Computational Aspects of Metabolic Processes:
Modeling, Analysis and Applications

Edward Vitkin
Computational Aspects of Metabolic Processes: Modeling, Analysis and Applications

Research Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Edward Vitkin

Submitted to the Senate of the Technion – Israel Institute of Technology

Kislev 5779 Haifa November 2018
To my beloved wife, Anastasia

I use this opportunity to thank everyone, who has helped me to successfully complete this long journey.

I want to thank my esteemed research adviser, Dr. Zohar Yakhini, for your enormous support and guidance. You’ve taught me a lot during these years — how to do research and how to present it, how to approach new problems and how to prioritize tasks. I am lucky and proud to be among your students and research team members. THANK YOU!

I want to thank my IBM managers Alex Melament and Yardena Peres for your support and trust. You’ve taught me how to address the research problems from the practical perspectives, how to close the gaps between theory and the practice and how to be a better team player. THANK YOU!

I want to thank my very good friend and coauthor Dr. Alex Golberg for your ideas and energy. Also, I’m grateful to my MSc adviser Dr. Tomer Shlomi and my colleagues from Technion, IBM and other places — Oe Solomon, Israel Steinfeld, Limor Leibovich, Shay Ben Shlazar, Leon Anavi, Igor Kostiner-Kronos, Yevgenia Tsimerman, Yael Mandel-Gutfreund, Amir Ben-Dor, Dan Geiger and many many others for our fruitful discussions, your invaluable help and simply for being important and pleasant contributors to this journey. THANK YOU ALL!

I want to thank my entire family for your believe and support. During my studies you’ve never shown any regret or doubt and always knew I would succeed. My grandparents — Yuri and Berta, Riva and Naum, my parents — Michael and Sophia, my mother in law — Lubov, my aunts, uncles, relatives and, of course, my many friends, who are also a family — I love you all. THANK YOU!

I am especially grateful to my beloved wife Anastasia and to my children Benjamin and Alma for being with me in general and especially during this journey. I am very lucky to have you on my side and I love you very much.
The research thesis was done under the supervision

of Dr. Zohar Yakhini

Computer Science Department,

Technion – Israel Institute of Technology

The generous financial help of the Technion – Israel Institute of Technology
is gratefully acknowledged
Abstract

This PhD dissertation addresses various aspects of simulating the metabolic behavior of living organisms. The main specific topics include modeling and optimizing bio-industrial fermentation processes, the refinement of metabolic network models leveraging high-throughput knockout data and the normalization and visualization of data from clinical metabolite measurements.

We start with presenting BioLego, which is a freely downloadable webservice, ready for installation in the Microsoft Azure Cloud environment. BioLego provides a friendly and intuitive interface that enables the simulation (modeling and evaluation of the expected performance) and the optimization of single and two-step fermentation processes. The BioLego fermentation simulator is a scalable distributed framework, providing means to the process designers for analyzing, predicting and comparing the expected efficiency of several fermentation scenarios of interest. It is based on a novel flexible modular modelling approach, enabling smooth generation of different multi-organism fermentation configurations consisting of independent encapsulated modules, representing individual organisms.

We continue with drill-down to genome-scale single organism cellular metabolic models. Gene-fitness assays, measuring organism behavior in different growth conditions after genetic modifications, lead to newly generated data that are relevant for metabolic modelling. We developed a methodology to leverage these data to improve gene-to-reaction associations in existing genome-scale models. Specifically, we use these data to assign genes to orphan reactions. We also show how to integrate these data with statistics gathered from other types of functional-genomic data, such as gene-expression profiles, to facilitate gene-to-reaction assignments.

Finally, we present an approach for measurement-based analysis of human metabolism using the example of the steroidogenesis pathway. We describe a novel method for normalizing highly biased (age, sex) data, such as metabolomic hormonal profiles of
children. We also describe a novel method for visualizing the analysis results both for patient-based and population-based perspectives.

The major contributions of this PhD research are as follows:

- **BioLego project, including**
  - Novel modular modelling approach, enabling smooth generation of different multi-organism fermentation configurations
  - Distributed framework based on Microsoft Azure Cloud environment, enabling high-scale simulations (modeling and evaluation of the expected performance) and optimization of single and two-step fermentation processes, including multiple knockouts in each organism.
  - Intuitive interface, requiring no programming skills from potential users (fermentation process designers).
  - Biological predictions for various fermentation scenarios supporting several published studies.

- **Gene-to-reaction assignment methodology for fitness assays, including:**
  - A novel methodology to leverage fitness assays for genes and promoters to improve existing gene-to-reaction assignment techniques
  - An integration of fitness assays statistics with other statistics based on gene-expression profiles
  - Biological predictions of candidate genes potentially encoding proteins that catalyze 107 orphan reactions in iJO1366 metabolic model of *E. coli*.

- **Analysis of clinical metabolomics profiles, including:**
  - A systems approach for analyzing steroid metabolomics data to identify a steroid-related disease signatures and patterns
  - A novel peer-group normalization approach, clearly overperforming current methods in the field
  - A novel visualization approach, addressing both patient-based and population-based studies
  - Analysis of data from Congenital Adrenal Hyperplasia due to 21-hydroxylase deficiency (CAH) and from childhood obesity subjects.
List of Publications

1. Vitkin, E., Gillis, A., Polikovsky, M., Bender, B., Golberg, A., Yakhini, Z. “Distributed flux balance analysis simulations of serial two-organism biomass fermentation”. *(Under preparation for Bioinformatics)*

*Also published as a selected paper with significant impact in Diabetes and Obesity section of World Biomedical Frontiers. (2015).*

Abstracts of these manuscripts can be found in Appendix A: Abstracts of published manuscripts

List of Conference Talks

2. Vitkin, E. “*High-throughput research in steroid metabolomics*” AnalytiX-2018. *(Invited Conference Talk)*
Table of Contents

Abstract ........................................................................................................................................... 1
List of Publications .......................................................................................................................... 3
List of Conference Talks ................................................................................................................... 3
Table of Contents ............................................................................................................................ 5
List of Equations .............................................................................................................................. 9
List of Figures .................................................................................................................................. 10
List of Tables .................................................................................................................................... 12
Abbreviations and notations ........................................................................................................... 14
Preface .............................................................................................................................................. 15

Chapter 1. Modeling fermentation processes for biorefinery design – BioLego .............. 17
  Section 1.1. Background ............................................................................................................. 19
  Section 1.2. General Methods .................................................................................................... 23
    1.2.1. Metabolic model of the cell ............................................................................................ 23
    1.2.2. Flux Balance Analysis .................................................................................................. 24
    1.2.3. Construction of genome-scale metabolic models .......................................................... 30
  Section 1.3. Methods developed for BioLego and related projects ........................................ 33
    1.3.1. Encapsulated microorganism metabolic models – intuition ........................................ 33
    1.3.2. The modular approach – principles of constructing single module .............................. 33
    1.3.3. The modular approach – mathematical formulation of single module ...................... 35
    1.3.4. The modular approach – principles of combining modules ......................................... 37
    1.3.5. The modular approach – mathematical formulation of two-organism fermentation process .......................................................................................................................... 40
    1.3.6. Analysis of feedstock media by product gradient and media sensitivity .................... 49
    1.3.7. Addressing changes in growth conditions. Adjustment of oxygen .............................. 50
    1.3.8. Fermentation pipeline optimization by reaction knock-outs ....................................... 50
    1.3.9. Scalable Architecture ...................................................................................................... 52
  Section 1.4. Experiments, Applications and Tools ................................................................. 55
1.4.1. Proposed design of distributed macroalgal biorefineries: Thermodynamics, bioconversion technology, and sustainability implications for developing economies. 55

1.4.2. BioLEGO — a web-based application for biorefinery design and evaluation of serial biomass fermentation ................................................................. 60

1.4.3. Thermochemical hydrolysis of green macroalgae Ulva feedstock for biorefinery: Taguchi orthogonal arrays experimental design approach for physical and chemical parameters optimization ........................................... 63

1.4.4. Macroalgae Biorefinery from Kappaphycus alvarezii: Conversion Modeling and Performance Prediction for India and Philippines as Examples .......... 68

1.4.5. Distributed flux balance analysis simulations of serial two-organism biomass fermentation ....................................................................................... 73

Chapter 2. Improving genome-scale metabolic models with fitness data ............ 81

Section 2.1. Background .................................................................................. 83

Section 2.2. Methods ...................................................................................... 85

2.2.1. Obtaining the fitness data ................................................................. 85

2.2.2. Fitness vectors for non-orphan reactions ........................................... 86

2.2.3. Fitness vectors for genes and promoters ............................................. 87

2.2.4. Reaction neighborhood ..................................................................... 87

2.2.5. Fitness vectors for orphan reactions. Gene–to-Reaction assignment based on reaction neighborhoods ................................................................. 89

2.2.6. Combination of gene and promoter fitness ........................................ 90

2.2.7. Uniform gene-to-reaction assignment null models............................. 91

2.2.8. Integration of fitness- and expression-based scores ............................ 91

2.2.9. Confidence of gene to reaction assignment ........................................ 93

Section 2.3. Results and Findings ................................................................... 94

2.3.1. Performance evaluation based on non-orphan reaction data ............... 94

2.3.2. Predicting genes for orphan reactions ................................................. 97

Chapter 3. Analysis of steroid metabolism ...................................................... 99

Section 3.1. Background ............................................................................... 99

Section 3.2. Methods and Subjects ............................................................... 101
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1.</td>
<td>Subjects</td>
<td>101</td>
</tr>
<tr>
<td>3.2.2.</td>
<td>GC-MS Data</td>
<td>101</td>
</tr>
<tr>
<td>3.2.3.</td>
<td>Peer-Group Normalization</td>
<td>101</td>
</tr>
<tr>
<td>3.2.4.</td>
<td>Analysis of Metabolites and Reactions</td>
<td>103</td>
</tr>
<tr>
<td>3.2.5.</td>
<td>Urine-to-Blood (U2B) Visualization</td>
<td>104</td>
</tr>
<tr>
<td>Section 3.3.</td>
<td>Results and Findings</td>
<td>105</td>
</tr>
<tr>
<td>3.3.1.</td>
<td>Evaluation of peer-group normalization</td>
<td>105</td>
</tr>
<tr>
<td>3.3.2.</td>
<td>Data normalization analysis and visualization in CAH patients</td>
<td>106</td>
</tr>
<tr>
<td>3.3.3.</td>
<td>Data normalization analysis and visualization in childhood Obesity</td>
<td>108</td>
</tr>
<tr>
<td>Chapter 4.</td>
<td>Discussion and Future directions</td>
<td>111</td>
</tr>
<tr>
<td>4.1.</td>
<td>Extending BioLego project</td>
<td>111</td>
</tr>
<tr>
<td>4.2.</td>
<td>Leveraging fitness assays for metabolic modeling</td>
<td>113</td>
</tr>
<tr>
<td>4.3.</td>
<td>Steroidogenesis analysis</td>
<td>113</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>115</td>
</tr>
<tr>
<td>Appendix A.</td>
<td>Abstracts of published manuscripts</td>
<td>121</td>
</tr>
</tbody>
</table>

Vitkin, E., Gillis, A., Polikovsky, M., Bender, B., Golberg, A., Yakhini, Z. “Distributed flux balance analysis simulations of serial two-organism biomass fermentation”. (Under preparation for Bioinformatics) ........................................ 121


List of Equations

Equation 1-1: Gene-to-Reaction Boolean mapping .................................................... 24
Equation 1-2: Basic CBM constraints .................................................................. 25
Equation 1-3: FBA framework ........................................................................... 26
Equation 1-4: FVA framework ........................................................................... 26
Equation 1-5: FBA framework with reaction knockouts ................................... 27
Equation 1-6: (A) OptKnock and (B) RobustKnock frameworks for knockout optimization ................................................................. 29
Equation 1-7: FVA framework for single module ............................................. 37
Equation 1-8: General FVA framework for serial two-step combination of two modules ........................................................................................................ 42
Equation 1-9: Transformation of bi-directional reaction $v_r$ ............................. 42
Equation 1-10: Constraint c3 from Equation 1-8 in its primal (a) and dual forms (b) .... 43
Equation 1-11: Constraint c4 from Equation 1-8 reformulated based on Equation 1-10. 43
Equation 1-12: Constraint c4 from Equation 1-8 in its primal (a) and dual forms (b) .... 43
Equation 1-14: Relaxed FVA framework for serial two-step combination of two modules ............................................................................................................ 45
Equation 1-15: FVA framework for single-step co-culture combination of two modules 47
Equation 1-16: Media sensitivity analysis .......................................................... 49
Equation 1-17: Media gradient analysis ............................................................. 49
Equation 1-18: FVA framework with reaction activity vectors for a serial two-step combination of two organisms ................................................................. 51
Equation 1-19: Wasted Energy ........................................................................... 58
Equation 1-20: Annual bioethanol production ..................................................... 70
Equation 2-1: Fitness score for non-coding region ............................................. 85
Equation 2-2: Fitness score for non-orphan reaction ......................................... 86
Equation 2-3: Z-normalized fitness score for non-orphan reaction .................. 86
Equation 2-4: Association Likelihood Score for a given Reaction-Gene pair .......... 90
Equation 2-5: Association Likelihood Score for a given Reaction-Promoter pair .... 90
Equation 2-6: Association Likelihood Score for a Reaction-(Gene&Promoters) pair .... 90
Equation 2-7: Combined Association Likelihood Score for a given Reaction-Gene pair 92
Equation 2-8: (A) Bayesian and (B) Unbiased confidence of $ALS(reaction, gene)$ .... 93
Equation 3-1: Distance in metadata space between two subjects ....................... 102
Equation 3-2: Threshold distance estimation ...................................................... 102
List of Figures

Figure 1-1: Coupling maximal growth with the production of target molecule ........ 28

Figure 1-2: Single-module for existing organism metabolic model. Each contains five chambers: Internal – consisting of the known metabolic model of the organism; Media – representing received biomass feedstock media; Product – representing desired target metabolite; Growth – consisting organism cellular components; Waste – representing all non-digested media residuals and molecules by organism as growth by-product ........ 34

Figure 1-3: Construction of the matrix $S_{Env}$. I stands for unitary matrix of appropriate dimension (varies from organism to organism). All empty cells contain 0-matrix of appropriate dimensions, which is not shown for visual compactness. The emphasized left side ($S_{Env-INISIDE}$) of the figure is used later for construction of multi-modular processes (see 1.3.5) ................................................................. 35

Figure 1-4: Combination of two single-organism modules to characterize two-step serial fermentation process. In addition to the modules of participating organisms, we added four auxiliary modules to simulate real industrial environment: Feedstock Media, Control Layer, Total Waste and Total Product (see 1.3.5) ................................................................. 37

Figure 1-5. Combination of two single-organism modules to characterize single-step fermentation process, designed as co-culturing of two organisms. Three external boundary modules are added to the process in addition to the modules of participating organisms: Feedstock Media, Total Waste and Total Product. ....................................................... 39

Figure 1-6: Construction of the matrix $S_{System}$ for serial two-step fermentation. I stands for unitary matrix of appropriate dimension (varies from organism to organism). All empty cells are equal to 0-matrix of appropriate dimension, which is not shown for visual compactness. For the same reason we omit all intra-module reactions described in Figure 1-3 ................................................................. 40

Figure 1-7: Construction of the matrix $S_{System}$ for single-step fermentation by two-organism co-culture. I stands for unitary matrix of appropriate dimension (varies from organism to organism). All empty cells are equal to 0-matrix of appropriate dimension, which is not shown for visual compactness. For the same reason we omit all intra-module reactions described in Figure 1-3 ................................................................. 48

