestingly, one aspect of their analysis involved a scenario computed by using a synthetic strain of *Pseudomonas* performing additional aerobic degradation pathways. Combining these computational tools allows designing effective remediation strategies, and provides at the same time insights into the ecological role of microbial communities within food webs.

In the past decade, genome-scale metabolic modelling has been successfully applied also for studying large-scale metabolic networks in microbes, with the aim of guiding rational engineering of biological systems, with applications in industrial and medical biotechnology, including antibiotic resistance [Milne et al., 2009]. Even though antibiotics remain an essential tool for treating animal and human diseases in the 21st century, antibiotic resistance among bacterial pathogens has garnered global interest in limiting their use, and to provide actionable strategies to search and support development of alternative antimicrobial substances [Nolte, 2014]. It is interesting to note that bacterial strains such as *Arthrobacter sp.* and *Gillisia sp.* CAL575, producing an array of molecules with potential antimicrobial activity vs human pathogenic *Burkolderia cepacia complex* strains [Orlandini et al., 2014, Fondi et al., 2012, Maida et al., 2014] were isolated from Antarctica. These strains represent useful models to unravel metabolic pathways responsible for the production of bioactive primary and/or secondary metabolites [Orlandini et al., 2014].

6 – Homology modelling

Computational methods such as homology modeling and molecular dynamic simulation can be employed for protein engineering and design. Directed evolution of enzymes and/or bacterial strains can be exploited for industrial processes [Adrio and Demain, 2014]. Protein modelling and molecular dynamic simulation can be applied to molecules from psychrophilic organisms to unravel the molecular mechanisms responsible for cold-adaptation. In the work of [Chiappori et al., 2012] a computational structural analysis based on molecular dynamics (MD) was performed for three β-tubulin isotypes from the Antarctic psychrophilic ciliate Euplotes focardii. Tubulin eterodimers (the building block of microbules composed of α-tubulin and β-tubulin) from psychrophilic eukaryotes can polymerize into microtubules at 4C, a temperature at which microtubules from mesophiles disassemble. The structural analysis based on MD indicated that all isotypes from E. focardii display different flexibility properties in the regions involved in the formation of longitudinal and lateral contacts during microtubule polymerization, with respect those from mesophilic organisms. A higher flexibility of these regions may facilitate the formation of lateral and longitudinal contacts among heterodimers for the formation of microtubules in an energetic unfavourable environment. Overall, homol-

ogy modelling plays a major role in the generation of testable hypothesis.

Conclusion

In this paper, we discussed the contribution of ”omics” technologies to understand many biological phenomena, from genotypes to the intricate mechanisms influencing the phenotypes of all living organisms, in particular in response to stress or for environmental adaptation. To date, a great deal of biological information has already been acquired through application of individual 'omics' approaches. However, ”multi-omic technology” will enable the integration of knowledge at different levels: from genes to proteins, from pathways to metabolic fluxes. This approach will be pivotal for understanding how the individual components in the system interact and influence the overall metabolism, phenotype, and adaptation. Multi-omic analyses represent a novel and powerful tool to generate discrete and testable biological hypotheses, with a possible framework to guide the design of systems biology experiments.

References

[Adrio and Demain, 2014] Adrio, J. L. and Demain, A. L. (2014). Microbial enzymes: Tools for biotechnological processes. *Biomolecules*, 4(1):117–139.

[Angione et al., 2015] Angione, C., Pratanwanich, N., and Lió, P. (2015). A hybrid of metabolic flux analysis and bayesian factor modeling for multi-omics temporal pathway activation. *ACS Synthetic Biology*, page DOI:10.1021/sb5003407.

[Chiappori et al., 2012] Chiappori, F., Pucciarelli, S., Merelli, I., Ballarini, P., Miceli, C., and Milanesi, L. (2012). Structural thermal adaptation of β-tubulins from the antarctic psychrophilic protozoan euplotes focardii. *Proteins: Structure, Function, and Bioinformatics*, 80(4):1154–1166.

[Faith et al., 2007] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8.

[Fondi et al., 2012] Fondi, M., Orlandini, V., Maida, I., Perrin, E., Papaleo, M. C., Emiliani, G., De Pascale, D., Parrilli, E., Tutino, M. L., Michaud, L., et al. (2012). Draft genome sequence of the volatile organic compound-producing antarctic bacterium arthrobacter sp. strain tb23, able to inhibit cystic fibrosis pathogens belonging to the burkholderia cepacia complex. *Journal of bacteriology*, 194(22):6334–6335.

[Herlemann et al., 2011] Herlemann, D. P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., and Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the baltic sea. *The ISME journal*, 5(10):1571–1579.

