Multi-omic data integration elucidates *Synechococcus* adaptation mechanisms to fluctuations in light intensity and salinity

Supreeta Vijayakumar and Claudio Angione

Department of Computer Science and Information Systems, Teesside University, UK

Abstract

Synechococcus sp. PCC 7002 is a fast-growing cyanobacterium which flourishes in freshwater and marine environments, owing to its ability to tolerate high light intensity and a wide range of salinities. Harnessing the properties of cyanobacteria and understanding their metabolic efficiency has become an imperative goal in recent years owing to their potential to serve as biocatalysts for the production of renewable biofuels. To improve characterisation of metabolic networks, genome-scale models of metabolism can be integrated with multi-omic data to provide a more accurate representation of metabolic capability and refine phenotypic predictions. In this work, a heuristic pipeline is constructed for analysing a genome-scale metabolic model of *Synechococcus* sp. PCC 7002, which utilises flux balance analysis across multiple layers to observe flux response between conditions across four key pathways. Across various conditions, the detection of significant patterns and mechanisms to cope with fluctuations in light intensity and salinity provides insights into the maintenance of metabolic efficiency.

Introduction

Metabolism is among the most important biological processes as balancing the production and consumption of metabolites is essential for maintaining life. Furthermore, it is currently the only biological layer that can be modelled genome-wide. Throughout the field of systems biology, there are a number of approaches which endeavour to capture the enormous complexity of biological systems by utilising mathematical modelling and computation to amalgamate the information required to build and refine predictive models of metabolism. The challenges presented by such an undertaking are numerous and persistent owing to the size, format, scale and variation of the disparate data types.

Constraint-based reconstruction and analysis (COBRA) methods are commonly used to express metabolic flux through biochemical pathways based on knowledge of reaction stoichiometry [1]. During flux balance analysis (FBA), a

pseudo-steady state is assumed to calculate all fluxes under time-invariance and spatial homogeneity for purposes of mass conservation. Mass-balance constraints are imposed on the system to identify a range of points representing all feasible flux distributions. A feasible phenotypic state in the solution space is then computed using linear programming with a set of values indicating the optimal conditions required to achieve a given objective function [2]. FBA is particularly suitable for modeling genome scale metabolic networks as the definition of kinetic parameters and metabolite concentrations is not a key requisite. To improve the characterisation of metabolic networks at the whole genome scale, genome-scale models of metabolism can be integrated with heterogeneous multi-omic data to provide a more accurate representation of metabolic capability. This is useful in refining phenotypic predictions across various environmental conditions.

In recent years, genome-scale metabolic models (GSMMs) have been integrated with multiple heterogeneous omic data types in a number of studies. This serves to exploit the large volume of experimental data being generated from high-throughput omics technologies, in order to improve the characterisation of metabolic networks at the whole genome scale. In doing so, additional constraints can be applied during flux balance analysis in order to shrink the solution space [3], thus providing a more accurate representation of metabolic capability as a greater number of factors can be considered to explain cellular behavior. This can prove useful in refining phenotypic predictions across various environmental conditions or engineering an organism in a way that optimises the production of a certain metabolite, which is highly applicable to fields such as industrial biotechnology and pharmacology.

Cyanobacteria are a group of photosynthetic prokaryotes for which it is imperative to adapt to constant fluctuations in temperature, salinity, light intensity (or irradiance), and nutrient availability, amongst other factors [4]. *Synechococcus* sp. PCC 7002 is a fast-growing cyanobacterium which flourishes in both freshwater and marine environments, owing to its ability to tolerate high light intensity and a wide range of salinities. Harnessing the properties of cyanobacteria has become an imperative goal in recent years owing to their potential to serve as biocatalysts for the production of renewable biofuels [5]. In an industrial setting, *Synechococcus* sp. PCC 7002 has been chosen as a model organism owing to its ease of genetic manipulation as well a tolerance for high salinity and slightly higher temperatures; these are highly desirable traits in micro-algae as this enables cultures to maintain a rapid growth rate in open raceway ponds as well as in photobioreactors, which operate at higher temperatures [6]. Recent studies have examined temporal variations in response to varying light intensity and associated conditional dependencies [7, 8]. These need to be accounted for as constraints in genome-scale metabolic models designed to simulate the phototrophic growth in cyanobacteria over diurnal cycles and tackle issues associated with resource allocation.