Figure 1-8: Architecture of MS Azure Cloud-based BioLego platform ....................... 54

Figure 1-9: Schematic representation of the serial two-step bioconversion of Ulva feedstock into bioethanol. Based on Model 3, Table 1-3 ................................................................. 58
Figure 1-10: Sensitivity analysis of fibers and protein content variation, expected with nutrients availability and seasonal changes, impact on (a). *S. cerevisiae* and *E.coli* growth rates (b). Ethanol yield and (c) Carbon Utilization yield based on the two-step fermentation scenario with *S. cerevisiae* as first organism................................. 72

Figure 1-11: Heatmap of the expected minimal and maximal ethanol production yields per knocked pair of reactions, in [%] of WT. Here x-axis corresponds to number of \( \text{minProd}(\text{act}) \) in Equation 1-12 and y-axis to number of \( \text{maxProd}(\text{act}) \) giving the observed WT percentage................................................................. 76

Figure 1-12: Experimental validation of BioLego predictions .................................. 78

Figure 2-1: The general flow for using fitness assays integrated together with additional data sources to improve the assignment of genes to orphan reactions ............ 81

Figure 2-2: Greedy removal of high-frequency metabolites...................................... 88

Figure 2-3: Number of reactions per metabolite in the iJO1366 model of *E. coli*. ..... 89

Figure 2-4: iJO1366 model genes covered by fitness and expression datasets.......... 92

Figure 2-5: Distribution of gene-to-reaction ALS scores ........................................ 93

Figure 2-6: Comparative accuracy of Association Likelihood Score values for first gene prediction in a leave-one-out validation on the 1556 adequate non-orphan reactions in iJO1366 metabolic model of *E. coli* ................................................................. 94

Figure 2-7: Comparative accuracy of Association Likelihood Score values for the second gene prediction in a leave-one-out validation on the 565 non-orphan reactions in iJO1366 metabolic model of *E. coli* with at least 2 encoding genes. ....................... 96

Figure 2-8: Performance as a function of number of measured fitness conditions .. 97

Figure 3-1: 2D example for the peer-group selection. Each “x” mark denotes control subject in AGE by BMI 2D-space; “+” marks the peer-subject. The ellipse demonstrates the peer-group chosen for the peer-subject. Since age has a higher weight than BMI, peer group membership is more sensitive to age than BMI ................................................. 103

Figure 3-2: (A) TNoM and (B) Student-TTest comparison between CAH-affected and control children................................................................. 106

Figure 3-3: (A) TNoM and (B) Student-TTest comparison between obese and control children................................................................. 109
List of Tables

Table 1-1: Inter-module interaction rules for two-step serial fermentation process........ 38

Table 1-2: Module interaction rules for single-step fermentation process, designed as co-culturing of two organisms ............................................................. 39

Table 1-3: Fermentation scenarios for production of ethanol based on U. lactuca. Column *MAX YIELD refers to full ulvan decomposition and the ability of S. cerevisiae to uptake xylose ................................................................. 59

Table 1-4: BioLego efficiency simulation results for corn cobs to ethanol single-step fermentation with different S. cerevisiae strains. RN1016 strain can digest xylose....... 60

Table 1-5: Efficiency of fermentation process for 2:1 mix of U. lactuca together with K. alvarezii biomasses by E. coli and WT S. cerevisiae. ................................................. 62

Table 1-6: Taguchi L16 orthogonal array for determination of optimum process parameters for thermochemical deconstruction of macroalgae Ulva biomass. ............ 63

Table 1-7: Carbohydrate components of U. lactuca averaged over 2 repetitions of the Taguchi L16 orthogonal array as defined in Table 1-6............................................. 64

Table 1-8: Predicted ethanol yields using Ulva hydrolysate fermentation by E. coli...... 65

Table 1-9: Predicted ethanol yields using Ulva hydrolysate fermentation by S. cerevisiae WT................................................................................................................. 65

Table 1-10: Predicted ethanol yields using Ulva hydrolysate fermentation by S. cerevisiae RN1016 ......................................................................................................... 66

Table 1-11: Predicted ABE (acetone, butanol, ethanol) yields using Ulva hydrolysates fermentation by C. acetobutylicum................................................................. 67

Table 1-12: K. alvarezii chemical composition used in the model as fermentation medium, based on (Fayaz et al., 2005; Rajasulochana et al., 2010; Lechat et al., 1997). 69

Table 1-13: Simulation results of fermentation of K. alvarezii biomass ...................... 70

Table 1-14: Sensitivity analysis of fermentation of K. alvarezii biomass in a serial two-step fermentation scenario with S. cerevisiae followed by E.coli ...................... 71
Table 1-15: Runtime statistics for different computational experiments with distributed BioLego

Table 2-1: Comparative accuracy of scores predicting first reaction gene assignments for 1556 adequate non-orphan reactions in iJO1366 metabolic model of E. coli

Table 2-2: Comparative accuracy of scores predicting second reaction gene assignments for 565 adequate non-orphan reactions in iJO1366 metabolic model of E. coli with at least 2 encoding genes.

Table 3-1: Counts of highly-correlated measurement inferred variables for each metadata dimension in the control population. These counts represent number of variables with correlation providing p-value <0.01 out of total 56 (31 urinal metabolites, 12 blood metabolites and 13 reaction activities).

Table 3-2: Most significant differences between control and CAH patients for the estimated values after normalization for gender, age, BMI, weight and height. Results are sorted according to TNoM p-values. P-value correction was done by Bonferroni correction.

Table 3-3: Most significant differences between control and obese subjects after normalization for gender and age. Results are sorted according to TNoM p-values. P-value correction was done by Bonferroni correction.
Abbreviations and notations

- CBM – Constraint Based Modeling
- FBA – Flux Balance Analysis
- FVA – Flux Variability Analysis
- MIRAGE – MetabolIc Reconstruction via functionAl Genomics
- CUY – Carbon Utilization Yield
- BPP – Bioethanol Production Potential
- CAH – Congenital Adrenal Hyperplasia due to 21-hydroxylase deficiency
- GC-MS – Gas Chromatography-Mass Spectrometry
- BMI – Body Mass Index
- U2B – Urine-to-Blood
- UI – User Interface
Preface

Organism metabolism is one of the best-studied biological systems. This is true both on the organism and on the cellular levels. Mathematical models describing various metabolic behaviors become more and more precise. There are many real-life applications that can benefit from leveraging such models. Unfortunately, the actual number of applications doing so is significantly lower than expected. One probable reason for this disparity is insufficient precision of the existing models. Another reason is a gap in understanding how models of metabolism can be leveraged for assistance in different relevant contexts.

This PhD dissertation aims to contribute to both abovementioned issues. First, we develop an end-to-end scenario of leveraging metabolic models to simplify the design of industrial bio-fermentation processes. Second, we contribute to improving the quality of models of metabolism both on the cellular and on the organism levels.

The overall structure of this dissertation is following. Chapter 1 presents BioLego, a webservice that enables simulation and optimization of bio-fermentation processes. Chapter 2 presents methodology to improve genome-scale cellular metabolic models using high-throughput gene-fitness assays. Chapter 3 presents a systems approach for analysis of human steroidogenesis pathway metabolites. Finally, in Chapter 4 we discuss future directions for each presented project.
Chapter 1. Modeling fermentation processes for biorefinery design – BioLego

Here we describe design and applications of BioLego – a freely downloadable webservice, ready to install in the Microsoft Azure Cloud environment. BioLego provides a friendly and intuitive interface that enables simulation (modeling and evaluation of the expected performance) and optimization of single and two-step fermentation processes.

We developed and implemented a flexible modular modelling approach enabling smooth generation of different multi-organism fermentation configurations consisting of independent encapsulated modules, representing individual organisms. BioLego webservice enables assessment of possible product yields that can be achieved from a given media according to various fermentation configurations. Moreover, it provides means for media content improvement by sensitivity and gradient analyses of received media components. Also, it provides the ability to identify the set of organism reactions, deletion of which can potentially improve the fermentation process. BioLego platform leverages the capabilities of Microsoft Azure Cloud (Microsoft) to allow high-scale simulations, which includes simultaneous analysis of multiple fermentation scenarios, comparison of fermentation potential of multiple feedstock biomasses and analysis of media and organism modifications.

The entire BioLego source code is freely downloadable and ready for Azure deployment at http://wassist.cs.technion.ac.il/~edwardv/STORAGE/biolego2_data/.

Chapter 1 is organized as following. Section 1.1 explains the problem setup we address. Section 1.2 sets mathematical foundations used to develop BioLego modeling. Section 1.3 addresses all novel mathematical aspects developed during this research. Finally, Section 1.4 describes BioLego usage as part of projects published in scientific literature.

Future directions of BioLego project are addressed in Chapter 4.
This chapter is based on the following materials:

- **Vitkin, E.**, Gillis, A., Polikovsky, M., Bender, B., Golberg, A., Yakhini, Z. “Distributed flux balance analysis simulations of serial two-organism biomass fermentation”. *(Under preparation for Bioinformatics)*


- *(Golberg et al., 2015):* Golberg, A., **Vitkin, E.** & Yakhini, Z. “Seaweed Biorefineries: Exergy, Fermentation, and Sustainability Implications; Example of potential production of Bioethanol from Kappaphycus Alvarezzi in Philippines” *Proceedings of ECOS (2015).*


Abstracts of these manuscripts can be found in Appendix A: Abstracts of published manuscripts

*We thank Microsoft Azure for Research for its generous grant, which was used to develop and extend BioLego platform.*
Section 1.1. Background

The economically efficient and socially and environmentally sustainable conversion of solar energy into valuable products is a major contemporary challenge for governments and businesses worldwide (Karp and Richter, 2011). Transportation fuels, electricity, heating, cooling, drinking water, food, animal feed, chemicals, and materials are all potential products of solar energy conversion. One of the pathways to convert solar energy into useful products is biorefining. Biorefineries integrate all the aspects of biomass processing flow, such as feedstock biomass growth (the capture of solar energy and carbon dioxide via plant photosynthesis), harvesting and fermentation together with distribution of the resulted biofuels and elimination of remaining waste (Star-coliBRi., 2011).

Clearly, the composition of the feedstock biomass and the ability of microorganisms to efficiently ferment it are two most critical factors influencing the efficiency biorefineries. While the first is heavily influenced by available local raw supply materials (Fatih Demirbas, 2009), the biorefinery engineers/designers have a relative flexibility in designing the fermentation setup. Although processes in chemical refineries rely on computation simulations for decades, the computational simulation and computer-assisted design of bio-fermentation processes are not sufficiently developed. For example, some existing tools are created to simulate organism feeding trajectories inside bioreactors. Others aim to predict optimal media concentrations to achieve better fermentation yields (Beg et al., 2003; Zhao et al., 2008; Rocha et al., 2014). Here we present a BioLego framework predicting the expected product yield given the composition of the feedstock biomass and the fermentation setup.

Two major problems are to be solved by fermentation designers. First is an elimination of biomass available at hand, when a set of available feedstocks is given and there is a need to identify a potentially interesting product and design a corresponding fermentation setup. Second is the opposite, when the desired product is known and there is a need to identify optimal feedstock and fermentation setup to achieve its maximal production. Clearly, the majority of biorefinery designers will tackle a mix of these problems – a
small set of possible feedstocks enforced by local agricultural specifications and a small set of products with potential interest to the local community.

Once the set of feedstocks which are possibly available for fermentation and the fermentation targets is identified, biorefinery designers need to select organisms for the fermentation process. Natural feedstock biomass is composed of many different molecules, such as monosaccharides (glucose, galactose, rhamnose, xylose, etc.), amino acids (valine, histidine, lysine, etc.), fatty acids (myristic, oleic, palmitic, etc.), fibers (cellulose, hemicellulose, lignin, ulvan, etc.) and others (Table 1-12). Thus, selection of the fermentation setup is a non-trivial issue, since many of the organisms cannot metabolize some of the existing biomass components, leading to significant amount of residual media and to low fermentation efficiency. For example, wild type *Saccharomyces cerevisiae*, which is a first-choice organism for bioethanol production, poorly utilizes carbohydrates such as *xylose*, *rhamnose* and *galactose*. One alternative to overcome this deficiency, and thereby to improve the bioethanol yields, is to leverage synthetic biology methodologies to enhance the fermentation ability of the chosen organism. For example, some studies aim to genetically modify *S. cerevisiae*, to improve uptake mechanisms for different monosaccharides (i.e. xylose). Studies in this direction are undertaken for several years but successful implementation remains an open challenge (van Maris *et al.*, 2006). Another approach is to induce or to increase the required functionality in the organism with broader digestion ability. For instance, Bond-Watts (Bond-Watts *et al.*, 2011) proposed different plasmid inserts into *E. coli* to introduce butanol-producing pathways. However, broader digestion ability of single organisms usually reduces the total yields of desired products. For example, *E.coli* is less efficient in production of ethanol from glucose than *S. cerevisiae* (Kim *et al.*, 2007; Talebnia *et al.*, 2005).

Fermentation by bacterial communities is a natural alternative to genetic modifications of selected organisms. Community members can be selected to naturally digest the broader range of existing biomass compounds and to further convert them to desired products. Indeed, fermentation by communities has some serious drawbacks, such as a need for expertise in growing of several organisms together with an understanding of inter-
organism interactions and of competition for resources. Mathematical modelling of the community-based fermentation process is also more complicated, since the natural inter-organism interactions are not-sufficiently understood (Zomorrodi and Maranas, 2012) (relatively to intra-organism metabolism) and since the mapping of metabolites between models representing individual organisms is not trivial due to historical multiplicity of naming conventions (Durinck et al., 2009; van Iersel et al., 2010). There are several approaches for mathematical modeling of the community behavior. For example, OptCom (Zomorrodi and Maranas, 2012) and its extension d-OptCom (Zomorrodi et al., 2014) methodologies propose a computational framework to describe different intra-species interactions grown together. This framework aims to describe trade-offs between individual vs. community level fitness criteria. Similarly, cFBA (Khandelwal et al., 2013) is a method that integrates inter-species interactions to achieve maximal growth rate of the entire bacterial community. The SUMEX (Zarecki et al., 2014) methodology demonstrates the advantage of maximizing the total molar output exchange minus input exchange of metabolites. As a result, higher SUMEX values correspond to better cellular catabolism ability. Another approach (Hanly and Henson, 2011) tests the fermentation efficiency of co-culturing species which require different media compounds.

One of the major technical problems in mathematical modeling of bacterial community behavior is mapping between different in silico models of organisms. Metabolic models of organisms are usually created by independent groups of organism experts; each of them with their own labeling conventions. Alternatively, models are developed using automatic reconstructions, which enforce same naming convention over the full set of resulted models. Unfortunately, such automatic models currently do not have sufficient quality and act only as initial drafts for further manual validation by separate expert groups. Thus, recognizing that metabolite A in Model1 is the same metabolite as B in Model2, is often a time-consuming and error-prone task.

Another challenge intrinsic to existing metabolic networks is predefined reaction directionality and, specifically, the bi-directional transporter reactions. In some growth conditions a certain reaction may insert a certain metabolite into the cell, while in others the same reaction is facilitated to secrete it to outside environment. Thus, simple
connection of two organisms by a group of shared metabolites can be problematic, since we cannot make any assumption regarding the directionality of the corresponding transport reactions. However, in certain industrial setups we can grow organisms in separate bioreactors, thus having a full control over the above-mentioned reaction directionality. Thus, existing mathematical modeling should be adapted to these realities.