[Maida et al., 2014] Maida, I., Fondi, M., Papaleo, M. C., Perrin, E., Orlandini, V., Emiliani, G., de Pascale, D., Parrilli, E., Tutino, M. L., Michaud, L., et al. (2014). Phenotypic and genomic characterization of the antarctic bacterium gillisia sp. cal575, a producer of antimicrobial compounds. *Extremophiles*, 18(1):35–49.

[Milne et al., 2009] Milne, C. B., Kim, P.-J., Eddy, J. A., and Price, N. D. (2009). Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnology journal*, 4(12):1653–1670.

[Nolte, 2014] Nolte, O. (2014). Antimicrobial resistance in the 21st century: a multifaceted challenge. *Protein and peptide letters*, 21(4):330–335.

[Orlandini et al., 2014] Orlandini, V., Maida, I., Fondi, M., Perrin, E., Papaleo, M. C., Bosi, E., de Pascale, D., Tutino, M. L., Michaud, L., Giudice, A. L., et al. (2014). Genomic analysis of three sponge-associated arthrobacter antarctic strains, inhibiting the growth of burkholderia cepacia complex bacteria by synthesizing volatile organic compounds. *Microbiological research*, 169(7):593–601.

[Taffi et al., 2014] Taffi, M., Paoletti, N., Angione, C., Pucciarelli, S., Marini, M., and Liò, P. (2014). Bioremediation in marine ecosystems: a computational study combining ecological modeling and flux balance analysis. *Frontiers in genetics*, 5.

[Taffi et al., 2015] Taffi, M., Paoletti, N., Liò, P., Pucciarelli, S., and Marini, M. (2015). Bioaccumulation modelling and sensitivity analysis for discovering key players in contaminated food webs: the case study of pcbs in the adriatic sea. *Ecological Modelling*, 306:205–215.

[Venter et al., 2004] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., et al. (2004). Environmental genome shotgun sequencing of the sargasso sea. *science*, 304(5667):66–74.

Proteome semantics application of natural language processing to peptide mass fingerprinting

Antonio Starcevic[(1)]

(1) SemGen Ltd. Lanite 5/D, 10000 Zagreb, Croatia,

Keywords:

Abstract. Peptide mass fingerprinting is a term which describes technique which utilizes ESI-TOF or MALDI-TOF mass spectrometer to accurately measure and then compare the masses of the peptides of the unknown protein to the theoretical peptide masses of each protein encoded in the genome. This technique has become a cornerstone for protein identification. Today, applications using peptide mass fingerprinting in biomedical analyses are a major driving force behind its rapid development. However, efficient and accurate analyses of usually large protein tandem mass spectrometry data sets require specialized software. In terms of final goal, which is data interpretation, the role of software and underlying algortihms is at least equally important as the technique itself, a fact which is often neglected. High-throughput mass spectrometry instruments can readily generate hundreds of thousands of spectra. This fact combined with the ever growing size of genomic databases imposes tremendous demands for potential successful softvare solutions. In fact, it is the process of comparing large-scale mass spectrometry data with enormous databases that remains the major bottleneck in proteomics. We propose a paradigm shift in form of a completely novel approach based on natural language processing which is not just another improvement of existing methods and algortihms. Instead of relying on peak intesity and using sequence alignments for database matching, we use newly developed concept of latent microbial proteome fingerptints for strain/species identification. Since our algorithm doesn't rely on sequence alignment but instead utilizes a concept of singular proteome fingerprints rather than sets of unrelated peptides, it offers an elegant solution for this most troubling step in proteome analyses.

Abandoning BLAST and other alignment based methods. results in far superior processing speed, accuracy and sensitivity. The above mentioned algortithm can be used to analyse not only proteomes but also metaproteomes coming from mixed microbe communities as in the case presented human urine samples taken from a hospital. The method itself is completely generic, not developed with any specific platform in mind, which makes it highly versatile, able to turn any existing device into highly efficient metaproteome analyzer without siginificant costs related to purchase of new equipment.

SNP-SHOT: Integrating Annotation Sources for Target Enrichment experiments

Ivano Zara[(1)], Ilena Li Mura[(1)], Andrea Telatin[(1)], Barbara Simionati[(1)]

(1) BMR Genomics, Italy

Keywords:

Abstract. Resequencing target genes of human (referred to as Target Enrichment) patients has become a common practice among geneticists and clinicians, both for the quick price drop and for the advantages in term of genomic privacy offered by the technique. The adoption rate of the method has also boosted the development of algorithms and platforms to perform the variant calling and then the functional annotation (and prioritizations) of variants. Variant annotation, in particular, is still a weak step, mostly because of the poor quality of public databases. Here we present SNP-SHOT, a variant annotation and query system focused on the establishment and maintenance of annotation databases. We worked on enrichment panels for cardiomyopathies, and imported as primary sources several Locus Specific DataBases (LSDB), most of the in the LOVD format, and noticed several discrepancies and errors in the data they report, thus adding several consistency filters in the import mechanism. SNP-SHOT database side also include the possibility of uploading user-specific database, with private set of variants and frequencies. The SNP-SHOT project thus spans from database update to single VCF upload and annotation. A web-based interface allows browsing and filtering variants of each individual, eventually comparing samples.