Methods

In this work, we present a heuristic pipeline for analysing a genome-scale metabolic model of the cyanobacterium *Synechococcus* sp. PCC 7002., which is detailed in Fig 1.
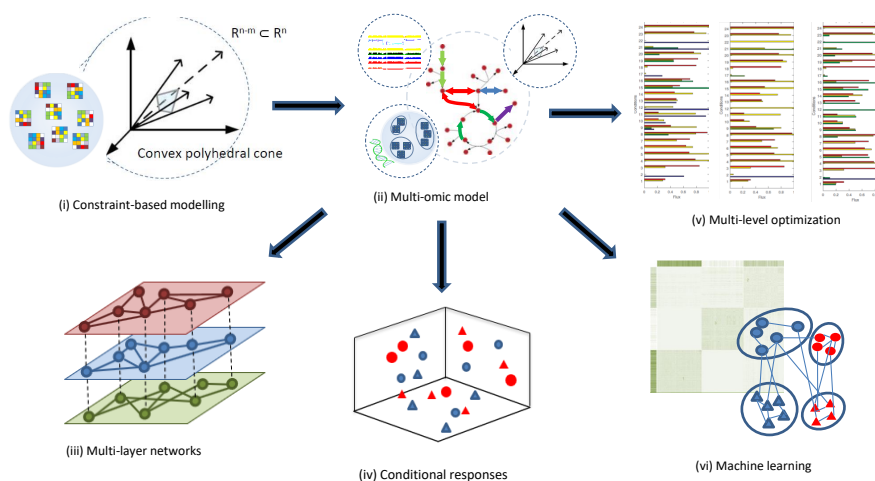

Fig. 1. Pipeline for prediction and classification of *Synechococcus* growth conditions. (i) Stoichiometric coefficients obtained from FBA used to map flux distribution as a convex polyhedral cone; (ii) Multi-omic model of textitSynechococcus sp. PCC 7002 produced by combining the genome-scale metabolic reconstruction with (iii) multi-layer networks of transcriptomic and fluxomic data; (iv) Phenotypic space depicting bacterial responses to varied conditions; (v) Utilisation of multiple objective functions by the model (vi) Unsupervised learning (e.g. PCA) can finally be used to detect latent patterns in unsupervised data by reducing dimensionality and identifying key contributions to variance in datasets.

We initiated our pipeline by mapping the flux distribution for phototrophic growth in *Synechococcus* sp. PCC 7002 using multi-omics flux balance analysis [9] and building condition-specific flux profiles using METRADE [10] and starting from a model recently published by Hendry et al [5].

Transcriptomic data was acquired in the form of RNA sequencing data from a series of studies previously conducted by Ludwig and Bryant [11, 4, 12]. Such data were compiled in an online repository known as Cyanomics (available at `http://lag.ihb.ac.cn/cyanomics/`)[13], an integrated omics analysis database containing omic data specific to *Synechococcus* sp. PCC 7002. The authors converted sequence data for various culture conditions to fastq format and used a Perl script to filter out low quality reads. A Python script had also been used to calculate the reads assigned per kilobase of target per million mapped reads

(RPKMs) as a measure of relative transcript abundance [13]. The reads had been mapped against the genome of *Synechococcus* sp. PCC using the Burrows-Wheeler algorithm [11]. We calculated fold change values to be centered around 1 by dividing the RPKM values under these conditions by the average expression of three standard control replicates for that gene.