One possible approach to allow better and predictable control of the inter-organism relations in bacterial communities originates from the idea of serial fermentation. In this approach, we trade off the possible benefits of inter-organism interactions in favor of increased control. Briefly, this means constructing a process where each organism is grown separately. First organism is grown on the initial media, while all non-utilized residual media components together with resulted grown and decomposed biomass of this organism are transferred to the next organism in chain. This approach has an advantage in the context of broader biomass components utilization allowing for tighter process control. In addition, serial fermentation removes the demand for community growing expertise (only single organism growing expertise is required), increases the process flexibility (the modification of one fermenting organism will have lower impact on the inter-organism interactions) and simplifies the mathematical process modeling. The last one is vital in the fermentation design stage, since the estimation of expected system efficiency prior to its implementation by simulation will reduce the number of experiments required for process optimization and will provide a new tool for the efficient process design.

Two-step fermentation process is a specific case of serial fermentation, in which only two organisms participate in the process. The implementation of this case is particularly interesting, since it provides practical perspectives on the general serial fermentation process drawbacks and advantages.
Section 1.2. General Methods

1.2.1. Metabolic model of the cell

Genome-scale metabolic network models leverage the existing knowledge of organism biochemistry and genetics to construct a framework for simulating intracellular metabolic processes. The core of the metabolic model contains two major elements: (1) information about the stoichiometry of the metabolic reactions and (2) associating protein-coding genes to the reactions that they catalyze (Edwards et al., 2002).

The stoichiometry is described in terms of metabolite conversion, i.e. metabolite \( m_1 \) is converted to metabolite \( m_2 \) by reaction \( r_1 \). This is usually represented by 2D stoichiometric matrix \( S \) of size \([\text{#Metabolites} \times \text{#Reactions}]\), wherein \( S_{m,r} \) corresponds to stoichiometric coefficient of metabolite \( m \) in the reaction \( r \). In addition, kinetic coefficients, boundary rates or reaction directionality may be known for certain cellular reactions. This information can also be included in the metabolic model.

Association of protein-coding genes to reactions is described in terms of gene-to-reaction mapping, i.e. gene \( g_1 \) encodes for protein \( p_1 \), which catalyzes reaction \( r_1 \). This is usually represented by a set of Boolean rules, combining genes to a protein by an \( \Lambda \) (AND) clause of binary indicators representing gene activity (existence, functionality, expression), meaning that all mentioned genes are required to create (activate) this protein. If a certain reaction can be catalyzed (activated) by one of several alternative proteins, then the activity status of this reaction is represented herein by an \( \vee \) (OR) clause of binary indicators of activities of each protein (Equation 1-1A). For instance, the regular expression for the reaction \( r_0 \), which is catalyzed either by a protein \( p_1 \), encoded by a single gene \( g_1 \), or by a protein \( p_2 \), encoded by a combination of genes \( g_2 \) and \( g_3 \), is represented in Equation 1-1B below.
Equation 1-1: Gene-to-Reaction Boolean mapping

\[(A) \text{activity}(r) = \bigvee_{p_i=1}^{\#Proteins(r)} \left\{ \bigwedge_{g_j^{p_i}=1}^{\#Genes(p_i)} \text{activity}(g_j^{p_i}) \right\} \]

such that: \(\text{activity}(g_j^{p_i}) = \begin{cases} 
    \text{TRUE} & \text{iff gene } g_j^{p_i} \text{ is active} \\
    \text{FALSE} & \text{otherwise}
\end{cases} \)

\[(B) \text{activity}(r_0) = \text{activity}(p_1) \lor \text{activity}(p_2) = (\text{activity}(g_1)) \lor (\text{activity}(g_2) \land \text{activity}(g_3)) \]

This Boolean representation allows simple addressing of gene and reaction activities. For example, when a gene gets knocked out for certain bacteria, such as in fitness assay experiments, the corresponding gene-related activity variable in the above regular expression (Equation 1-1A) is changed to FALSE, which affects the resulted activity state of the matching reaction. One possible implication of this representation, which is relevant in our context, is presented in Equation 1-5 later.

1.2.2. Flux Balance Analysis

BioLego mathematical simulations of biomass utilization and target molecule production yields are performed using the Flux Balance Analysis (FBA) approach. This is a subclass of Constraint-Based Modeling (CBM) mathematical modeling approaches. CBM analyzes internal reaction fluxes based solely on simple physical-chemical constraints, such as reaction stoichiometry and metabolic flux constraints, without requiring exact enzyme kinetic data. Specifically, such a methodology enables the prediction of organism growth rates based only on reaction stoichiometry and directionality. FBA-based approaches have a wide range of applications including phenotype analysis, bioengineering and metabolic model reconstructions.

As already mentioned above (see 1.2.1), the reaction stoichiometry in a metabolic model is represented by a stoichiometric matrix \(S\), wherein \(S_{m,r}\) corresponds to stoichiometric coefficient of metabolite \(m\) in the reaction \(r\). The vector of metabolic fluxes that are carried by the model reactions is denoted as \(v\). One common assumption for the vector \(v\) is so-called mass-balance constraint (Equation 1-2A), meaning that for each metabolite \(m\)
the sum of reaction fluxes producing it is equal to the sum of reaction fluxes consuming
it. Another common constraint is some knowledge about maximal \( v^{UB} \) and/or minimal
feasible fluxes \( v^{LB} \) for each reaction \( r \) (Equation 1-2B). The specific context may
impose additional constraints.

**Equation 1-2: Basic CBM constraints**

\[
\begin{align*}
(A) \quad S * \vec{v} & = 0 \\
(B) \quad v^{LB} \leq v_r \leq v^{UB}
\end{align*}
\]

Although \( v^{UB} \) and \( v^{LB} \) are usually set to \( \pm \infty \) for most of reactions due to lack of
knowledge, the solution space is usually not unbounded; it is limited by constraining the
feedstock media uptake rate. In our specific case of BioLego simulations, the knowledge
of actual media uptake rate is not critical, since we are interested in the total conversion
yield (in %) of biomass into target product, rather than in specific reaction rates.
Therefore, we assume the uptake rate of 1gDW*h\(^{-1}\) of media (1 gram of media dry weight
per one hour), which is enforced by updating the boundaries of media transporter
reactions in Equation 1-2B and calculate the biomass-to-product conversion yield
accordingly. Note, that such uptake rate is very similar to the commonly accepted. For
example, common glucose uptake rate which is accepted for *E. coli* is 1.8gDW*h\(^{-1}\) for
glucose minimal media (Orth *et al.*, 2011).

FBA is a particular case of the CBM framework, assuming that the metabolic network of
the studied organism or system is regulated (e.g. by evolutionary processes) to maximize
some function, which is usually an organism growth rate for unicellular organisms
(Raman and Chandra, 2009).

The FBA framework is presented in Equation 1-3 below:
Equation 1-3: FBA framework

\[ BM_{MAX} = \max_{\vec{v}} \{ v_{Growth} \} \]

s.t.: \[
\begin{align*}
    S * \vec{v} &= 0 \\
    v_r^{LB} &\leq v_r \leq v_r^{UB} \\
    \sum v_{\text{media transporters}} &= 1 \frac{gDW(\text{media})}{h}
\end{align*}
\]

Here, \( v_{Growth} \) is an artificial growth reaction converting all the organism cellular components into a single output variable representing a unit of biomass; \( v_{\text{media transporters}} \) is a generic name for all transporter reactions and \( BM_{MAX} \) is an estimated maximal organism growth rate under the given constraints. Indeed, there may be many possible sets of fluxes that both satisfy all the CBM constraints and maximize \( v_{Growth} \). That is – the optimum is often attained by a set of possible solution vectors, \( \vec{v} \).

Note that under fluxes that maximize growth rate as in Equation 1-3, each non-growth reaction (e.g., ethanol production) may have a range of possible values. Modelling often aims to estimate this range for reactions of interest, such as reactions producing target metabolites. This is formally described in Flux Variability Analysis (FVA) (Mahadevan and Schilling, 2003) framework, as presented in Equation 1-4. Notice that there is an additional constraint enforcing maximal growth rate together with setting the adequate optimization target:

Equation 1-4: FVA framework

\[
\max_{\vec{v}} \min \{ v_{Target} \}
\]

s.t.: \[
\begin{align*}
    v_{Growth}^{LB} &= BM_{MAX} : \text{Solution of Equation 1-3} \\
    S * \vec{v} &= 0 \\
    v_r^{LB} &\leq v_r \leq v_r^{UB} \\
    \sum v_{\text{media transporters}} &= 1 \frac{gDW(\text{media})}{h}
\end{align*}
\]

In practice we solve this optimization task by first determining \( BM_{MAX} \), solving Equation 1-3, and then setting it as a lower limit to \( v_{Growth} \). A range of allowed suboptimal growth rates can also be obtained through this approach, by relaxing this constraint to \( \alpha * BM_{MAX} \) for some \( 0< \alpha<1 \).
The deletion of certain reactions, also referred to as *reaction knockouts*, as obtained by genetic manipulations or by other means, may have a major impact (either positive or negative) on the estimated production yield. First, the deleted reaction can be vital for producing organism biomass (organism growth). In this case $BM_{\text{MAX}}$, which is a maximal solution of Equation 1-3, will be 0. Second, the deleted reaction can be important, but non-vital for organism growth. In this case a decrease in $BM_{\text{MAX}}$ value will be observed. Finally, the deleted reaction may have no any effect on the organism growth, thus leading to no change in $BM_{\text{MAX}}$. Notably, the value of $BM_{\text{MAX}}$ cannot increase with the deletion of any reaction, since the CBM solution space after the deletion is a sub-space of the original CBM space of valid solutions.

Mathematically, to simulate a reaction knockout, Equation 1-3 is modified to become Equation 1-5, where the reaction boundary constraints (Equation 1-2B) are multiplied by the activity status of the reaction, thus effectively nullifying flux values for the deleted reactions (Burgard *et al.*, 2003; Tepper and Shlomi, 2010).

**Equation 1-5: FBA framework with reaction knockouts**

$$BM_{\text{MAX}} = \max_v \{v_{\text{Growth}}\}$$

s. t.: 

$$S \cdot \vec{v} = 0$$

$$\text{activity}_r \cdot v^L_r \leq v_r \leq \text{activity}_r \cdot v^U_r$$

$$\sum v_{\text{media transporters}} = 1 \frac{gDW(\text{media})}{h}$$

A very important generalization/extension of the abovementioned FVA framework (Equation 1-4) with incorporation of the reaction activity constraints is the search for the optimal set of reaction knockouts, leading to the maximal production rate of the target molecule. A formulation of this task was initially proposed by Burgard (Burgard *et al.*, 2003).

First, as we mention previously, each non-growth reaction, such as reactions producing the production target metabolite, may have a range of possible values under the assumption of maximal growth. If the resulting maximal growth rate is below some predefined threshold (i.e. 1e-5), we treat the organism as non-viable and estimate this
range to be 0.00. Otherwise, the estimation of the target production yield is performed as before (Equation 1-4) with the updated value of $BM_{MAX}$ (Equation 1-18).

Notably, the expected production resulted from the reaction knockout can both decrease or increase as compared to the wild type. The latter case can be explained as follows – some part of media components, which is normally used for production of biomass, is now unavailable for this purpose. The other part, which is still available, can be utilized for the product of interest. For instance, suppose organism biomass is composed of two metabolites $A$ and $B$, and it requires stoichiometric ratio of 1:1. If after certain reaction knockout, the amount of $A$ is decreased by 50%, this implies that 50% of $B$ cannot be utilized for biomass production and can be redirected to other directions, such as production of the target molecule.

![Diagram](image)

**Figure 1-1: Coupling maximal growth with the production of target molecule**
We return to the formulation first proposed by Burgard et al as the OptKnock Algorithm (Equation 1-6A) (Burgard et al., 2003). Later this formulation was extended by Tepper and Shlomi to account for alternative pathways as RobustKnock Algorithm (Equation 1-6B) (Tepper and Shlomi, 2010).

Equation 1-6: (A) OptKnock and (B) RobustKnock frameworks for knockout optimization

(A) \[
\max_{activity[i] \in \{0;1\}} \left\{ \max_{\vec{v}} \{v_{Target}\} \right\} \leftarrow \text{OptKnock (Burgard et al., 2003)}
\]

(B) \[
\max_{activity[i] \in \{0;1\}} \left\{ \min_{\vec{v}} \{v_{Target}\} \right\} \leftarrow \text{RobustKnock (Tepper and Shlomi, 2010)}
\]

\[
N - \|activity\| \leq K : K \in \{0,1,2,...\} \leftarrow \text{Maximal number of Knockouts}
\]

\[
v_{Growth}^{LB} = \max_{\vec{v}} \{v_{Growth}\}
\]

s.t.: \[
S * \vec{v} = 0
\]

s.t.: \[
activity_r * v_r^{LB} \leq v_r \leq activity_r * v_r^{UB}
\]

\[
\sum v_{media\ transporters} = \frac{gDW(media)}{h}
\]

Notably, the increased minimal target production (Equation 1-6B) means that a configuration is obtained (through genetic modifications), which couples the reaction that produces the target molecule with the maximal growth rate scenario of the organism (Tepper and Shlomi, 2010). This is depicted on Figure 1-1, where the knockout of reaction R6 enforces the organism to utilize the alternative pathway producing the target molecule M10 as a by-product of maximal growth.
1.2.3. Construction of genome-scale metabolic models

There are two major computational challenges in the construction of genome-scale metabolic networks: (i) the identification of missing reactions in a metabolic network, also referred to as “gap-filling”; and (ii) the association of genes with network reactions, with special interest in orphan reactions, i.e. reactions without any known associated gene in this organism. Clearly, these two challenges are interconnected. The prediction of reaction existence is more reliable, if it is supported by related genomic information. And the prediction of gene-to-reaction assignment is irrelevant, if there is a significant doubt in the organism’s need for this reaction.

The general schema for the construction of genome-scale metabolic model works as following:

- Start from some seed of known reactions, genes and metabolites
- Extend it to complete model, which should adhere to a set of predefined conditions, such as possibility of organism grow, possibility to activate certain (or all) reactions, etc. This is done by applying gap-filling and gene-to-reaction assignment logic together with leveraging all existing knowledge about the target organism, which may include (among many other types of data):
  - Sequence data of the target organism as well as sequence data of other organisms (phylogenetic relationships).
  - Gene-expression profiles, which are expression levels of genes measured over several conditions.
  - Gene-fitness assays, which are essentiality organism growth rates observed after the deletion of genes measured over several conditions.

The reason we specifically mention here these particular types of data (phylogenetic, gene-expression and gene-fitness) is the MIRAGE algorithm, which was developed during my MSc research with Prof. Tomer Shlomi. MIRAGE (MetabolIc Reconstruction via functionAl GEnomics) is an automatic reaction gap-filling algorithm which identifies missing network reactions by integrating metabolic flux analysis together with functional genomics data. Specifically, to reconstruct a metabolic network model for an organism of
interest, MIRAGE starts from a core set of reactions, whose presence is established via strong genomic evidence. Then, to enable the activation of this core set, MIRAGE algorithm identifies most appropriate reactions from other species, whose presence is supported by genes assigned based on functional genomics knowledge, such as phylogenetic and gene-expression data (Vitkin 2012; Vitkin and Shlomi, 2012).