MODELLING METABOLIC ADAPTATIONS TO COLD SHOCK AND SUBSTRATES SWITCHING

Marco Fondi [(1)], Emanuele Bosi[(1)], Luana Presta[(1)], Pietro Lio'[(2)], Renato Fani[(1)]

(1) ComBo (Florence Computational Biology group), Dep. of Biology, University of Florence
Via Madonna del Piano 6, 50019 Sesto Fiorentino, Florence

(2) Computer Laboratory, Cambridge University
CB3 0FD Cambridge, UK

Keywords: Microbial adaptation, Metabolic Modelling, Multi-omics, FBA.

Abstract. Extremophilic microbes have adapted specific features to survive in ecological niches characterized by harsh conditions such as, for example, very low/high temperature. Studying the mechanisms exploited by these microorganisms to overcome the selective pressure acting in such ecological niches is stimulating from a basic research viewpoint and because of biotechnological applications. The Antarctic strain *Pseudoalteromonas haloplanktis* TAC125 is one of the model organisms of cold-adapted bacteria and is currently exploited as a new alternative expression host for numerous biotechnological applications. However, several features concerning its metabolic landscape remain obscure, including, for example, the effect of a temperature down shift on its overall metabolic activity. Moreover, this microbe does not possess a phosphoenolpyruvate-dependent phosphotransferase system for the transport and first metabolic step of carbohydrates degradation. The growth on arabinose and glycerol as sole carbon sources is also not possible. As a consequence, the volumetric product yields are still poor due to low cell densities and process development aimed at fed-batch optimized growth with this organism is a challenge. Here, we applied metabolic reconstruction and functional modelling to shed light on these topics. More in detail, a genome-scale metabolic model of *P. haloplanktis* TAC125 was reconstructed from genomic data and validated comparing constraints-based modeling outcomes with experimentally determined growth rates and large scale growth phenotype data (Phenotype Microarray). Furthermore, the model was used to globally investigate possible metabolic adjustments of *P. haloplanktis* TAC125 during growth at low temperature by means of robustness analysis and functional integration of protein abundance data into the reconstructed network. Finally, we simulated growth of *P. haloplanktis* TAC125 in a complex medium (resembling the one commonly during fed-batch experiments) and preliminary described the deep reprogramming of its overall metabolic landscape following the substrate switching observed *in vivo*. Our results represent a valuable platform for a further understanding of *P. haloplanktis* TAC125 cellular physiology at the system level and the design of more focused strategies for its possible biotechnological exploitation.

1 Scientific Background

Living organisms continuously need to adapt to fluctuating levels of nutrients, exogeneous toxic compounds and environmental stresses (e.g. temperature, pH etc.). In most cases, this adaptation process involves systemic changes at different cellular levels, including gene expression, protein synthesis regulation and metabolic reactions. Understanding the mechanisms underlying such phenotypic switches is one of the most intriguing and challenging topics in modern day biology and one of the key aims of systems biology. Genome scale metabolic modeling represents a valuable tool in this

context. Indeed, this *in silico* approach can be adopted to quantitatively simulate chemical reactions fluxes within the cell, including metabolic adjustments in response to external perturbations (e.g. temperature downshift). Genome annotations are usually transformed into models by defining the boundaries of the system, a biomass assembly reaction, and exchange fluxes with the environment [1]. Constraint-based modelling methods (e.g. Flux Balance Analysis, FBA) can then be used to compute the resulting balance of all the active cellular reactions in the cell and to simulate the maximal growth of a cell in a given environmental condition [2, 3].

We used constraint-based metabolic modeling and multi-omics integration for gaining a system-level understanding of the metabolic lanscape of the Antarctic bacterium *Pseudoaltermonas haloplanktis* TAC125 (PhTAC125). PhTAC125 has been isolated from sea water sampled along the Antarctic ice-shell, a permanently cold environment, it is capable of growing in a wide temperature range (4 to 25 C°) and its lowest observed doubling time was detected at 20 C°[4]. Unfortunately, the metabolic consequences of growth at low temperature are still largely unexplored, despite they migth reveal biologically relevant features in the context, for example of the consequences of global ocean warming on microbial life.