These condition-specific expression profiles were loaded into the model using METRADE [10], for which FBA was carried out using the COBRA Toolbox in MATLAB. For the standard growth conditions, A+ medium was utilised as the culture medium [14] with a temperature of 38°C, continuous illumination at 250 μmol photons $m^{-2}s^{-1}$, and sparging with 1% (v/v) CO_2 in air; harvestation was performed when the optical density (OD) 730nm reached 0.7. Specific growth conditions which deviate from the standard conditions are recorded in Table 1, along with a reference to the original paper and the predicted biomass output.

Following this, three reactions involved in energy metabolism (ATP maintenance, photosystem I and photosystem II) were selected to serve as secondary objectives for a bi-level optimisation problem, which was formulated as follows:

$$
\begin{aligned}
&\max && g^\intercal v \\
&\text{such that} && \max f^\intercal v, \quad Sv = 0, \\
&v^{\min}\varphi(\Theta) \le v \le v^{\max}\varphi(\Theta),
\end{aligned}
\tag{1}
$$

where f is the primary objective function (biomass) and g is the secondary objective function. f and g are Boolean vectors of weights selecting the reactions in v whose flux rate will be considered as the objective. $v^{\min}$ and $v^{\max}$ are vectors which represent the lower and upper limits for the flux rates in v for the unconstrained model. Gene set expression of the reactions associated with the fluxes in v are represented by the vector Θ. φ is a function which maps the expression level of each gene set to a coefficient for the lower and upper limits of the corresponding reaction, and is defined as follows:

$$
\varphi(\Theta) = [\mathbf{1} + \gamma \, |log(\Theta)|]^{\mathrm{sgn}(\Theta-\mathbf{1})}.
\tag{2}
$$

The flux distributions calculated for four primary reactions (ATP, Photosystem II, Photosystem I and Biomass) under three pairs of objectives are detailed in Fig 2. In order to better visualise the differences in flux between conditions, flux values were normalised by dividing by the maximal flux for that reaction across all conditions.

Results and Discussion

The multi-omic data used in the multi-layer network consists of gene expression profiles in the transcriptomic layer and steady state flux distributions in the fluxomic layer. In Fig 1, the nodes in each layer represent environmental conditions such as light intensity and salinity whereas the dashed lines represent interactions between layers of data.

ID	Condition	Condition specifics	Ref.	Biomass
1	Dark oxic	Incubated in darkness prior to harvest, sparged in N_2	[11]	14.04
2	Dark anoxic	Incubated in darkness prior to harvest	[11]	0.00
3	High light	Illuminated at 900 μmol photons m^{-2} s^{-1} prior to harvest	[11]	37.25
4	OD 0.4	Harvested at OD 730nm = 0.4	[11]	34.54
5	OD 1.0	Harvested at OD 730nm = 1.0	[11]	30.35
6	OD 3.0	Harvested at OD 730nm = 3.0	[11]	28.18
7	OD 5.0	Harvested at OD 730nm = 3.0	[11]	30.73
8	Low O_2	Sparged in N_2	[11]	42.64
9	Low CO_2	Sparged with air [0.035% (v/v) CO_2]	[12]	17.85
10	N-limited	Cells washed in medium A (lacking NO_3-) and resuspended	[12]	9.59
11	S-limited	Cells washed with $MgCl_2$	[12]	9.53
12	PO_4-limited	Cells washed w/o (PO_4^{3-}), allowed to grow to OD = 0.7 until harvestation	[12]	0.00
13	Fe-limited	Cells washed in medium A with 720 μM deferoxamine mesylate B added at OD 0.35	[12]	21.65
14	NO_3-	Standard growth in medium A (lacking $NaNO_3$) with 25 mM HEPES, 1 μM NiSO4, 12 mM $NaNO_3$	[12]	20.36
15	NH_3	Standard growth in medium A (lacking $NaNO_3$) with 25 mM HEPES, 1 μM $NiSO_4$ and 10 mM NH_4Cl	[12]	23.91
16	$CO(NH_2)_2$	Standard growth in medium A (lacking $NaNO_3$) with 25 mM HEPES, 1 μM $NiSO_4$ and 10 mM $CO(NH_2)_2$	[12]	31.04
17	Heat Shock	1h heat shock at 47°C	[4]	0.00
18	22°C	Standard growth at 22°C	[4]	35.10
19	30°C	Standard growth at 30°C	[4]	36.26
20	Oxidative stress	5 μM methyl viologen added for 30 minutes prior to harvestation	[4]	33.55
21	Mixotrophic	Medium A+ supplemented with 10 mM glycerol	[4]	0.00
22	Low salt	Medium A+ containing 3 mM NaCl and 0.08 mM KCl	[4]	0.00
23	High salt	Medium A+ containing 1.5 M NaCl and 40 mM KCl	[4]	33.64