During my PhD research, we further contributed to this approach by developing a methodology to incorporate in the process of gene-to-reaction assignment additional novel type of data, called *gene-fitness assays*. These assays, which are measured over several growth conditions, include levels of organism growth rates received after the knockout of the gene or promoter of interest. Details on the proposed gene-to-reaction assignment methodology and its integration with other functional-genomic data types (paving a natural way for utilization of gene-fitness profiles as part of MIRAGE algorithm) are described in Chapter 2.
Section 1.3. Methods developed for BioLego and related projects

1.3.1. Encapsulated microorganism metabolic models – intuition

Mathematical modelling of the fermentation process based on bacterial community is difficult due to lack of understanding of inter-organism interactions and due to nomenclature differences. To offset these issues we propose a flexible modular modelling approach making the generation of different fermentation configurations akin to combining LEGO (Home - LEGO.com) pieces. In our case each LEGO piece, called module, contains an encapsulated metabolic model of an individual organism received as a preliminary input. The underlying metabolic model can be taken from literature or constructed automatically (Henry et al., 2010; Vitkin and Shlomi, 2012). In the following sections, we address the mathematical aspects of using this approach in formulating BioLego optimization tasks.

1.3.2. The modular approach – principles of constructing single module

Inspired by the compartmentalized consortium approach (Taffs et al., 2009), we address both the issue of mapping between naming conventions and the issue of reaction directionality (see Section 1.1) by encapsulating the existing organism metabolic model with an additional layer, we call envelope. This reduces the problem of mapping between each newly introduced metabolic model to all others to the problem of mapping between this model and the envelope (Figure 1-2), thus significantly diminishing the inter-model mapping problem, although not eliminating it. To implement this approach we created a dataset of new metabolite identifiers based on the proposed convention of KEGG DB (Ogata et al., 1999) and are continuously updating it with the integration of new underlying metabolic models in the BioLego website.
Figure 1-2: Single-module for existing organism metabolic model. Each contains five chambers: Internal – consisting of the known metabolic model of the organism; Media – representing received biomass feedstock media; Product – representing desired target metabolite; Growth – consisting organism cellular components; Waste – representing all non-digested media residuals and molecules by organism as growth by-product.

For each organism we create a single module, composed of two principle parts – the Internal chamber, which is essentially an existing organism metabolic model, and the four chambers of the envelope layer: Media, Growth, Waste and Product. Media refers to the set of media compounds received by organism; Growth refers to the set of molecules comprising the organism biomass; Waste refers to all the compounds either non-digested from media or extracted as by-product of metabolic processes; and Product refers to the desired product. Obviously, each such module should be constructed dynamically according to the setup-specific media composition and the defined target product molecule.

The inter-chamber metabolic fluxes are defined to be unidirectional, as presented on Figure 1-2. Also, note that there is no direct metabolic flux connecting the Internal chamber and the out-of-module space, which allows full control over all the system inputs and outputs akin to the expected industrial process. There is a clear mathematical detachment between the digested metabolite $A$ and extracted metabolite $A$, although practically it can be the same molecule. This feature is a key to enabling simple modelling that faithfully represents a controllable process even for multi-organism fermentation scenarios.
1.3.3. The modular approach – mathematical formulation of single module

Several adjustments to the general CBM constraints (Equation 1-2) are required to create the encapsulation of the existing metabolic models as described above (Figure 1-2). First, the encapsulation of the underlying organism stoichiometric model ($S$-matrix from Equation 1-2A) is created, resulting in the matrix $S_{Env}$ (Figure 1-3).

For each exterior non-product metabolite from the fermented media we add a row both into Media and into Waste chambers. Another row(s) is created for the desired product as a Product chamber. Finally, additional rows are created for each molecule consisting organism biomass to construct Growth chamber.

![Figure 1-3: Construction of the matrix $S_{Env}$. 1 stands for unitary matrix of appropriate dimension (varies from organism to organism). All empty cells contain 0-matrix of appropriate dimensions, which is not shown for visual compactness. The emphasized left side ($S_{Env-INSIDE}$) of the figure is used later for construction of multi-modular processes (see 1.3.5)]
Once all novel rows are introduced into the matrix, we add reaction columns corresponding to the permitted inter-chamber flux directionality (Figure 1-2, Figure 1-3):

- Media → Internal
- Media → Waste
- Internal → Waste
- Internal → Product
- Internal → Growth

For instance, to allow Internal → Growth flux, we add columns (reactions) transferring the biomass metabolites from the original metabolic model to the metabolites of Growth chamber (Figure 1-3), thus representing the decomposition process of organism cellular components possibly existing between the bioreactors.

In addition, module interface columns ($S_{Env-OUTSIDE}$) are added to simulate metabolite exchange with the environment (Figure 1-2, Figure 1-3):

- Outside → Media
- Waste → Outside
- Product → Outside
- Growth → Outside

These interface columns, composing $S_{Env-OUTSIDE}$ part of the $S_{Env}$, are omitted during the construction of the multi-organism (i.e. two-step) fermentation processes. In other words, to simulate the multi-organism fermentation pipes we use only $S_{Env-INISIDE}$ part of the matrix $S_{Env}$ (see 1.3.5)

The upper and lower boundaries for each of the newly added inter-chamber reactions in the system are set to $[0; +\infty]$, thus mathematically enforcing unidirectionality of the existing inter-chamber metabolic flux.

To estimate the fermentation characteristics of a single module, i.e. the fermentation potential of a single organism, the bi-level FVA (Mahadevan and Schilling, 2003) framework is formulated as in Equation 1-7, where we check that $BM_{MAX} \geq \varepsilon$ and return 0.0 otherwise.
Equation 1-7: FVA framework for single module

1) LET: \( BM_{MAX} = \max_{\bar{v}} \{ v_{Growth} \} \)

\[
\begin{align*}
S_{Env} \times \bar{v} &= 0 \\
\nu_r^{LB} &\leq v_r \leq \nu_r^{UB} \quad : \forall r \text{ in original organism model} \\
0 &\leq v_r \leq +\infty \quad : \forall r \text{ in inter – chamber reactions} \\
\sum v_{Outside\rightarrow Media} &= 1 \frac{gDW(media)}{h}
\end{align*}
\]

IF \( BM_{MAX} < \varepsilon \):

SET: \( \max \& \min_{\bar{v}} \{ v_{Target} \} = 0 \) and FINISH

2) Compute: \( \max \& \min_{\bar{v}} \{ v_{Target} \} \)

\[
\begin{align*}
\nu_r^{LB} &= BM_{MAX} \\
S_{Env} \times \bar{v} &= 0 \\
\nu_r^{LB} &\leq v_r \leq \nu_r^{UB} \quad : \forall r \text{ in original organism model} \\
0 &\leq v_r \leq +\infty \quad : \forall r \text{ in inter – chamber reactions} \\
\sum v_{Outside\rightarrow Media} &= 1 \frac{gDW(media)}{h}
\end{align*}
\]

1.3.4. The modular approach – principles of combining modules

Figure 1-4: Combination of two single-organism modules to characterize two-step serial fermentation process. In addition to the modules of participating organisms, we added four auxiliary modules to simulate real industrial environment: Feedstock Media, Control Layer, Total Waste and Total Product (see 1.3.5).
Once we create module with the envelope layer around each organism metabolic model (see 1.3.2 and 1.3.3), any combination of these envelopes is as simple as combining LEGO (Home - LEGO.com) pieces. All that is needed is to define the inter-organism relationship rules of the desired fermentation process.

For example, the serial two-step fermentation process can be enforced by the following rules (Table 1-1), corresponding to bold arrows in Figure 1-4:

<table>
<thead>
<tr>
<th>Rule ID</th>
<th>Rule (Arrow in Figure 1-4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1.1</td>
<td>Feedstock Media $\rightarrow$ Media (Org1)</td>
</tr>
<tr>
<td>Rule 1.2a</td>
<td>Waste (Org1) $\rightarrow$ Control Layer</td>
</tr>
<tr>
<td>Rule 1.2b</td>
<td>Control Layer $\rightarrow$ Media (Org2)</td>
</tr>
<tr>
<td>Rule 1.3a</td>
<td>Growth (Org1) $\rightarrow$ Control Layer</td>
</tr>
<tr>
<td>Rule 1.3b</td>
<td>Control Layer $\rightarrow$ Media (Org2)</td>
</tr>
<tr>
<td>Rule 1.4</td>
<td>Product (Org1) $\rightarrow$ Total product</td>
</tr>
<tr>
<td>Rule 1.5</td>
<td>Waste (Org2) $\rightarrow$ Total waste</td>
</tr>
<tr>
<td>Rule 1.6</td>
<td>Growth (Org2) $\rightarrow$ Total waste</td>
</tr>
<tr>
<td>Rule 1.7</td>
<td>Product (Org2) $\rightarrow$ Total product</td>
</tr>
</tbody>
</table>

Table 1-1: Inter-module interaction rules for two-step serial fermentation process

For comparison, the single-step fermentation process, designed as co-culturing of two organisms, is represented by the set of rules in Figure 1-5 and Table 1-2. Note, that Control Layer do not exist here, since we cannot intervene the metabolite exchange between organisms. Rule 1.3 (from Table 1-1) transferring “Growth (Org 1)” to “Media (Org 2)” does not exist in the co-culture system, since cellular components of Organism1 cannot be decomposed and returned to media in the real time. In addition, in the co-culture system, we introduce additional possible metabolite flux directions, transferring particles secreted by the second organism as media to the first, as well as enabling the entire initial feedstock media to the second organism (Rule 2.7 and Rule 2.6 respectively). It is important to emphasize the fact that there is no direct flux between the Internal chambers of two organisms’ modules.
<table>
<thead>
<tr>
<th>Rule ID</th>
<th>Rule Description</th>
<th>Comparison to Table 1-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 2.1</td>
<td>Feedstock Media $\rightarrow$ Media (Org1)</td>
<td>Same as rule 1.1</td>
</tr>
<tr>
<td>Rule 2.2</td>
<td>Waste (Org1) $\rightarrow$ Media (Org2)</td>
<td>Same as rule 1.2a &amp; 1.2b</td>
</tr>
<tr>
<td>Rule 2.3</td>
<td>Waste (Org1) $\rightarrow$ Total waste</td>
<td>Novel</td>
</tr>
<tr>
<td>Rule 2.4</td>
<td>Growth (Org1) $\rightarrow$ Total waste</td>
<td>Novel</td>
</tr>
<tr>
<td>Rule 2.5</td>
<td>Product (Org1) $\rightarrow$ Total product</td>
<td>Same as rule 1.4</td>
</tr>
<tr>
<td>Rule 2.6</td>
<td>Feedstock Media $\rightarrow$ Media (Org2)</td>
<td>Novel</td>
</tr>
<tr>
<td>Rule 2.7</td>
<td>Waste (Org2) $\rightarrow$ Media (Org1)</td>
<td>Novel</td>
</tr>
<tr>
<td>Rule 2.8</td>
<td>Waste (Org2) $\rightarrow$ Total waste</td>
<td>Same as rule 1.5</td>
</tr>
<tr>
<td>Rule 2.9</td>
<td>Growth (Org2) $\rightarrow$ Total waste</td>
<td>Same as rule 1.6</td>
</tr>
<tr>
<td>Rule 2.10</td>
<td>Product (Org2) $\rightarrow$ Total product</td>
<td>Same as rule 1.7</td>
</tr>
</tbody>
</table>

Table 1-2: Module interaction rules for single-step fermentation process, designed as co-culturing of two organisms

Figure 1-5. Combination of two single-organism modules to characterize single-step fermentation process, designed as co-culturing of two organisms. Three external boundary modules are added to the process in addition to the modules of participating organisms: Feedstock Media, Total Waste and Total Product.
1.3.5. The modular approach – mathematical formulation of two-organism fermentation process

Simulation of the two-organism fermentation process introduces additional level of complexity. We developed a mathematical modeling formulation and framework for this task, as presented here (Figure 1-4 and Figure 1-5).

Serial two-step fermentation: General mathematical formulation

First, we address the combination of two single-organism modules to characterize two-step serial fermentation process (Figure 1-4). For this purpose, we combine both used organism modules into a single stoichiometric system to create the matrix $S_{System}$ (Figure 1-6) based on the appropriate set of rules (Table 1-1):

![Figure 1-6: Construction of the matrix $S_{System}$ for serial two-step fermentation.](image)

$I$ stands for unitary matrix of appropriate dimension (varies from organism to organism). All empty cells are equal to 0-matrix of appropriate dimension, which is not shown for visual compactness. For the same reason we omit all intra-module reactions described in Figure 1-3.
The matrix $S_{\text{System}}$ is created based on $S_{\text{Env-INSIDE}}$ parts of single-module matrix $S_{\text{Env}}$ of both participating organisms. The module-level interface parts, describing interactions with the environment ($S_{\text{Env-OUTSIDE}}$ in 1.3.3) are omitted.

Four additional auxiliary modules are created to cover the interactions with the industrial environment: \textit{Feedstock Media}, \textit{Control Layer}, \textit{Total Product} and \textit{Total Waste}. Respectively, they correspond to the composition of the input feedstock biomass, any compound manipulations between the organisms (i.e. addition of oxygen, vitamins or other molecules non-existing in original feedstock but essential for second organism), total product received from all participating organisms and the remaining undigested particles, which includes organism biomasses. Columns representing inter-module metabolic fluxes are introduced according to Table 1-1). Together with this, system-level interface columns describing metabolic exchanges with the environment are also created (Figure 1-4 and Figure 1-6):

- From \textit{Outside} to \textit{Feedstock Media}
- From \textit{Outside} to \textit{Control Layer}
- From \textit{Total Waste} to \textit{Outside}
- From \textit{Total Product} to \textit{Outside}

Again, the reaction boundaries for all newly added reactions are set to $[0; +\infty]$, like in inter-chamber case.

The resulting \textit{Total Product} maximization and minimization estimations are formulated according to Equation 1-8. Briefly, we want to estimate the maximum and minimum values for $v_{\text{Total Product}}$ under the constraint that the 1\textsuperscript{st} organism grows at its maximum for the given conditions and the 2\textsuperscript{nd} organism grows at its maximum on the remainders from the 1\textsuperscript{st} one. Indeed, both organisms should grow with the rate of at least $\varepsilon$ (constraint $c5$), otherwise we switch to the single module scenario (Equation 1-7). For the simplicity of the following discussion we omit the constraint $c5$ from further notations.
Equation 1-8: General FVA framework for serial two-step combination of two modules

Compute: \( \max_{\tilde{v}} \min \{ v_{total\ prod} = v[Org1\ prod] + v[Org2\ prod] \} \)

\[
\begin{align*}
\text{s.t.:} & \quad \text{c1). } S_{System} * \tilde{v} = 0 \\
& \quad \text{c2). } v^{LB} \leq \tilde{v} \leq v^{UB} \\
& \quad \text{c3). } v[BM_{Org1}] \text{ is maximal subject to [c1 and c2]} \\
& \quad \text{c4). } v[BM_{Org2}] \text{ is maximal subject to [c1 and c2 and given } v_{Org1} \text{ from c3]} \\
& \quad \text{c5). } v[BM_{Org1}] \geq \varepsilon \text{ AND } v[BM_{Org2}] \geq \varepsilon \\
\end{align*}
\]

where \( c_2 = \begin{cases} 
2.1). & v_r^{LB} \leq v_r \leq v_r^{UB} : \forall r \text{ in original organism models} \\
2.2). & 0 \leq v_r \leq +\infty : \forall r \text{ in novel envelope reactions} \\
2.3). & \sum_{Outside\rightarrow Feedstock\ Media} v_{Out} = 1 \frac{gDW(\text{media})}{h} 
\end{cases} \)

Serial two-step fermentation: Transformation to single-level optimization

The Equation 1-8 represents three-level optimization framework, for which a constraint \( c_4 \) is an internal bi-level optimization. The direct solution, (from inside to outside, as implemented in Equation 1-7) is infeasible, because in general case constraint \( c_3 \) has infinite number of fluxes giving maximal growth value. To handle this, we redesign the formulation to be single-level, using the primal-dual approach in a spirit of the proposed by Burgard et al in OptKnock algorithm (Burgard et al., 2003).