Furthermore, PhTAC125 has recently attracted the interest of biotechnologists, as it has been suggested as an alternative host for the soluble overproduction of heterologous proteins, given its capability to grow fast at low temperatures [5, 6, 7, 8]. Nevertheless the volumetric product yields are still poor due to low cell densities. The establishment of a fed-batch culture system relying on complex media [6] has revealed the presence of several metabolic switches during which PhTAC125 selectively uses one (or few) carbon source(s) to sustain growth. Again, the reprogramming of the whole metabolic network during the different growth phases in a complex medium is largely unknown, despite it may have to important biological/biotechnological drawbacks.

2 Materials and Methods

2.1 Metabolic reconstruction and modelling

An initial draft metabolic reconstruction of the strain PhTAC125 was constructed using RAST annotation system with default parameters [9] and then thoroughly inspected following the main steps listed in [1]. The Flux Balance Analysis (FBA) method was employed to simulate flux distribution in different conditions and under various simulated perturbation [2].

Up- and down-regulation ratios of protein expression were mapped onto the PhTAC125 metabolic model using MADE (Metabolic Adjustment by Differential Expression) [13]. The visualization of the changes in reaction fluxes in the two conditions was performed using iPath 2.0 [10].

2.2 Robustness analysis

Robustness analysis consists in the calculation of suboptimal cellular growth (using FBA) when the reaction flux of a given reaction is varied around the optimal value. In this context, we perturbed the flux through each reaction whose corresponding genes i) were included in the model and ii) showed a significant change in expression during proteomics experiments. Results of this analysis can be visualized as a plot of the reaction flux (x-axis) versus the cellular growth rates (y-axis).

3 Results

3.1 Model reconstruction and validation

The metabolic network of PhTAC125 was initially obtained from its genome annotation and integrated with additional functional information. The reconstructed genome-scale metabolic model (named iMF721) encompasses information on 721 ORFs (20.7%

of the PhTAC125 protein encoding genes), 1133 metabolites and 1322 reactions (Tab. 1).

Table 1: Experimental Data.

PhTAC125 model	
N. of genes	721
N. of reactions	1322
N. of metabolites	1133
Gene-associated reactions	1146

The model includes non-gene-associated reactions accounting for i) the biomass assembly reaction (which also takes into consideration non-growth-associated ATP costs), ii) 48 reactions which filled gaps in the metabolic network (19 added by the AUTOCOMPLETION function of the RAST annotation system, and 29 added during the manual evaluation of the model), iii) 85 exchange reactions allowing the simulation of external conditions (e.g. nutrients exchange) and iv) 17 spontaneous reactions. Additionally, during the gap-filling process sink and demand reactions were added to the model when necessary.

Accordingly, an *in silico* minimal growth medium was defined using exchange reactions present in the model and biomass optimization was selected as the model objective function (O.F.). More in detail, lower bounds of exchange reactions accounting for all the salts present in Schatz medium [11] were set to -1000 mmol/g*h-1, in order to mimic non-limiting conditions. Each of the aforementioned amino acids was then chosen as the unique carbon source of this in silico medium.

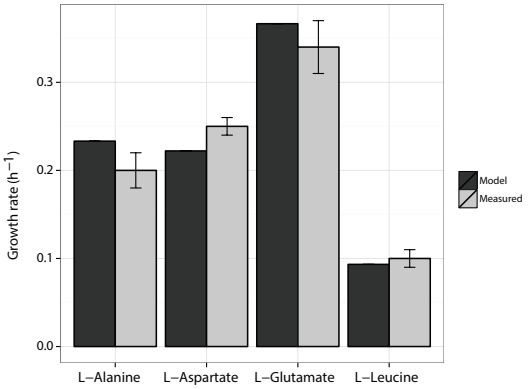


Figure 1: Comparison between model-predicted growth rates and experimentally determined ones on 4 different carbon sources: L-Alanine, L-Aspartate, L-Leucine and L-Glutamate.

The PhTAC125 enzymatic capacity for these compounds was calculated as the ratio of the growth rate to the biomass yield in batch experiments [2] and was set to 0.7, 3.6, 2.5 and 3.4 (mmol/g*h^{-1} for leucine, alanine, aspartate and glutamate, respectively. The predicted growth rates were compared to those experimentally determined for PhTAC125 (Fig. 1.), revealing an overall agreement between experimentally determined growth rates and *in silico* predictions. Furthermore, the iMF721 growth phenotypes predictions on a set of C sources were compared with large scale growth data obtained from Phenotype Microarray experiments. We found that in 84% of the cases (54 out of 64) the outcomes of *in silico* simulations correctly matched growth phenotypes assessed by *in vivo* experiments.

The quantitative and qualitative evaluations of the predictive capability of the model reconstructed herein falls within the range of those from most of the metabolic reconstructions available to date, supporting iMF721 as being a reliable reconstruction of the central metabolism of this bacterium.