Table 1. Growth and stress conditions for *Synechococcus* sp. PCC 7002.
We map the flux distribution for phototrophic growth in *Synechococcus* sp. PCC 7002
using multi-omics flux balance analysis [9]. We then build condition-specific models
starting from a model recently published by Hendry et al (2016) [5], and predict the
biomass (growth rate, h^{-1}) in all experimental conditions. The unconstrained model
representing the optimal growth condition produced 44.04 h^{-1} of biomass.

From the transcriptomic studies listed, there are a number of genes which were not transcribed in the controls but specifically transcribed under perturbed conditions. Many of these genes have yet to be assigned a particular functional category or encode hypothetical proteins, but many more have been linked to specific pathways and compounds and some have been associated with the adaptation of *Synechococcus* sp. PCC 7002 to atypical environmental or growth conditions. The unconstrained model (simulating an optimal growth condition) produces 44.04 h^{-1} of biomass. Therefore, FBA correctly predicts reduced growth in sub-optimal conditions. Some growth is still maintained in some harsh conditions, although severely impaired or null under the dark anoxic, heat shock, phosphate-limited, mixotrophic and low salt conditions. More specifically, inducing a heat-shock response triggers transcriptional activation of heat shock proteins and alternative mechanisms to ensure survival when the main biomass production pathways are inhibited [15].

Although a large number of studies express the maximisation of biomass as the only objective when performing FBA, it is imperative to recognise that in reality most organisms have multiple objectives to satisfy. The addition of multi-level optimisation to our pipeline enables the consideration of more than one objective function and expands the phenotypic solution space so that there are a greater number of feasible optimal points. Specifically, when calculating the flux distribution across conditions, biomass was chosen as the primary objective and the secondary objective set to ATP maintenance, photosystem I and photosystem II. Biomass was chosen to represent the maximisation of growth rate and cellular yields [16], which is a critical consideration for the production of biofuels by cyanobacteria as this informs the substrate uptake rates and maintenance requirements indicating the fundamental growth requirements of the cell. The secondary objectives are key pathways involved in energy metabolism during photosynthesis. Simulating the cost of ATP maintenance can help to examine the energy required for sustaining metabolic activity even in the absence of growth. Incorporation of the photoexcitation reactions occurring within photosystems I and II can characterise how flux under various conditions reflects the light harvestation and energy transfer via photon absorption through these complexes. Thus, solving the linear programming problem between multiple pairs of objectives helps to resolve trade-offs by considering the conditions and constraints affecting each of these objectives.