First, we transform all bidirectional reactions to two unidirectional. Specifically, each bidirectional reaction \( v_r \) is transformed to two reactions \( v_{r1} \) and \( v_{r2} \) as in Equation 1-9. All the relevant constraints are updated respectively.

Equation 1-9: Transformation of bi-directional reaction \( v_r \)

\[
v_r : [v_r^{LB} \leq v_r \leq v_r^{UB}] \rightarrow \begin{cases} 
v_r \geq 0, v_{r2} \geq 0 : [v_r^{LB} \leq v_r = v_{r1} - v_{r2} \leq v_r^{UB}] 
\end{cases}
\]
Now, let’s present c3 from Equation 1-8 in its standard primal (Equation 1-10a) and dual (Equation 1-10b) forms.

**Equation 1-10:** Constraint c3 from Equation 1-8 in its primal (a) and dual forms (b)

\[
\begin{align*}
\text{a. max}_v \{ v[BM_{Org1}] = c_1^T \ast \bar{v} \} & \quad \text{b. min}_{u_1} \{ b_1^T \ast \bar{u}_1 \} \\
\begin{align*}
(a1) \quad & S_{System} \ast \bar{v} = 0 \\
(a2) \quad & A_1 \ast \bar{v} \leq b_1 : A_1 = \begin{bmatrix} I \end{bmatrix}, b_1 = \begin{bmatrix} v^{UB} \end{bmatrix} \\
(a3) \quad & v[r] \geq 0 : \forall r
\end{align*}
\end{align*}
\]

\[
\begin{align*}
\text{s.t.} & : [b1]. \quad A_1^T \ast \bar{u}_1 \geq c_1 \\
& : [b2]. \quad u_1[m] \geq 0 : \forall m
\end{align*}
\]

Since media is limited the Equation 1-10a has a finite optimal solution. Therefore, according to strong duality theorem, the solutions of primal and dual forms are equal. Therefore, we can reformulate the constraint c4 from Equation 1-8 as in Equation 1-11:

**Equation 1-11:** Constraint c4 from Equation 1-8 reformulated based on Equation 1-10

\[
\begin{align*}
\max_{v, u_1} \{ v[BM_{Org2}] \} & \quad \max_{v, u_1} \{ v[BM_{Org2}] \} \\
\begin{align*}
& : [1]. \quad S_{System} \ast \bar{v} = 0 \\
& : [2]. \quad v^{LB} \leq \bar{v} \leq v^{UB} \\
& : [3a1]. \quad S_{System} \ast \bar{v} = 0 \\
& \quad \vdots \\
& : [3a2]. \quad A_1 \ast \bar{v} \leq b_1 \\
& : [3a3]. \quad \bar{v} \geq 0 \\
& : [3b1]. \quad A_1^T \ast \bar{u}_1 \geq c_1 \\
& : [3b2]. \quad \bar{u}_1 \geq 0
\end{align*}
\end{align*}
\]

Now, we repeat the process above for the constraint c4 from Equation 1-8. First, we present it in its standard primal (Equation 1-12a) and dual (Equation 1-12b) forms.

**Equation 1-12:** Constraint c4 from Equation 1-8 in its primal (a) and dual forms (b)

\[
\begin{align*}
\text{a. max}_{u_1} \{ v[BM_{Org2}] = c_2^T \ast \bar{u}_1 \} & \quad \text{b. min}_{u_2} \{ b_2^T \ast \bar{u}_2 \} \\
\begin{align*}
\begin{align*}
(a1) \quad & S_{System} \ast \bar{v} = 0 \\
(a2) \quad & c_1^T \ast \bar{v} = b_1^T \ast \bar{u}_1 \\
(a3) \quad & A_2 \ast \bar{u}_1 \leq b_2 : A_2 = \begin{bmatrix} A_1, 0 \end{bmatrix}, b_2 = \begin{bmatrix} b_1 \end{bmatrix} \\
(a4) \quad & \bar{u}_1 \geq 0
\end{align*}
\end{align*}
\end{align*}
\]

\[
\begin{align*}
\text{s.t.} & : [b1]. \quad A_1^T \ast \bar{u}_2 \geq c_2 \\
& : [b2]. \quad \bar{u}_2 \geq 0
\end{align*}
\]
Again, from the strong duality theorem we deduce that the solutions of primal and dual forms are equal. Therefore, we can reformulate the Equation 1-8 as single-level optimization framework, presented at Equation 1-13, where all the variables are defined as previously.

**Equation 1-13:** Single-level optimization FVA framework for serial two-step combination of two modules based on Equation 1-11 and on Equation 1-12

\[
\max_{v, u_1, u_2} \min \{ v_{Total\ Product} \} \\
\begin{align*}
\text{subject to} & \\
& c1. S_{System} \ast \bar{v} = 0 \\
& c2. v^B \leq \bar{v} \leq v^U \\
& c3.1. S_{System} \ast \bar{v} = 0 \\
& c3.2. c_1^T \ast \bar{v} = b_1^T \ast \bar{u}_1 \\
& c3.3. A_1 \ast \bar{v} \leq b_1 \\
& c3.4. A_1^T \ast \bar{u}_1 \geq c_1 \\
& c3.5. \bar{v} \bar{u}_1 \geq 0 \\
& c4. a1. S_{System} \ast \bar{v} = 0 \\
& c4. a2. c_1^T \ast \bar{v} = b_1^T \ast \bar{u}_1 \\
& c4. a3. A_2 \ast \bar{v} \bar{u}_1 \leq b_2 \\
& c4. a4. \bar{v} \bar{u}_1 \geq 0 \\
& c4. b1. A_2^T \ast \bar{u}_2 \geq c_2 \\
& c4.2. \bar{u}_2 \geq 0
\end{align*}
\]

\[
\max_{v_1, u_2} \min \{ v_{Total\ Product} \} \\
\begin{align*}
\text{subject to} & \\
& k1. S_{System} \ast \bar{v} = 0 \\
& k2. c_1^T \ast \bar{v} = b_1^T \ast \bar{u}_1 \\
& k3. c_2^T \ast \bar{v} \bar{u}_1 = b_2^T \ast \bar{u}_2 \\
& k4. A_2 \ast \bar{v} \bar{u}_1 \leq b_2 \\
& k5. A_2^T \ast \bar{u}_2 \geq c_2 \\
& k6. \bar{v} \bar{u}_1 \geq 0
\end{align*}
\]

To summarize, we estimate maximal and minimal values for reaction \( v_{Total\ Product} \) under all possible values of reaction vector \( \bar{v} \) and auxiliary vectors \( \bar{u}_1 \) and \( \bar{u}_2 \) subject to network stoichiometry (\( k1 \)), maximal growth rate of the 1st organism (\( k2 \)), maximal growth rate of the 2nd organism on the left-overs from the 1st (\( k3 \)) and vector boundaries (\( k4,k5,k6 \)).

**Serial two-step fermentation: Practical perspective**

In practice, to simplify the implementation process we relax the original formulation of serial two-step fermentation (Equation 1-8) to four single-level optimization tasks as presented in Equation 1-14. Briefly, we first estimate the maximal growth rate of the first organism in the pipe and switch to the single-step fermentation with second organism only, in case first organism grows too slowly, i.e. \( BM_{Org1} < \varepsilon \). Second, akin to Equation
1-7, we estimate maximal production of the first organism during its maximal growth. Then we estimate the maximal growth rate of the second organism under the two constraints of 1) maximal growth rate of the first organism and 2) maximal production rate of the first organism. The intuition behind this step is that second organism fluxes should not have any impact on the fluxes in the first. Thus, maximal solution space should be ready to use at the reactions of interest within the first organism. Again, if the second organism grows too slowly ($BM_{Org2|Org1} < \varepsilon$), we switch to the single-step estimations. Finally, we estimate the expected range of Total Product under the growth constraints for both organisms. Note, that at this last step we do not assume any constraint on $v_{Org1 product}$ reaction.

Equation 1-14: Relaxed FVA framework for serial two-step combination of two modules

1) LET: $BM_{Org1} = \max_{\vec{v}} \{v_{Org1 growth}\}$

$$\begin{bmatrix}
S_{System} * \vec{v} = 0 \\
v_{LB} \leq v_r \leq v_{UB}^{LB} & : \forall r \text{ in original organism model} \\
0 \leq v_r \leq +\infty & : \forall r \text{ in novel envelope reactions} \\
\sum v_{Outside\rightarrow Feedstock Media} = 1 \frac{gDW(media)}{h}
\end{bmatrix}$$

IF $BM_{Org1} < \varepsilon$ :

Compute $v_{Total Product}$ based only on the second organism module (Equation 1-7) and FINISH

2) LET: $MaxProd_{Org1} = \max_{\vec{v}} \{v_{Org1 product}\}$

$$\begin{bmatrix}
S_{System} * \vec{v} = 0 \\
v_{LB} \leq v_r \leq v_{UB}^{LB} & : \forall r \text{ in original organism model} \\
0 \leq v_r \leq +\infty & : \forall r \text{ in novel envelope reactions} \\
\sum v_{Outside\rightarrow Feedstock Media} = 1 \frac{gDW(media)}{h} \\
v_{Org1 growth}^{LB} = BM_{Org1}
\end{bmatrix}$$

3) LET: $BM_{Org2|Org1} = \max_{\vec{v}} \{v_{Org2 growth}\}$
\[
\begin{align*}
S_{\text{System}} \ast \bar{v} &= 0 \\
v_r^{LB} &\leq v_r \leq v_r^{UB} : \forall r \text{ in original organism model} \\
0 &\leq v_r \leq +\infty : \forall r \text{ in novel envelope reactions} \\
\text{s.t.:} & \sum v_{\text{Outside} \rightarrow \text{Feedstock Media}} = 1 \frac{gDW(\text{media})}{h} \\
v_{\text{Org1 growth}}^{LB} &= BM_{\text{Org1}} \\
v_{\text{Org1 product}}^{LB} &= \text{MaxProd}_{\text{Org1}} \\
\end{align*}
\]

**IF** $BM_{\text{Org2}|\text{Org1}} < \varepsilon$ :

**Compute** $v_{\text{Total Product}}$ **based only on the first organism module** (Equation 1-7) and **FINISH**

**4) Compute:** \[
\text{max}&\text{min}\{v_{\text{Total Product}} = v_{\text{Org1 product}} + v_{\text{Org2 product}}\}
\]

\[
\begin{align*}
S_{\text{System}} \ast \bar{v} &= 0 \\
v_r^{LB} &\leq v_r \leq v_r^{UB} : \forall r \text{ in original organism model} \\
0 &\leq v_r \leq +\infty : \forall r \text{ in novel envelope reactions} \\
\text{s.t.:} & \sum v_{\text{Outside} \rightarrow \text{Feedstock Media}} = 1 \frac{gDW(\text{media})}{h} \\
v_{\text{Org1 growth}}^{LB} &= BM_{\text{Org1}} \\
v_{\text{Org2 growth}}^{LB} &= BM_{\text{Org2}|\text{Org1}} \\
\end{align*}
\]

**Co-culture of two organisms**

Implementation of the scenario with co-culturing two-organisms in a single step fermentation process (Figure 1-5) is very similar to above. Here we construct the matrix $S_{\text{System}}$ according to the rules from Table 1-2, as in Figure 1-7.

The FVA equation is updated accordingly, to correctly address this setup and specifically the selected inter-organism collaboration objective. Last one is currently an open issue and should be carefully adapted for specific bacterial community. For example, for a naïve assumption that the given community aims to maximize its growth (which is indeed a clear oversimplification), we get Equation 1-15.
Equation 1-15: FVA framework for single-step co-culture combination of two modules

1) LET: \(BM_{MAX} = \max_{\vec{\nu}} \{v_{org1 growth} + v_{org2 growth}\}\)

\[
\begin{align*}
S_{System} \cdot \vec{\nu} &= 0 \\
v_r^{LB} &\leq v_r \leq v_r^{UB} : \forall r \text{ in original organism model} \\
0 &\leq v_r \leq +\infty : \forall r \text{ in novel envelope reactions} \\
\sum v_{Outside\rightarrow Media} &= 1 \frac{gDW(media)}{h}
\end{align*}
\]

\[
\text{IF } BM_{MAX} < \varepsilon : \\
\text{SET: max\&min}_p \{v_{TotalProduct}\} = 0 \text{ and FINISH}
\]

2) Compute: \(\max_{\vec{\nu}} \min \{v_{Total Product} = v_{Org1 product} + v_{Org2 product}\}\)

\[
\begin{align*}
S_{System} \cdot \vec{\nu} &= 0 \\
v_r^{LB} &\leq v_r \leq v_r^{UB} : \forall r \text{ in original organism model} \\
0 &\leq v_r \leq +\infty : \forall r \text{ in novel envelope reactions} \\
\sum v_{Outside\rightarrow Feedstock\ Media} &= 1 \frac{gDW(media)}{h} \\
v_{org1 growth} + v_{org2 growth} &\geq BM_{MAX}
\end{align*}
\]

The estimation is very similar to the single-step based on one organism (Equation 1-7). First, estimate the maximal community growth rate and stop with result of \(TotalProduct=0.0\) if too slow \((BM_{MAX} < \varepsilon)\). Otherwise, we estimate the expected range of \(TotalProduct\) under the resulted growth constraint.
Figure 1-7: Construction of the matrix $S_{System}$ for single-step fermentation by two-organism co-culture. I stands for unitary matrix of appropriate dimension (varies from organism to organism). All empty cells are equal to 0-matrix of appropriate dimension, which is not shown for visual compactness. For the same reason we omit all intra-module reactions described in Figure 1-3.

It is important to emphasize that the proposed modular approach results in significant simplicity of the construction of different fermentation scenarios. It also allows simple and smooth replacement of the participating individual models or organisms.
1.3.6. Analysis of feedstock media by product gradient and media sensitivity

Biomass fermentation pipeline can benefit from minor modifications of the feedstock media, for example from selection of species with biomass containing more of one certain component and less of another. In BioLego we provide options to check the effect of cancelling certain media components, i.e. media sensitivity analysis, and to check the effect of minor increase of certain media components, i.e. (right) gradient analysis on the resulting production rate.

During the media sensitivity analysis, we iteratively remove media components one at a time by nullifying the value of appropriate reaction boundary \(v^{UB}\) in Media chamber (Equation 1-7, Equation 1-14) and repeating the estimation process. The resulting sensitivity value is a ratio of the new and the original production yields (Equation 1-16).

**Equation 1-16: Media sensitivity analysis**

\[
\text{Sensitivity}(\text{compound}) = \frac{v_{\text{TotalProduct}}\{s.t.: v_{\text{Outside}\rightarrow\text{Media}(\text{compound})} = 0\}}{v_{\text{TotalProduct}}\{\text{original setup}\}}
\]

For the gradient analysis we iteratively increase the amount of media components one at a time by adding small \(\delta\) (i.e. \(\delta=1\text{mg per 1g of growth media}\)) value to the appropriate reaction boundary \(v^{UB}\) in media chamber (Equation 1-7, Equation 1-14) and repeating the estimation process. The resulting gradient component is equal to the difference between new and the original production yields (Equation 1-17).