3.2 Modelling the metabolic response to cold shock

In order to globally examine changes in PhTAC125s metabolism resulting from the temperature transition, we integrated two protein abundance datasets (obtained after growth at 4 and 18 C°, respectively) with constraints-based modelling of the iMF721 model. Up- and down-regulation ratios of protein expression were combined with the iMF721 metabolic model using MADE (Metabolic Adjustment by Differential Expression) [13]. Briefly, MADE creates a sequence of binary expression states that matches the most statistically significant changes in the series of gene expression measurements and, as such, it does not require an arbitrarily imposed gene expression threshold. The resulting gene states produce functioning models that simulate the real metabolic functional state of the cell, given the input expression values. Accordingly, this approach allows the identification of two distinct metabolic models (functional metabolic states), i.e. the original (optimal) iMF721 model and the one derived from simulating growth at lower temperature. These two models will differ in that some of their reactions will be (completely) turned on or off according to the measured levels of their corresponding proteins.After mapping information on up- and down-regulated genes onto the iMF721 model, the predicted (adjusted) growth rate decreased to 0.48 h-1; the 31% downshift in respect to the growth rate at 18 C°($0.69\ h^{-1}$) is compatible with experimental evidences on the reduced biomass production of PhTAC125 at lower temperature [12].

The overall scenario emerging from comparing the reactions fluxes from computational simulations (Fig. 2.) at the two temperatures suggests that PhTAC125 depresses its general metabolism following a switch between 18 and 4 C°, compatibly with i) the amino acids enriched nutritional environment of both in silico simulations and proteomics experiments, ii) the number of up- vs. down-regulated genes and iii) the reduced growth rate of this strain at 4 C°in respect to 18 C°. In this context, amino acids degradation and fatty acids metabolism seem to cover an important role. Amino acids, in particular, could be used as important carbon and energy sources.

3.3 Modelling the metabolic switches during the growth in a complex medium

The growth of PhTAC125 on complex medium (peptone) is characterized by several switches among the different C sources available in the growth medium (Fig. 3 A-E). These six steps can be summarized as follows: assimilation of P1) Ser, Asn, Glu, P2) Ser, Asn , Glu, Asp, Thr, P3) Glu, P4) Ala, Leu, Lys, Gly, Tyr, P5) Ala, Leu, Lys, Gly, Ile, Tyr, P6) Ala, Leu, Lys, Gly, Ile, Tyr, Phe, Val, P7) His. To explore the metabolic consequences at the whole cellular level of these relatively rapid changes in the preferred source, we implemented a (time resolved) modelling framework to account the flux distribution in each of the six phases. Predicted growth rates matched those derived experimentally (Fig. 3 F).

We then specifically analysed the changes in carried flux for each reaction, across the seven switches identified. Overall, we found that more than 200 reactions varied their carried flux across the whole time course (about 15% of the whole PhTAC125 metabolic network). These reactions were named ”switching reactions”, whereas those maintaining a fixed flux regardless of the used C source were named ”core” reactions. Calculating the Pearson correlation within the set of changing reactions, modules of apparently incompatible reactions were found (Fig. 4), parallel to the presence of co-occurring modules (i.e. reactions that are probably working in a concerted manner). The scenario emerging from this (non trivial) reactions associations points toward the

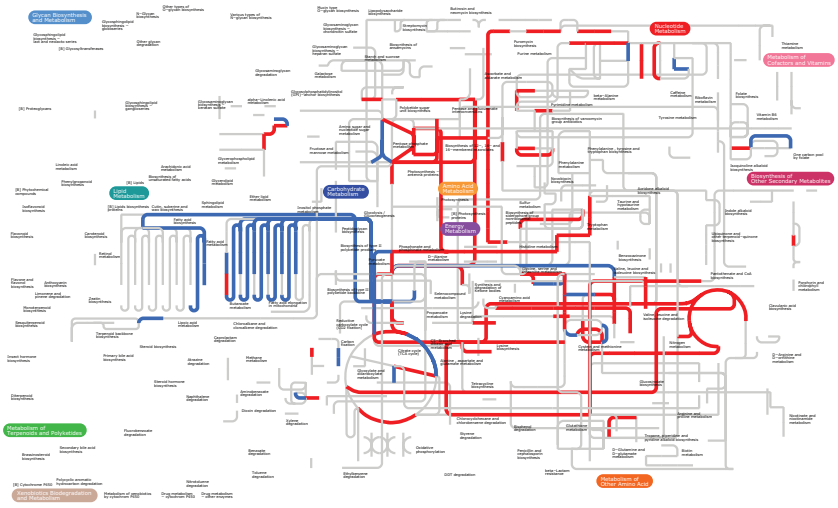


Figure 2: Schematic representation of changes in fluxes distribution after the shift from high 18 to 4C temperature. Red and blue lines indicate a decrease or an increase (of at least a factor 2) in reaction fluxes when shifting between the two conditions, respectively. Grey lines represent reactions for which a significant change in fluxes was not observed.