Fig 2 predicts that the biomass flux through Conditions 1-23 are always lower than in Condition 24, the standard control. In Fig 2, we show that when ATP is set as the secondary objective, the highest fluxes through the ATP maintenance reaction are among conditions which limit growth such as dark anoxic (608.5), mixotrophic (1000), low salinity (1000) or nutrient limitations (ranging from 323.95 to 1000 for nitrogen, sulfur, phosphate and iron-limited conditions). It is likely that this is in order to maintain minimal cellular function when there is no growth (no flux through the biomass pathway) or energy transfer through the photosystems. Lack of light is likely to be a greater contributory factor to decrease in growth as low oxygen concentration does not seem to stunt growth,
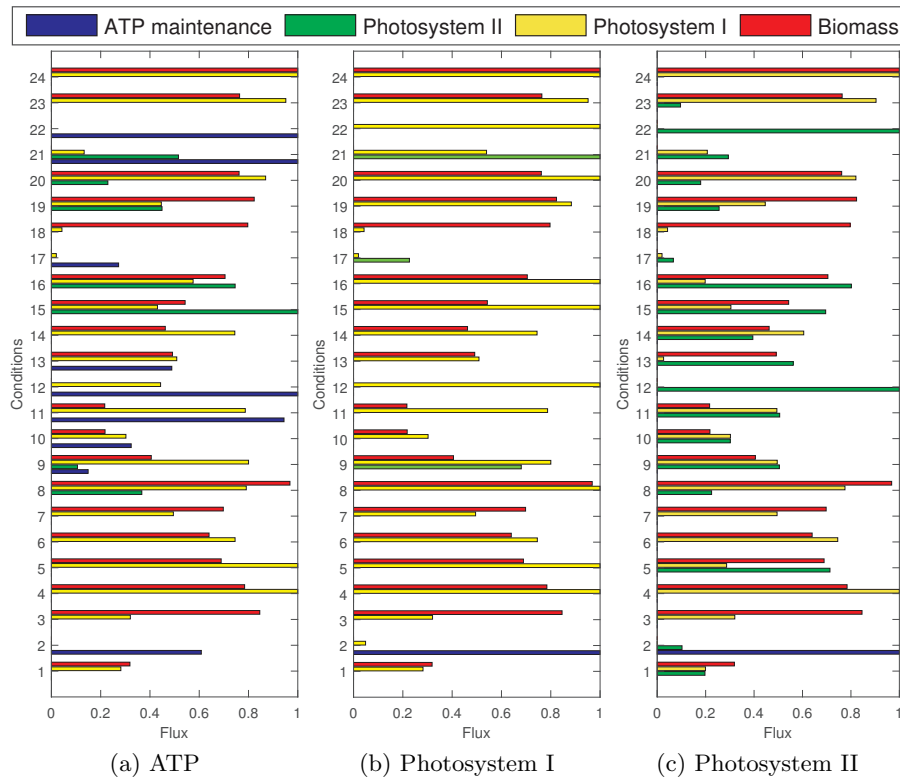
Fig. 2. This figure shows flux distributions for four key reactions: ATP maintenance, photosystem I, photosystem II and biomass when running FBA using three different pairs of objectives (where f = biomass, g = ATP/Photosystem I/Photosystem II). Conditions 1-23 correspond to those detailed in Table 1; Condition 24 is the standard control. Flux values were normalised by dividing by the maximal flux in each reaction across all conditions.

seeing as the proportional decrease in biomass is low relative to the standard control conditions. On the other hand, there appears to be little to no flux for either the biomass or photosynthetic pathways in the dark anoxic condition. This is supported by Vu et al (2013) [17], which states that lower yields under dark conditions may be due to the limited generation of energy (ATP) and reductant (NADPH) from glycogen in the absence of photoautotrophic growth. Equal reduction in transcript levels for photosynthetic apparatus was previously observed in all macronutrient-limited conditions studied [12]. We surmise that phosphate limitation has the greatest effect as there is no biomass production across any of our objectives for this condition. This is in line with the findings of Ludwig and Bryant (2012) [12], where perturbations caused by phosphate limitation had a greater impact on the global transcription pattern than observed for high irradiance or dark treatments of *Synechococcus* sp. PCC 7002.