**Equation 1-17: Media gradient analysis**

\[
\text{Gradient}(\text{compound}) = \left[\frac{v_{\text{TotalProduct}}\{s.t.: v_{\text{Outside}\rightarrow\text{Media}(\text{compound})}^{+} = \delta\}}{v_{\text{TotalProduct}}\{\text{original setup}\}}\right] / \delta
\]

Media sensitivity and gradient analysis methods are targeted to identify minor media adjustments that can result from changes in feedstock sub-species, fermentation fertilizers or from changes in feedstock biomass growing condition. To evaluate the expected efficiency associated with more significant media changes, i.e. obtained by mixing different feedstocks or by other major feedstock updates, we recommend re-evaluating the fermentation process with the expected new media composition(s).
1.3.7. Addressing changes in growth conditions. Adjustment of oxygen

Ability to efficiently simulate alternative growth conditions for the same fermentation scenario is one of key features of the proposed framework. We demonstrate it by providing an option to check the expected fermentation yield for various input amounts of oxygens. As part of the input to the simulation, user selects the expected amount of oxygen – either anaerobic or aerobic, and if latter – then to which extent. Each such supplied oxygen amount $\alpha$ results in a specific dedicated optimization task with a modified value of oxygen transporter reaction boundary: $v_{\text{Outside} \rightarrow \text{Media(Oxygen)}}^{UB} = \alpha$.

1.3.8. Fermentation pipeline optimization by reaction knock-outs

Biomass fermentation pipeline can be optimized not only by a smart selection of participating organisms and their ordering, but also by internal organism modifications. Our software supports the evaluation of the effect of canceling certain reactions on the resulting production rates. As before, to simulate knockouts for the given set of reactions we nullify the bits corresponding to the reactions in this set in the corresponding reaction activity vector. Specifically, the integration of reaction activity vectors within serial two-step fermentation (Equation 1-14) leads to framework formulated in Equation 1-18.

In the BioLego system we provide an option to simulate any number of simultaneously deleted reactions in each organism in the pipe. Specifically, the user has an option to select in each organism a specific reaction to knockout or to perform an exhaustive scanning of all possible reactions and reaction combinations. This selection can be performed in a nested scheme, whereby at each repetition the new knockouts are added to the previously selected ones.
Equation 1-18: FVA framework with reaction activity vectors for a serial two-step combination of two organisms

For each reaction activity vector \( \text{act} \) of interest:

1) LET: \( BM_{\text{Org1}} = \max_{\vec{v}} \{ v_{\text{Org1 growth}} \} \)

\[
\begin{align*}
\text{s.t.:} & \\
& S_{\text{System}} \ast \vec{v} = 0 \\
& \text{act}_r \ast v_{r_{LB}}^{LB} \leq v_r \leq \text{act}_r \ast v_{r_{UB}}^{UB} : \forall r \\
& \sum v_{\text{Outside} \rightarrow \text{Feedstock Media}} = 1 \frac{gDW(\text{media})}{h} \\
\end{align*}
\]

IF \( BM_{\text{Org1}} < \varepsilon \):

Compute \( v_{\text{Total Product}} \) based only on the second organism module updated according to \( \text{act} \) and FINISH

2) LET: \( \text{MaxProd}_{\text{Org1}} = \max_{\vec{v}} \{ v_{\text{Org1 product}} \} \)

\[
\begin{align*}
\text{s.t.:} & \\
& S_{\text{System}} \ast \vec{v} = 0 \\
& \text{act}_r \ast v_{r_{LB}}^{LB} \leq v_r \leq \text{act}_r \ast v_{r_{UB}}^{UB} : \forall r \\
& \sum v_{\text{Outside} \rightarrow \text{Feedstock Media}} = 1 \frac{gDW(\text{media})}{h} \\
& v_{\text{Org1 growth}}^{LB} = BM_{\text{Org1}} \\
& v_{\text{Org1 product}}^{LB} = \text{MaxProd}_{\text{Org1}} \\
\end{align*}
\]

3) LET: \( BM_{\text{Org2} \mid \text{Org1}} = \max_{\vec{v}} \{ v_{\text{Org2 growth}} \} \)

\[
\begin{align*}
\text{s.t.:} & \\
& S_{\text{System}} \ast \vec{v} = 0 \\
& \text{act}_r \ast v_{r_{LB}}^{LB} \leq v_r \leq \text{act}_r \ast v_{r_{UB}}^{UB} : \forall r \\
& \sum v_{\text{Outside} \rightarrow \text{Feedstock Media}} = 1 \frac{gDW(\text{media})}{h} \\
& v_{\text{Org1 growth}}^{LB} = BM_{\text{Org1}} \\
& v_{\text{Org2 growth}}^{LB} = BM_{\text{Org2} \mid \text{Org1}} \\
\end{align*}
\]

IF \( BM_{\text{Org2} \mid \text{Org1}} < \varepsilon \):

Compute \( v_{\text{Total Product}} \) based only on the first organism module updated according to \( \text{act} \) and FINISH

4) Compute: \( \max\text{Prod}(\text{act}) \& \min\text{Prod}(\text{act}) = \max_{\vec{v}} \min\{ v_{\text{Total Product}} \} \)

\[
\begin{align*}
\text{s.t.:} & \\
& S_{\text{System}} \ast \vec{v} = 0 \\
& \text{act}_r \ast v_{r_{LB}}^{LB} \leq v_r \leq \text{act}_r \ast v_{r_{UB}}^{UB} : \forall r \\
& \sum v_{\text{Outside} \rightarrow \text{Feedstock Media}} = 1 \frac{gDW(\text{media})}{h} \\
& v_{\text{Org1 growth}}^{LB} = BM_{\text{Org1}} \\
& v_{\text{Org2 growth}}^{LB} = BM_{\text{Org2} \mid \text{Org1}} \\
\end{align*}
\]
1.3.9. Scalable Architecture

Simulating each scenario of interest can reach few seconds. Working on a single machine becomes infeasible in cases when the number of such scenarios reaches millions (i.e. simulation of knockouts). We address this challenge by developing a flexible and scalable system leveraging MS Azure Cloud capabilities.

Figure 1-8 below presents the simplified snapshot of the resulted distributed architecture. The entire system is transparent to the user. Once the system is set up, no specific programming skills, such as Matlab capabilities, are required. User provides the input in the HTML and/or XLS (i.e. CSV) format and, after the simulation ends, receives the results in the HTML and CSV formats.

On the background we first place a request on the entrance server, which can run either locally (as in our implementation) or remotely. It verifies the request correctness and generates setup-specific requests per each desired fermentation configuration (selected pipe of organisms, media and production target) to Azure Cloud side via dedicated request submission queue (setup-requestq). This stage is parallelized into several instances on Azure Cloud side to support simultaneous construction of multiple fermentation setups. The resulting model, describing requested fermentation setup, is placed in a dedicated blob (file storage entity) in the Azure Cloud and the local server is notified by dedicated notification queue (setup-resultq).

On receiving such notification, the local server generates a set of tasks specific to each fermentation setup, which may involve slight modifications of the resulted model. Specifically, it may include an update in reaction boundaries for several model reactions, thus giving means to media gradient and sensitivity analyses (see 1.3.6); to estimation within same setup of various oxygen amounts (see 1.3.7); and to simulation of sets of reaction knockouts (see 1.3.8). The generated set of single tasks is sent to Azure Cloud side for the evaluation via a dedicated queue (task-requestq). This stage is also parallelized into several instances to support tasks multiplicity. The details of estimation results (i.e. flux values of specific reactions) are placed in a dedicated result blob and the
local server is notified with result summary via dedicated notification queue (task-resultq).

Running the media sensitivity and gradient analyses generates few tens of tasks, which is a typical number of media components. Simultaneous estimation of various fermentation scenarios for multiple input media might generate hundreds or even few thousands of tasks. However, number of tasks to be estimated during the analysis of reaction knockouts may reach millions, since the number of reactions in a single organism is few thousands and we provide an option to estimate the impact of simultaneous knockouts in both organisms. To address such computationally challenging estimation problems, the BioLego architecture on the local server side includes few additional components dedicated to handling large-scale tasks. Specifically, we parallelize both the sending of task calculation requests and the processing of the received single-task results.

The entire simulation framework is implemented in Perl for management of model and media files, Matlab for FBA formulations and GNU Linear Programming Kit (GLPK - GNU Project - Free Software Foundation (FSF)) for solving resulting linear programming problems.
Runtime experiments and the quality of computational predictions are addressed later in 1.4.5.

All the code of BioLego system, including local server side, Microsoft Azure Cloud side and Microsoft Azure persistent content together with an installation guide is freely available for download at:

http://wassist.cs.technion.ac.il/~edwardv/STORAGE/biolego2_data/

Figure 1-8: Architecture of MS Azure Cloud-based BioLego platform
Section 1.4. Experiments, Applications and Tools

This section briefly describes published experimental work related to BioLego

1.4.1. Proposed design of distributed macroalgal biorefineries: Thermodynamics, bioconversion technology, and sustainability implications for developing economies.

In this study we presented the advantages of serial two-step methodology using the example of modeling green macroalgae *U. lactuca* bioconversion into bioethanol with metabolic models. In the first step, we simulated the ethanol fermentation from the decomposed *U. lactuca* biomass by *S. cerevisiae*. In the second step, fermentation leftovers and the *S. cerevisiae* biomass resulted from the first step are fermented by *E. coli* to produce additional ethanol (see Figure 1-9).

This study was published in BioFPR journal in 2014 *(Golberg, Vitkin et al., Biofuels, Bioproducts and Biorefining 2014)*

Macroalgae as an energy crop

The proper choice of raw biomass material is critical to ensure the efficient production of transportation biofuels. First generation feedstocks include sugar cane, sugar beets, starch-bearing grains, and conventional vegetable oil crops; while first generation fuel products include ethanol and biodiesel (Tanaka, 2011). Second and third generation biofuel technologies, currently in research and development, utilize animal fat, lignocellulosic biomass, and algae feedstocks; and produce hydrotreated vegetable oil, cellulosic-ethanol, biomass-to-liquids (BtL)-diesel, bio-butanol, and advanced drop-in replacement fuels such as fatty-acid ethyl esters, alkanes, alkenes, terpenes and methyl ketones (Tanaka, 2011; Keasling and Chou, 2008; Wargacki *et al.*, 2012; Bokinsky *et al.*, 2011; Steen *et al.*, 2010; Peralta-Yahya *et al.*, 2011). A large portfolio of thermochemical and biological conversion technologies is under development in several large-scale biomass programs. Despite the promise of biofuel to satisfy a significant portion of the future demand for transportation fuel, the perceived competition between “energy crops”
and “food crops” for land and water resources is a growing concern (Tanaka, 2011; WTO, 2011). Furthermore, the extent to which land erosion, potable water consumption, fertilizers, pesticides, biodiversity and climate change impact biofuel sustainability have yet to be evaluated (Pimentel, D., 2008).

Concerns over a net energy balance, a land and potable water use, and environmental hazards, question the sustainability of a corn/sugar cane biofuel future (Tanaka, 2011; Soetaert Wim, 2011; Pandey et al., 2011; WTO, 2011; Pimentel, D., 2008; Varghese, 2007; Gerbens-Leenes et al.). While lignocellulosic feedstock alternatives do not directly compete with the food supply, they may not be able to fully address the concerns above. In many locations around the world, arable land and potable water are scarce, but fuel demand exists. Furthermore, the technology required to release and ferment sugars from lignocellulose is still challenging (Brown and Brown, 2013; Stephanopoulos, 2007; Sanderson, 2011; Limayem and Ricke, 2012).

Macroalgae, which contain very little lignin and do not compete with food crops for arable land or potable water, are promising candidates for future transportation fuel feedstocks (Wargacki et al., 2012; Roesijadi et al., 2010; van der Wal et al., 2013; Bruhn et al., 2011; Kraan, 2013). Both developed and developing countries have recently reported efficient conversion of macroalgae to transportation fuels (Wargacki et al., 2012; Kambhaty et al., 2012). Various green, red and brown macroalgae species are under evaluation for inland and off shore cultivation (Wargacki et al., 2012; Roesijadi et al., 2010; Potts et al., 2012). The macroalgae from Ulva sp. is of particular interest due to its fast growth rates (Bruhn et al., 2011; Yantovski E, 2008) and high carbohydrate content. For example, Yantovski (Yantovski E, 2008) has proposed land grown Ulva for the production of 100% of electricity supply in Israel. Nevertheless, current technological immaturity at large-scale (~100 km² of cultivation are required per GW of power) limit the immediate implementation of this program.

Limitations of macroalgal biofuels feedstock are the availability of enzymes to decompose specific macroalgae polysaccharides (such as ulvan), as well as the lack of tractable microorganisms that can efficiently convert the resulting mono-saccharides into
biofuels. Chemical composition of Ulva sp. introduces additional challenge for its fermentation. Due to high carbohydrate content, theoretically up to 60% of the dry biomass in Ulva sp. could be fermented into ethanol (Nikolaisen, 2010). However, xylose and rhamnose, resulting from hemicellulose and ulvan decomposition, often comprise up to 29% of biomass (Marc Lahaye* and Audrey Robic†, 2007). The yeast S. cerevisiae, one of the most promising organisms for biomass-derived biofuel production, poorly utilizes these sugars. Current reported ethanol yields for S. cerevisiae Ulva bioconversion are only ~14% of biomass dry weight (dw) (Nikolaisen, 2010). To make macroalgae Ulva competitive with lignocellulosic ethanol production, average yields should be increased to 20-25% (Tao et al., 2011). The multiple economic, cultivation, and environmental advantages of Ulva feedstock (listed above) provided improvements in biomass processing technology, would make Ulva sp. a preferable biofuel feedstock, especially in coastal regions. In coastal areas, climatic conditions may be suitable for offshore Ulva cultivation, while the production of lignocellulosic plant biomass may be limited due to limited land and fresh water resources.

**Modelling U. lactuca fermentation setup**

Three possible scenarios of U. lactuca bioconversion into ethanol were tested Table 1-3. Analysis of these three models provides estimates of:

- Carbon uptake, defined as percentage of U. lactuca biomass carbons digested by organism.
- Expected emission of CO2 emission, defined as percentage of U. lactuca biomass carbons.
- Carbon utilization yields (CUY), defined as ratio of number of carbons in produced with the number of carbons in the biomass of U. lactuca.
- Wasted energy during the bioconversion process, as defined in Equation 1-19, where $m_{\text{ethanol}}$ is the mass of ethanol produced, $m_{\text{Ulva}}$ is the mass of Ulva consumed, and $H$ is the maximum heat of combustion, which is $19kJ \, g^{-1}$ for Ulva (Yantovski E, 2008) and $30kJ \, g^{-1}$ for ethanol (Domalski, 1972).
Equation 1-19: Wasted Energy

\[
Energy_{\text{waste}} = 1 - \frac{m_{\text{Ethanol}} \cdot H_{\text{Ethanol}}}{m_{\text{Ulva}} \cdot H_{\text{Ulva}}}
\]

Figure 1-9: Schematic representation of the serial two-step bioconversion of Ulva feedstock into bioethanol. Based on Model 3, Table 1-3.