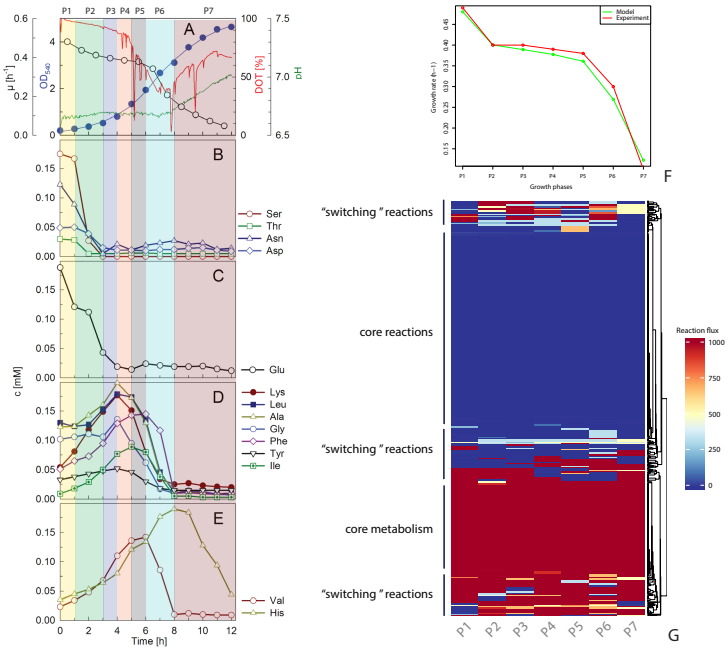


Figure 3: (A-E) The seven phases of PhTAC125 growth in complex medium (peptone). (F) The fitting of the model to the experimentally measured growth rates across the six phases. (G) Core and swithcing reactions during the seve gorwth phases.

presence of complex regulatory and metabolic interaction networks that allow a rapid adaptation of PhTAC125 in a nutritionally rapidly changing environment.

4 Conclusion

Here we have presented the genome-scale metabolic model reconstruction of the strain *P. haloplanktis* TAC125 (iMF721) [14]. To the best of our knowledge, this represents the first metabolic reconstruction of a bacterium isolated from Antarctica. Evidence synthesis from multiple data sources (including proteomics, phenomics, mod-

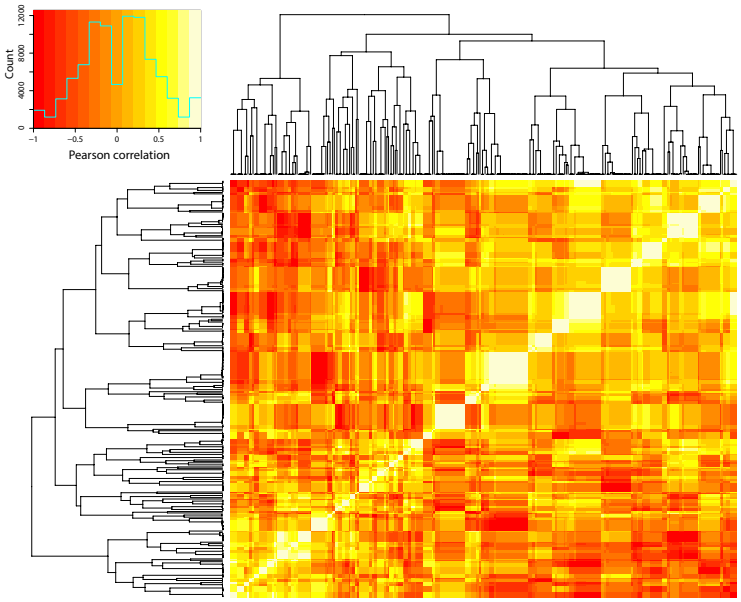


Figure 4: Correlation matrix bewteen all the identified "switching reaction identified during FBA modeling in complex medium."

elling and physiology) allowed the study of the metabolic response of PhTAC125 to temperature down shift and to nutrients switching in a complex growth medium at the system level. The biologically consistent predictions derived from our modelling framework represent a robust platform for future strategies aimed at the biotechnological exploitation of this strain.

Acknowledgments

The work described in this publication was financially supported by two PNRA (Piano Nazionale per la Ricerca in Antartide) grants (PNRA 2013/B4.02 and PNRA 2013/AZ1.04).

References

[1] Thiele, I., and Palsson, B.O. "A protocol for generating a high-quality genome-scale metabolic reconstruction." *Nat Protoc*, 5:93–121, 2010

[2] Varma, A., and Palsson, B.O. "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110". *Appl Environ Microbiol*, 60:3724–3731, 1994.