Synechococcus sp. PCC 7002 is known to possess one of the greatest tolerances for high light intensity among cyanobacteria (with an upper limit of approximately 2000 μmol photons m^{-2} s^{-1}) [18]. This is evident from our predictions for all three pairs of objectives, where flux through the biomass pathway during high light intensity is only slightly lower than the control condition (37.25 compared to 44.04), even though we observed flux under light intensity of 900 μmol photons m^{-2} s^{-1}. Although fluxes through photosystem II are disrupted, fluxes through photosystem I are still maintained (80.11 in the high light intensity condition as opposed to 250 in the control). Heat shock results in weak fluxes through both photosystems for all three objective pairs and triggers flux through the ATP maintenance reaction when it is set as an objective, presumably to retain cellular function. It was reported by Ludwig and Bryant (2012)[4] that transcript levels for genes encoding photosystem I decreased slightly in cells grown at high salinity and remained constant at low salinity; on the other hand, it was found that transcript levels for genes encoding photosystem II did not change in response to fluctuations in salinity. For the high salinity condition, we have observed that fluxes through biomass and photosystem I remain high for our three objective pairs, whereas flux is only maintained in the low salinity condition for the reaction set as the secondary objective g. However, flux through photosystem II for the low salinity condition (250.00) is much higher than flux for the high salinity condition (24.06) when the reaction for photosystem II is set as the secondary objective, and there is no flux through the photosystem II reaction in the control condition.

Conclusions

Complex metabolic and phenotypic outcomes as a result of adaptation to a changing environment are difficult to predict from gene expression only. The unified measure of bacterial responses computed by the condition-specific models allows for the detection of coordinated responses shared between different data types as well as the variation in responses across differing growth conditions.

In this work, a heuristic pipeline was constructed for analysing a genome-scale metabolic model of the cyanobacterium *Synechococcus* sp. PCC 7002, which utilises flux balance analysis to obtain flux distributions with multi-level optimisation using linear programming. The four reaction fluxes obtained show clear differences in pathway activity across the various conditions and also between the three pairs of objectives used. This enables the detection of latent, biologically significant patterns and adaptive mechanisms to fluctuations in light intensity and salinity. The aim is to elucidate how *Synechococcus* maintains metabolic efficiency at the cellular level whilst assessing multiple cellular objectives. Considering the vast dimensionality of multi-omic models, the identification of biologically meaningful information can prove to be challenging. As a non-parametric statistical technique, principal components analysis (PCA) can be incorporated into our workflow for identifying patterns in metabolic fluxes within multi-omic models [19].

In this regard, this study integrates transcriptomics with metabolomics and elucidates the unique mechanisms utilised by *Synechococcus* sp. PCC 7002 to adapt to changes in light intensity and salinity to maintain metabolic efficiency for phototrophic growth and light-dependent photosynthesis in a multi-omic fashion. As a result, by predicting and classifying its metabolic profiles in such growth conditions, our approach sheds light on the adaptation process undergone by the cyanobacterium to enable its survival across a wide range of environments and stress conditions.