All the simulations were performed under anaerobic conditions assuming a 1g of \textit{Ulva lactuca} uptake for 1g dry weight of organism in 1hour. Both Model 2 and Model 3 are two-step models (Figure 1-9) that assume that \textit{E. coli} digests all \textit{U. lactuca} remnants, such as rhamnose, together with all \textit{S. cerevisiae} biomass.
Model 1 consists of a one-step fermentation by wild type *S. cerevisiae*, with no ulvan decomposition or xylose uptake by *S. cerevisiae*. This model predicts ethanol yields of 15.6–19.9% of dry weight (dw), which is comparable to the experimentally-determined 14% yield (Nikolaisen, 2010).

Model 2 consists of a two-step fermentation by *S. cerevisiae* and then by *E. coli*, with no ulvan decomposition or xylose uptake by *S. cerevisiae*. This approach predicts significantly increase in total ethanol yield to 22.8 - 27.2 %dw. This advancement would make macroalgae *Ulva* competitive with lignocellulosic ethanol production, which have average yields of about 23-25% (Singh *et al.*, 2010).

Model 3 consists of a two-step fermentation combined with full ulvan polysaccharide degradation and xylose utilization by *S. cerevisiae*. Currently, there is neither commercially available technology for full ulvan decomposition and nor *S. cerevisiae* strain capable of efficiently utilizing xylose. Thus, the technology required for Model 3 has yet to be realized. Here it presents a theoretical maximal potential of *Ulva* bioconversion process for comparative purposes.

---

Table 1-3: Fermentation scenarios for production of ethanol based on *U. lactuca*. Column *MAX YIELD* refersto full ulvan decomposition and the ability of *S. cerevisiae* to uptake xylose

<table>
<thead>
<tr>
<th>Model #</th>
<th>Step 1 <em>S. cerevisiae</em></th>
<th>Step 2 <em>E. coli</em></th>
<th>MAX YIELD*</th>
<th>Ethanol Production (min-max) [%dw]</th>
<th>Carbon utilization (min-max) [%]</th>
<th>Wasted Energy (range)</th>
<th>Carbon Uptake [%C]</th>
<th>Emission of CO2 [%C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>15.6 – 19.9</td>
<td>19.00 – 24.23</td>
<td>0.754 – 0.686</td>
<td>40.25</td>
<td>11.28</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>22.8 – 27.2</td>
<td>27.77 – 33.12</td>
<td>0.640 – 0.571</td>
<td>68.17</td>
<td>16.59</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>31.0 – 34.8</td>
<td>37.75 – 42.38</td>
<td>0.511 – 0.451</td>
<td>83.55</td>
<td>22.24</td>
</tr>
</tbody>
</table>

*Technion - Computer Science Department - Ph.D. Thesis PHD-2018-12 - 2018*
1.4.2. BioLEGO — a web-based application for biorefinery design and evaluation of serial biomass fermentation.

In this study we originally presented the initial version of BioLego webservice together with the mathematical formulation of serial (two-step) fermentation process. This paper addresses the details of all the mathematical modeling aspects of the two-step fermentation process as relevant for processes that require tight control. The implementation of the serial two-step case is of particular interest, since it provides practical perspectives on the general serial fermentation process drawbacks and advantages.

This study was published in the journal Technology in 2015 (Vitkin et al., Technology 2015)

Here we summarize the experiments performed in this study. All the implementation details were covered in Section 1.2 and Section 1.3.

Simulation of Corn Cobs single-step fermentation

We compared corn cobs single-step fermentation efficiency between two different yeast species – wild type (WT) and recombinant (RN1016), which has a *Piromyces xylose-isomerases* (Harhangi et al., 2003) genome facilitating the fermentation of both glucose and xylose to ethanol. Moreover, we tested the potential of slight corn cobs biomass changes to enhance this process using media gradient analysis (Equation 1-17).

<table>
<thead>
<tr>
<th>S. cerevisiae strain</th>
<th>Growth Rate [h⁻¹]</th>
<th>Ethanol Production (min-max) [g/Kg]</th>
<th>Carbon utilization (min-max) [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>0.0498</td>
<td>132.5 – 164.8</td>
<td>20.9 – 26.0</td>
</tr>
<tr>
<td>RN1016</td>
<td>0.0642</td>
<td>268.7 – 300.8</td>
<td>42.4 – 47.5</td>
</tr>
</tbody>
</table>

Table 1-4: BioLego efficiency simulation results for corn cobs to ethanol single-step fermentation with different *S. cerevisiae* strains. RN1016 strain can digest xylose.

The 3 repetitions of simulation for both species took on average 5.54 minutes on the Technion CSM server (6 cores of 2.53GHz, 8Gb total memory). Results (Table 1-4) show that WT yeast should produce 132.5-164.8g ethanol per 1kg of corn cobs (g/Kg). These
estimations are well-correlated with the experimental results demonstrated by van Eylen et al (van Eylen et al., 2011), who showed yield of 158g/Kg. On the other hand, for the RN1016 yeast we predict production rate of 268.7-300.8g/Kg, which is significantly higher than the reported 197g/Kg (van Eylen et al., 2011). This disagreement is expected because the assumption of highly efficient xylose digestion efficiency by *S. cerevisiae* is an open challenge (van Maris et al., 2006).

The gradient analysis of corn cob biomass composition yielded some interesting results. As can be expected, our simulations suggest decreasing the relative weight of all non-digested particles, while increasing glucose or glucose & xylose relative weights for WT and for RN1016 yeasts respectively. Interestingly, system suggestions for maximizing minimal versus maximal ethanol productions differ for several amino acids. For example, for maximizing minimal ethanol production our system proposes to increase the proportion of valine, lysine and isoleucine, while it is unnecessary for maximizing maximal ethanol production.

**Simulation of mixed media two-step fermentation**

We also evaluated the efficiency of two-step fermentation process for 2:1 mix of *U. lactuca* and *K. alvarezii* algal biomasses by *E. coli* and WT *S. cerevisiae*. We estimated both two-step configuration directions, i.e. *E. coli* followed by *S. cerevisiae* and vice versa. In addition, we compared the efficiency of two-step fermentation process to the single step fermentation by each organism separately and by the organism co-culture (Figure 1-5 and Table 1-2).
### Table 1-5: Efficiency of fermentation process for 2:1 mix of *U. lactuca* together with *K. alvarezii* biomasses by *E. coli* and WT *S. cerevisiae*.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Growth Rate <em>E. coli</em> [h⁻¹]</th>
<th>Growth Rate <em>S. cerevisiae</em> [h⁻¹]</th>
<th>Ethanol Production (min-max) [g/Kg]</th>
<th>Carbon utilization (min-max) [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT <em>S. cerevisiae</em></td>
<td>–</td>
<td>0.0469</td>
<td>142.4 – 148.1</td>
<td>20.7 – 21.6</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>0.1016</td>
<td>–</td>
<td>95.5 – 95.5</td>
<td>13.9 – 13.9</td>
</tr>
<tr>
<td>WT <em>S. cerevisiae</em> followed by <em>E. coli</em></td>
<td>0.0501</td>
<td>0.0469</td>
<td>168.4 – 170.8</td>
<td>24.5 – 24.9</td>
</tr>
<tr>
<td><em>E. coli</em> followed by WT <em>S. cerevisiae</em></td>
<td>0.1016</td>
<td>0.0001</td>
<td>95.5 – 99.8</td>
<td>13.9 – 14.5</td>
</tr>
<tr>
<td>WT <em>S. cerevisiae</em> together with <em>E. coli</em></td>
<td>0.1081</td>
<td>0.0</td>
<td>90.4 – 97.7</td>
<td>13.2 – 14.2</td>
</tr>
</tbody>
</table>

The 3 repetitions of simulation for both species at this version of BioLego for all 5 configurations took on average 31 sec on the Technion CSM server (6 cores of 2.53GHz, 8Gb total memory). Simulation results (Table 1-5) predict that two-step fermentation starting with WT *S. cerevisiae* should produce 168.4-170.8g ethanol per 1kg of media mix, which has clear advantage over all other scenarios, i.e. ethanol production yield of 142.4-148.1g of ethanol for WT *S. cerevisiae* alone or around 90-100g/Kg of ethanol for *E. coli* consisting scenarios. Note that last includes evaluation of the co-culturing setup estimation, predicting 90.4-97.7g ethanol per 1kg of media mix.
1.4.3. Thermochemical hydrolysis of green macroalgae *Ulva* feedstock for biorefinery: Taguchi orthogonal arrays experimental design approach for physical and chemical parameters optimization

The purpose of this study was to understand how various parameters, such as temperature, time, acid and solid percentage, influence the efficiency of *U. lactuca* cell wall deconstruction. This efficiency was estimated by the resulting amounts of extracted monosaccharides together with estimates of the potential production of ethanol, butanol and acetone resulted from their fermentations as predicted by BioLego system.

This study was published in Scientific Reports (Nature Publishing Group) in 2016 *(Jiang, Linzon, Vitkin et al., Scientific Reports 2016)*

**Study overview**

In the first stage, quantities of major carbohydrate components of the dried *U. lactuca* biomass were quantified for 2 repetitions of 16 different hydrolysis conditions. The exact hydrolysis conditions were defined based on Taguchi orthogonal arrays technique *(Rao et al., 2008)* for hydrolysis temperature (ranging 100-134°C), hydrolysis time (ranging 30-60min), acid percentage (ranging 0.0-5.0%) and the percentage of solid part (ranging 5-25%), as presented in Table 1-6. Specifically, we quantified the amounts of Rhamnose, of Glucose, of Xylose and of Glucuronic acid (Table 1-7).

<table>
<thead>
<tr>
<th>Condition #</th>
<th>T [℃]</th>
<th>Time [min]</th>
<th>Acid [%]</th>
<th>Solid [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>30</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>45</td>
<td>0.5</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>60</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>45</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>121</td>
<td>30</td>
<td>0.5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>121</td>
<td>45</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>121</td>
<td>60</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>121</td>
<td>30</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>134</td>
<td>30</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>134</td>
<td>45</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>11</td>
<td>134</td>
<td>60</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>134</td>
<td>60</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>134</td>
<td>30</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>121</td>
<td>45</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>100</td>
<td>60</td>
<td>0.5</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>134</td>
<td>45</td>
<td>0</td>
<td>25</td>
</tr>
</tbody>
</table>

*Table 1-6: Taguchi L16 orthogonal array for determination of optimum process parameters for thermochemical deconstruction of macroalgae *Ulva* biomass.*
### Table 1-7: Carbohydrate components of *U. lactuca* averaged over 2 repetitions of the Taguchi L16 orthogonal array as defined in Table 1-6

<table>
<thead>
<tr>
<th>Condition #</th>
<th>Rhamnose [gr/kg]</th>
<th>Glucose [gr/kg]</th>
<th>Xylose [gr/kg]</th>
<th>Glucuronic acid [gr/kg]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.81</td>
<td>0.01</td>
<td>0.02</td>
<td>12.88</td>
</tr>
<tr>
<td>2</td>
<td>4.12</td>
<td>0.06</td>
<td>0.10</td>
<td>29.70</td>
</tr>
<tr>
<td>3</td>
<td>32.79</td>
<td>70.62</td>
<td>14.47</td>
<td>172.91</td>
</tr>
<tr>
<td>4</td>
<td>63.67</td>
<td>76.81</td>
<td>8.08</td>
<td>57.18</td>
</tr>
<tr>
<td>5</td>
<td>0.53</td>
<td>0.05</td>
<td>0.05</td>
<td>10.18</td>
</tr>
<tr>
<td>6</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>5.93</td>
</tr>
<tr>
<td>7</td>
<td>65.65</td>
<td>75.02</td>
<td>4.48</td>
<td>48.03</td>
</tr>
<tr>
<td>8</td>
<td>67.31</td>
<td>93.53</td>
<td>12.16</td>
<td>51.60</td>
</tr>
<tr>
<td>9</td>
<td>74.30</td>
<td>87.89</td>
<td>11.39</td>
<td>41.98</td>
</tr>
<tr>
<td>10</td>
<td>56.36</td>
<td>82.15</td>
<td>3.24</td>
<td>30.56</td>
</tr>
<tr>
<td>11</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>3.81</td>
</tr>
<tr>
<td>12</td>
<td>52.71</td>
<td>63.31</td>
<td>6.94</td>
<td>40.20</td>
</tr>
<tr>
<td>13</td>
<td>63.19</td>
<td>84.44</td>
<td>3.77</td>
<td>38.17</td>
</tr>
<tr>
<td>14</td>
<td>59.92</td>
<td>67.78</td>
<td>7.68</td>
<td>45.16</td>
</tr>
<tr>
<td>15</td>
<td>0.19</td>
<td>0.01</td>
<td>0.01</td>
<td>8.21</td>
</tr>
<tr>
<td>16</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>1.36</td>
</tr>
</tbody>
</table>

In the second stage, we used the BioLego framework to predict the minimal and maximal ethanol production rates based on the fermentation of the received 16 media (together with *U. lactuca* media without any carbohydrates) as achieved from fermentation (i) with *E. coli*; (ii) with *S. cerevisiae WT*; (iii) with *S. cerevisiae RN1016* (yeast with ability of xylose digestion); and (iv) with *C. acetobutylicum*. In addition, minimal and maximal acetate and butanol production rates were evaluated for *C. acetobutylicum*.

**Simulation results**

For *E. coli*, the maximum predicted production rate of ethanol was found in experiments 8 and 9 (19.459-20.833 [gr ethanol kg\(^{-1}\) *Ulva DW*]) (Table 1-8). This corresponds to the highest extraction yields of glucose, rhamnose and xylose with thermochemical hydrolysis.