[3] Schilling, C.H., Edwards, J.S., Letscher, D., and Palsson, B.O. "Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems." *Biotechnol Bioeng*, 71:286–306, 2000

[4] Medigue, C., Krin, E., Pascal, G., Barbe, V., Bernsel, A., Bertin, P.N. et al. "Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis*". *Genome Res*,15:1325–1335, 2005

[5] Duilio, A., Tutino, M.L., and Marino, G. "Recombinant protein production in Antarctic Gram-negative bacteria.". *Methods Mol Biol*,267:225–237, 2004

[6] Wilmes, B., Kock, H., Glagla, S., Albrecht, D., Voigt, B., Markert, S. et al."Cytoplasmic and periplasmic proteomic signatures of exponentially growing cells of the psychrophilic bacterium *Pseudoalteromonas haloplanktis* TAC125.". *Appl Environ Microbiol*,77:1276–1283, 2011

[7] Rippa, V., Papa, R., Giuliani, M., Pezzella, C., Parrilli, E., Tutino, M.L. et al. " Regulated recombinant protein production in the Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125.". *Methods Mol Biol*,824:203–218, 2012.

[8] Corchero, J.L., Gasser, B., Resina, D., Smith, W., Parrilli, E., Vazquez, F. et al. " Unconventional microbial systems for the cost-efficient production of high-quality protein therapeutics.". *Biotechnol Adv*,31:140–153, 2013.

[9] Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T. et al. "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)." *Nucleic acids research*,42:D206–2143, 2014.

[10] Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P."iPath2.0: interactive pathway explorer." *Nucleic Acids Res* ,39:W412–415, 2011.

[11] "An effective cold inducible expression system developed in *Pseudoalteromonas haloplanktis* TAC125." *J Biotechnol*,127: 199-210, 2007.

[12] Piette, F., D'Amico, S., Struvay, C., Mazzucchelli, G., Renaut, J., Tutino, M.L. et al. "Proteomics of life at low temperatures: trigger factor is the primary chaperone in the Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125." *Mol Microbiol* ,76: 120-132, 2010.

[13] Jensen, P.A., and Papin, J.A. "Functional integration of a metabolic network model and expression data without arbitrary thresholding." *Bioinformatics* ,27: 541-547, 2011.

[14] Fondi, M, Maida, I., Perrin, E., Mellera, A., Mocali, S., Parrilli, E., Tutino, M.L., Li, P., Fani, R. "Genome-scale metabolic reconstruction and constraint-based modelling of the Antarctic bacterium *Pseudoalteromonas haloplanktis*TAC125." *Environmental Environ Microbiol.* ,17(3):751–766, 2014.

Comparative Analysis of Differentially Expressed Pathways in Mouse with ALS

Baarbatu Can(1), Arda Durmaz(1), Osman Uur Sezerman(1)

(1) Epigenetiks Genetik Biyoinformatik Yazlm A.
Teknopark stanbul, Sanayi Mh. Teknopark Blv. No: 1/4A Kuluka Merkezi, Pendik, 34906, Istanbul/Turkey

Keywords:

Abstract.

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease characterized by death of motor neurons [1]. Although several relations between specific genes and ALS have been identified, the underlying molecular mechanism is still waiting to be discovered. In this study, we used significantly expressed genes represented in [2] and carried out pathway analysis using PANOGA [3] to understand the difference between fast and slow progressing ALS. Significantly altered genes for each type of disease phenotype generated on mouse models are mapped to a protein-protein interaction network with 10174 genes and 61070 interactions [4]. Then active subnetworks, containing most of the affected genes, are identified. Then affected Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (http://www.genome.jp/kegg/pathway.html) in these active subnetworks are found and assigned a significance value using hyper geometic statistics. This procedure is repeated three times and the most significantly identified pathway is reported. In slow progressing ALS, Renin-angiotensin system is the most significantly affected pathway with a p value of 5.0e-8 and the relation of this pathway with motor neuron degradation and ALS is explained in [5]. In fast processing ALS, we found Pentose phosphate pathway is the most significantly altered pathway with a p value of 3.8e-9. Pentose phosphate pathway plays an important role in cell cycle and energy supply and operates together with Nrf2/ARE Pathway which is pointed out as an important biomarker for neurodegenerative diseases as explained in [6].

References

[1] Hardiman, Orla, Leonard H. van den Berg, and Matthew C. Kiernan. "Clinical diagnosis and management of amyotrophic lateral sclerosis." Nature Reviews Neurology 7.11 : 639-649, 2011.