References

1. Daniel Hyduke, Jan Schellenberger, Richard Que, Ronan Fleming, Ines Thiele, Jeffery Orth, Adam Feist, Daniel Zielinski, Aarash Bordbar, Nathan Lewis, et al. Cobra toolbox 2.0. *Protocol Exchange*, pages 1–35, 2011.
2. Ali Ebrahim, Elizabeth Brunk, Justin Tan, Edward J O'brien, Donghyuk Kim, Richard Szubin, Joshua A Lerman, Anna Lechner, Anand Sastry, Aarash Bordbar, et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nature Communications*, 7, 2016.
3. Jennifer L Reed. Shrinking the metabolic solution space using experimental datasets. *PLoS Comput Biol*, 8(8):e1002662, 2012.
4. Marcus Ludwig and Donald A Bryant. Synechococcus sp. strain pcc 7002 transcriptome: acclimation to temperature, salinity, oxidative stress, and mixotrophic growth conditions. *Frontiers in microbiology*, 3:354, 2012.
5. John I Hendry, Charulata B Prasannan, Aditi Joshi, Santanu Dasgupta, and Pramod P Wangikar. Metabolic model of synechococcus sp. pcc 7002: Prediction of flux distribution and network modification for enhanced biofuel production. *Bioresource Technology*, 213:190–197, 2016.
6. Anne M Ruffing, Travis J Jensen, and Lucas M Strickland. Genetic tools for advancement of synechococcus sp. pcc 7002 as a cyanobacterial chassis. *Microbial Cell Factories*, 15(1):190, 2016.
7. Marco Rügen, Alexander Bockmayr, and Ralf Steuer. Elucidating temporal resource allocation and diurnal dynamics in phototrophic metabolism using conditional fba. *Scientific Reports*, 5, 2015.

8. Alexandra-M Reimers, Henning Knoop, Alexander Bockmayr, and Ralf Steuer. Evaluating the stoichiometric and energetic constraints of cyanobacterial diurnal growth. *arXiv preprint arXiv:1610.06859*, 2016.

9. Claudio Angione, Max Conway, and Pietro Lió. Multiplex methods provide effective integration of multi-omic data in genome-scale models. *BMC bioinformatics*, 17(4):257, 2016.

10. Claudio Angione and Pietro Lió. Predictive analytics of environmental adaptability in multi-omic network models. *Scientific reports*, 5, 2015.

11. Marcus Ludwig and Donald A Bryant. Transcription profiling of the model cyanobacterium synechococcus sp. strain pcc 7002 by next-gen (solid) sequencing of cdna. *Frontiers in microbiology*, 2:41, 2011.

12. Marcus Ludwig and Donald A Bryant. Acclimation of the global transcriptome of the cyanobacterium synechococcus sp. strain pcc 7002 to nutrient limitations and different nitrogen sources. *Frontiers in microbiology*, 3:145, 2012.

13. Yaohua Yang, Jie Feng, Tao Li, Feng Ge, and Jindong Zhao. Cyanomics: an integrated database of omics for the model cyanobacterium synechococcus sp. pcc 7002. *Database*, 2015:bau127, 2015.

14. S Edward Stevens and Ronald D Porter. Transformation in agmenellum quadruplicatum. *Proceedings of the National Academy of Sciences*, 77(10):6052–6056, 1980.

15. Hema Rajaram, Akhilesh Kumar Chaurasia, and Shree Kumar Apte. Cyanobacterial heat-shock response: role and regulation of molecular chaperones. *Microbiology*, 160(4):647–658, 2014.

16. Adam M Feist and Bernhard O Palsson. The biomass objective function. *Current opinion in microbiology*, 13(3):344–349, 2010.

17. Trang T Vu, Eric A Hill, Leo A Kucek, Allan E Konopka, Alexander S Beliaev, and Jennifer L Reed. Computational evaluation of synechococcus sp. pcc 7002 metabolism for chemical production. *Biotechnology journal*, 8(5):619–630, 2013.

18. Qian Xiong, Jie Feng, Si-ting Li, Gui-ying Zhang, Zhi-xian Qiao, Zhuo Chen, Ying Wu, Yan Lin, Tao Li, Feng Ge, et al. Integrated transcriptomic and proteomic analysis of the global response of synechococcus to high light stress. *Molecular & Cellular Proteomics*, 14(4):1038–1053, 2015.

19. Elizabeth Brunk, Kevin W George, Jorge Alonso-Gutierrez, Mitchell Thompson, Edward Baidoo, George Wang, Christopher J Petzold, Douglas McCloskey, Jonathan Monk, Laurence Yang, et al. Characterizing strain variation in engineered e. coli using a multi-omics-based workflow. *Cell systems*, 2(5):335–346, 2016.