[2] Nardo G, Iennaco R, Fusi N, Heath PR et al. Transcriptomic indices of fast and slow disease progression in two mouse models of amyotrophic lateral sclerosis. Brain Nov;136(Pt 11):3305-32. 2013

[3] Bakir-Gungor, B., Egemen, E., Sezerman, O. U. PANOGA: a web-server for identification of SNP targeted pathways from genome-wide association study data Bioinformatics, 2013

[4] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. The human disease network. Proceedings of the National Academy of Sciences of the United States of America 104: 86858690. 2007

[5] Kawajiri, M., et al. "Reduced angiotensin II levels in the cerebrospinal fluid of patients with amyotrophic lateral sclerosis." Acta neurologica Scandinavica 119.5 : 341-344, 2009.

[6] Calkins, Marcus J., et al. "The Nrf2/ARE Pathway as a Potential Therapeutic Target in Neurodegenerative Disease." ANTIOXIDANTS & REDOX SIGNALING 11.3 2009.

Psychrophilic protein modeling

Federica Chiappori[(1)]

(1) Istituto di tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Milano

Keywords:

Abstract. Tubulin dimers of psychrophilic eukaryotes can polymerize into microtubules at 4^oC, a temperature at which microtubules from mesophiles disassemble. This unique capability requires changes in the primary structure and/or in post-translational modifications of the tubulin subunits. To contribute to the understanding of mechanisms responsible for microtubule cold stability, here we present a computational structural analysis based on molecular dynamics (MD) and experimental data of three -tubulin isotypes, named EFBT2, EFBT3, and EFBT4, from the Antarctic protozoon Euplotes focardii that optimal temperature for growth and reproduction is 4^oC. The unique characteristics of the primary and tertiary structures of psychrophilic -tubulin isotypes seem responsible for the formation of microtubules with distinct dynamic and functional properties. Folding assistance is a fundamental requirement for tubulin. Here, we report a pilot folding analysis of a divergent beta-tubulin isotype, named EFBT3, from the Antarctic psychrophilic ciliate Euplotes focardii. To attain its native monomeric structure, beta-tubulin needs the assistance of the eukaryotic class II chaperonin CCT and cofactor A (CofA). We demonstrated that the rare Cys281 of EFBT3 is critical for the folding reaction. Model predictions indicate that EFBT3 binds to CofA differently from yeast beta-tubulin, suggesting a diverse folding mechanism that may be correlated with microtubule cold adaptation. Superoxide dismutases (SODs) are ubiquitous enzymes, which catalyse the disproportion of superoxide to molecular oxygen and peroxide. This reaction preempts the oxidizing chain reaction, preventing several cell damage caused by reactive oxygen species. E. focardii has two Cu/Zn SODs isoforms which display about 40% of sequence identity. In order to investigate the structural adaptation and the effect of ions at living low temperature and at mesophilic temperature, we performed a computational structural analysis based on molecular dynamics (MD) at 4C and 27C, and complexed or un-complexed with Cu and Zn ions, for comparing the ions effect on protein structure and the structure flexibility at different temperature.

Antimicrobial compounds from Antarctic bacteria

Pietro Tedesco[(1)], Fortunato Palma Esposito[(1)], Antonio Mondini[(1)], Glen Brodie[(2)], Renato Fani[(3)], Marcel Jaspars[(2)] and Donatella de Pascale[(1)]

(1) Institute of Protein Biochemistry, CNR, Naples, Italy

(2) University College of Aberdeen, The School of Natural and Computing Sciences, Aberdeen

(3) Department of Biology, University of Florence, Florence, Italy

Keywords:

Abstract. The increasing alarm of multidrug resistant bacteria in the last 20 years, led scientific community to the discovery of novel source of antimicrobials compounds. The bioprospecting from marine and extreme environments has yielded a noteworthy number of novel molecules from a wide range of organisms. Antarctica is the one of the most extraordinary places on Earth and exhibits many distinctive features. It is Earths southernmost continent and it is the coldest, driest, and windiest place on the planet. Thus, Antarctica hides organisms, which have evolved unique characteristics to face these harsh environmental conditions. In particular, Antarctic microorganisms are known to produce novel secondary metabolites that are valuable in a range of applications. Herein, we report on the development of a six-step biodiscovery pipeline starting with the collection of environmental samples and isolation of novel bacteria, to the chemical identification of the bio-assay guided purification of compounds with antimicrobial and antibiofilm activities. Antarctic sub-sea sediments were used to isolate more 1000 bacteria. The novel isolates were subjected to primary screening to determine their bioactivity against a selected panel of human pathogens (Staphylococcus aureus, Pseudomonas aeruginosa, Klebsiella pneumonia, Burkholderia cenocepacia). Isolates, positive to the first screening, were used to produce crude extracts from microbial exhausted culture broths.

A bioassay-driven purification was performed using crude extracts of the most promising isolates. LC-MS and NMR then structurally resolved the purified bioactive compounds.

12th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics

CNR Research Area, Naples, Italy
September 10-12, 2015

Conference Proceedings