For *S. cerevisiae WT* the maximum predicted ethanol rates are 49.323-51.929 [gr ethanol kg\(^{-1}\) *Ulva DW*] were also observed for fermentation simulation of hydrolysates from experiments 8 and 9 (Table 1-9). This corresponds to the highest extraction yields of glucose, rhamnose and xylose with thermochemical hydrolysis. Previous experimental studies reported on 62 gr ethanol production per kg of *Ulva pertusa* after enzymatic hydrolyses (Yanagisawa *et al.*, 2013, 2011).
<table>
<thead>
<tr>
<th>Condition #</th>
<th>E. coli growth rate (h⁻¹)</th>
<th>Min Ethanol (g/Kg)</th>
<th>Max Ethanol (g/Kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.051</td>
<td>13.094</td>
<td>14.294</td>
</tr>
<tr>
<td>4</td>
<td>0.034</td>
<td>13.757</td>
<td>15.955</td>
</tr>
<tr>
<td>5</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>0.032</td>
<td>12.531</td>
<td>14.879</td>
</tr>
<tr>
<td>8</td>
<td>0.038</td>
<td>18.883</td>
<td>20.833</td>
</tr>
<tr>
<td>9</td>
<td>0.036</td>
<td>17.358</td>
<td>19.459</td>
</tr>
<tr>
<td>10</td>
<td>0.029</td>
<td>14.319</td>
<td>16.798</td>
</tr>
<tr>
<td>11</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>12</td>
<td>0.027</td>
<td>10.605</td>
<td>13.151</td>
</tr>
<tr>
<td>13</td>
<td>0.031</td>
<td>14.819</td>
<td>17.158</td>
</tr>
<tr>
<td>14</td>
<td>0.035</td>
<td>11.631</td>
<td>14.119</td>
</tr>
<tr>
<td>15</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Hydrolysate free medium</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 1-8: Predicted ethanol yields using Ulva hydrolysate fermentation by E. coli

<table>
<thead>
<tr>
<th>Condition #</th>
<th>S. cerevisiae WT growth rate (h⁻¹)</th>
<th>Min Ethanol (g/Kg-1)</th>
<th>Max Ethanol (g/Kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>8.753</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
<td>8.774</td>
</tr>
<tr>
<td>3</td>
<td>0.011</td>
<td>33.205</td>
<td>41.349</td>
</tr>
<tr>
<td>4</td>
<td>0.012</td>
<td>36.149</td>
<td>44.209</td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>0.000</td>
<td>8.769</td>
</tr>
<tr>
<td>6</td>
<td>0.000</td>
<td>0.000</td>
<td>8.746</td>
</tr>
<tr>
<td>7</td>
<td>0.012</td>
<td>35.297</td>
<td>43.381</td>
</tr>
<tr>
<td>8</td>
<td>0.014</td>
<td>44.107</td>
<td>51.929</td>
</tr>
<tr>
<td>9</td>
<td>0.014</td>
<td>41.421</td>
<td>49.323</td>
</tr>
<tr>
<td>10</td>
<td>0.013</td>
<td>38.693</td>
<td>46.674</td>
</tr>
<tr>
<td>11</td>
<td>0.000</td>
<td>0.000</td>
<td>8.746</td>
</tr>
<tr>
<td>12</td>
<td>0.010</td>
<td>29.724</td>
<td>37.974</td>
</tr>
<tr>
<td>13</td>
<td>0.013</td>
<td>39.780</td>
<td>47.731</td>
</tr>
<tr>
<td>14</td>
<td>0.011</td>
<td>31.852</td>
<td>40.038</td>
</tr>
<tr>
<td>15</td>
<td>0.000</td>
<td>0.000</td>
<td>8.751</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>0.000</td>
<td>8.746</td>
</tr>
<tr>
<td>Hydrolysate free medium</td>
<td>0.000</td>
<td>0.000</td>
<td>8.746</td>
</tr>
</tbody>
</table>

Table 1-9: Predicted ethanol yields using Ulva hydrolysate fermentation by S. cerevisiae WT

Among all simulations, the maximum production of ethanol was achieved with Ulva fermentation using S. cerevisiae RN1016 (which model was constructed by inclusion of Xylose isomerase-related set of reactions into Yeast5 model of S. cerevisiae WT (Heavner et al., 2012)) (Table 1-10). For this organism, in simulations 1, 2, 5, 6, 11, 15 and 16 the predicted ethanol production rates were the same as for media without any monosaccharides (0–8.746 g kg⁻¹) derived from hydrolysates. For other simulations we
observed an increase in minimal ethanol production rates from 0.00 up to 54.134 g ethanol kg\(^{-1}\) \textit{Ulva} (in exp. 8) and an increase in maximal ethanol production rate from 8.746 g ethanol kg\(^{-1}\) \textit{Ulva} up to 57.500 g ethanol kg\(^{-1}\) \textit{Ulva} (in exp. 8). The achieved predicted results seem reasonable, since the major carbohydrates source for ethanol is glucose, reaching maximal concentration of 93.53 g glucose kg\(^{-1}\) \textit{Ulva} in exp. 8 (Table 1-7). Also, these findings are supported in previous work (Nikolaisen \textit{et al.}, 2011), which predicted 160g of glucose per kg of \textit{Ulva}, with glucose concentration of 310g ethanol kg\(^{-1}\) \textit{Ulva} (keeping similar glucose-to-ethanol mass ratio of 2:1).

<table>
<thead>
<tr>
<th>Condition #</th>
<th>\textit{S. cerevisiae RN1016} growth rate (h(^{-1}))</th>
<th>Min Ethanol (g/Kg)</th>
<th>Max Ethanol (g/Kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>8.762</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
<td>8.820</td>
</tr>
<tr>
<td>3</td>
<td>0.013</td>
<td>44.609</td>
<td>47.890</td>
</tr>
<tr>
<td>4</td>
<td>0.013</td>
<td>43.441</td>
<td>47.741</td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>0.000</td>
<td>8.792</td>
</tr>
<tr>
<td>6</td>
<td>0.000</td>
<td>0.000</td>
<td>8.746</td>
</tr>
<tr>
<td>7</td>
<td>0.012</td>
<td>38.909</td>
<td>45.448</td>
</tr>
<tr>
<td>8</td>
<td>0.016</td>
<td>54.134</td>
<td>57.500</td>
</tr>
<tr>
<td>9</td>
<td>0.015</td>
<td>51.063</td>
<td>54.450</td>
</tr>
<tr>
<td>10</td>
<td>0.013</td>
<td>40.960</td>
<td>48.174</td>
</tr>
<tr>
<td>11</td>
<td>0.000</td>
<td>0.000</td>
<td>8.746</td>
</tr>
<tr>
<td>12</td>
<td>0.011</td>
<td>35.896</td>
<td>41.096</td>
</tr>
<tr>
<td>13</td>
<td>0.014</td>
<td>42.621</td>
<td>49.473</td>
</tr>
<tr>
<td>14</td>
<td>0.012</td>
<td>38.753</td>
<td>43.428</td>
</tr>
<tr>
<td>15</td>
<td>0.000</td>
<td>0.000</td>
<td>8.755</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>0.000</td>
<td>8.746</td>
</tr>
<tr>
<td>Hydrolysate free medium</td>
<td>0.000</td>
<td>0.000</td>
<td>8.746</td>
</tr>
</tbody>
</table>

Table 1-10: Predicted ethanol yields using \textit{Ulva} hydrolysate fermentation by \textit{S. cerevisiae RN1016}

For \textit{C. acetobutylicum}, in simulations with growth rates not close to zero, the nonzero ABE production rate (defined as sum over acetone, butanol and ethanol predicted production rates) varied from 1.08 to 2.91 gr ABE kg\(^{-1}\) \textit{Ulva} (Table 1-11). The ABE predicted yield (calculated as ratio of the sum of ABE products to total sugars extracted with hydrolysis) of predicted ABE fermentation of \textit{Ulva} hydrolysate by \textit{C. acetobutylicum} was in 0.02-0.17 range. In the reported experimental ABE, yield of \textit{Ulva lactuca} hydrolysate fermentation by \textit{C. acetobutylicum} was in the 0.03-0.32 range for hydrolysates with various supplements such as glucose, xylose and nutrients as in CM2 medium and 0.08 for the hydrolysate without supplements (van der Wal \textit{et al.}, 2013). In our simulations, the maximum predicted ABE yield is 23.955-26.830 gr ABE kg\(^{-1}\) \textit{Ulva}.
However, these results were shown for experiments 2, 5 and 15, where the growth rates of the organisms were close to 0.00 and these results should therefore be treated with care, as they depend on the organism survival at this specific medium. Specifically, in our simulations (Equation 1-3, Equation 1-4) we first maximize organism growth rate, utilizing all available media components in favor of this task and, only then, estimate the molecule production range. This is the reason that media with more sugars does not necessarily lead to higher ABE production but rather to higher organism growth rates. On the other hand, in cases when the amount of monosaccharides (specifically, glucose) is very low, it is a limiting factor to the *C. acetobutylicum* growth, leading to near-ε (Equation 1-3) growth rates (like in exp. 2, Table 1-11). As a result, the amount of media components not utilized by biomass-constructing reaction (v<sub>Growth</sub> in Equation 1-3) is relatively high and the simulations predict high target molecule production estimation in such low growth rate scenarios.

<table>
<thead>
<tr>
<th>Condition #</th>
<th><em>C. acetobutylicum</em> growth rate (h&lt;sup&gt;-1&lt;/sup&gt;)</th>
<th>Min Ethanol (g/Kg)</th>
<th>Max Ethanol (g/Kg)</th>
<th>Min Acetone (g/Kg)</th>
<th>Max Acetone (g/Kg)</th>
<th>Min Butanol (g/Kg)</th>
<th>Max Butanol (g/Kg)</th>
<th>Min ABE (g/Kg)</th>
<th>Max ABE (g/Kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.00</td>
<td>1.72</td>
<td>0.00</td>
<td>2.18</td>
<td>0.00</td>
<td>8.28</td>
<td>0.00</td>
<td>12.18</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.00</td>
<td>3.37</td>
<td>0.00</td>
<td>4.25</td>
<td>0.00</td>
<td>16.34</td>
<td>0.00</td>
<td>23.96</td>
</tr>
<tr>
<td>3</td>
<td>0.079</td>
<td>1.35</td>
<td>1.35</td>
<td>0.28</td>
<td>0.28</td>
<td>0.00</td>
<td>1.63</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.085</td>
<td>1.68</td>
<td>1.68</td>
<td>0.30</td>
<td>0.30</td>
<td>0.00</td>
<td>1.97</td>
<td>1.97</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>0.00</td>
<td>6.07</td>
<td>0.00</td>
<td>4.30</td>
<td>0.00</td>
<td>16.46</td>
<td>0.00</td>
<td>26.83</td>
</tr>
<tr>
<td>6</td>
<td>0.000</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.083</td>
<td>1.58</td>
<td>1.58</td>
<td>0.30</td>
<td>0.30</td>
<td>0.00</td>
<td>1.88</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.099</td>
<td>2.56</td>
<td>2.56</td>
<td>0.35</td>
<td>0.35</td>
<td>0.00</td>
<td>2.91</td>
<td>2.91</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.094</td>
<td>2.26</td>
<td>2.26</td>
<td>0.33</td>
<td>0.33</td>
<td>0.00</td>
<td>2.59</td>
<td>2.59</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.089</td>
<td>1.96</td>
<td>1.96</td>
<td>0.32</td>
<td>0.32</td>
<td>0.00</td>
<td>2.27</td>
<td>2.27</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.000</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.073</td>
<td>0.94</td>
<td>0.94</td>
<td>0.14</td>
<td>0.55</td>
<td>0.00</td>
<td>1.08</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.091</td>
<td>2.08</td>
<td>2.08</td>
<td>0.33</td>
<td>0.33</td>
<td>0.00</td>
<td>2.40</td>
<td>2.40</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.077</td>
<td>1.20</td>
<td>1.20</td>
<td>0.24</td>
<td>0.35</td>
<td>0.00</td>
<td>1.44</td>
<td>1.61</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.000</td>
<td>0.00</td>
<td>3.43</td>
<td>0.00</td>
<td>4.34</td>
<td>0.00</td>
<td>16.55</td>
<td>0.00</td>
<td>24.32</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Hydrolysate free medium</td>
<td>0.000</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

Table 1-11: Predicted ABE (acetone, butanol, ethanol) yields using *Ulva* hydrolysates fermentation by *C. acetobutylicum*

The goal of this work was to develop models to for optimization of biorefinery based on macroalgae K. alvarezii, which produces carrageenan, ethanol, fertilizer and biogas for the local coastal communities in developing countries. We specifically focus on the species *Kappaphycus alvarezii*, which is of particular interest as it is currently one the most widely cultivated macroalgae species (Valderrama et al., 2013).

This study was published in BioEnergy Research journal in 2017 (Golberg, Vitkin et al., ECOS 2015; Ingle, Vitkin et al., BioEnergy Research 2017).

**Kappaphycus alvarezii**

*K. alvarezii* species shows relatively higher growth rate compared to other *Kappaphycus* macroalgae with biomass yield ranging from 12-45 dry tons ha$^{-1}$ year$^{-1}$ (Tee et al., 2015; Subba Rao and Mantri, 2006). Carrageenan derived from this macroalgae makes *K. alvarezii* highly valuable so its cultivation provides jobs and commercial opportunities to numerous poor communities in coastal areas of developing countries (Sahu et al., 2011). After carrageenan extraction, approximately 60 to 70% resultant solid fraction is considered today as waste (Uju et al., 2015). However, this waste contains high concentrations of carbohydrates, which can be hydrolyzed to monosaccharides and then converted into biofuels (Lee et al., 2016; Khambhaty et al., 2012; Hargreaves et al., 2013). In addition, the production of liquid biofertilizer (sap) from fresh *K. alvarezii* prior to drying and processing for carrageenan production has been reported (Eswaran et al., 2005). This biofertilizer has proven to have many benefits on local crop yield and resistance and is easy to produce and use allowing the consideration of such process for rural applications.

Macroalgae derived biomass mainly consists of high amounts of various polysaccharides, like cellulose and carrageenan in *K. alvarezii* (Lechat et al., 1997). Other prevailing molecules are carbohydrates (up to 27%), amino acids (16%) and fatty acids (1%)
(Fayaz et al., 2005). Table 1-12 details the *K. alvarezii* biomass composition used in simulations.

<table>
<thead>
<tr>
<th>Class</th>
<th>Component</th>
<th>Weight (g/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbohydrates</td>
<td>d-Galactose</td>
<td>189.56</td>
</tr>
<tr>
<td></td>
<td>3.6-anhydrogalactose</td>
<td>126.64</td>
</tr>
<tr>
<td></td>
<td>6-O-methyl galactose</td>
<td>9.01</td>
</tr>
<tr>
<td></td>
<td>d-Glucose</td>
<td>92.30</td>
</tr>
<tr>
<td></td>
<td>d-Xylose</td>
<td>5.09</td>
</tr>
<tr>
<td></td>
<td>d-Mannose</td>
<td>4.52</td>
</tr>
<tr>
<td></td>
<td>L-Arabinose</td>
<td>0.46</td>
</tr>
<tr>
<td>Amino Acids</td>
<td>Aspartic acid</td>
<td>5.65</td>
</tr>
<tr>
<td></td>
<td>Glutamic acid</td>
<td>11.20</td>
</tr>
<tr>
<td></td>
<td>Asparagine</td>
<td>10.45</td>
</tr>
<tr>
<td></td>
<td>Serine</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>Glycine</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>Threonine</td>
<td>5.07</td>
</tr>
<tr>
<td></td>
<td>Arginine</td>
<td>5.20</td>
</tr>
<tr>
<td></td>
<td>Alanine</td>
<td>7.06</td>
</tr>
<tr>
<td></td>
<td>Cystine</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>Tyrosine</td>
<td>3.42</td>
</tr>
<tr>
<td></td>
<td>Histidine</td>
<td>11.50</td>
</tr>
<tr>
<td></td>
<td>Valine</td>
<td>3.14</td>
</tr>
<tr>
<td></td>
<td>Iso-leucine</td>
<td>11.26</td>
</tr>
<tr>
<td></td>
<td>Phenyl alanine</td>
<td>14.54</td>
</tr>
<tr>
<td></td>
<td>Leucine</td>
<td>9.07</td>
</tr>
<tr>
<td></td>
<td>Lysine</td>
<td>20.75</td>
</tr>
<tr>
<td></td>
<td>Proline</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td>Tryptophan</td>
<td>11.10</td>
</tr>
<tr>
<td>Fats</td>
<td>myristic acid</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>myristoleic acid</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>pentadecanoic acid</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>palmitoleic acid</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>heptadecanoic acid</td>
<td>2.53</td>
</tr>
<tr>
<td></td>
<td>cis-heptadecanoic acid</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>oleic acid</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>linoleic acid</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>linolenic acid</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>arachidic acid</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>eicosanoic acid</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>eicosadienoic acid</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>arachidonic acid</td>
<td>0.10</td>
</tr>
<tr>
<td>Other</td>
<td>Sulfate-remainders</td>
<td>140.41</td>
</tr>
<tr>
<td></td>
<td>Ash</td>
<td>197.00</td>
</tr>
<tr>
<td></td>
<td>ascorbic acid</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td>vitamin A</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Protein-remainders</td>
<td>25.70</td>
</tr>
</tbody>
</table>

Table 1-12: *K. alvarezii* chemical composition used in the model as fermentation medium, based on (Fayaz et al., 2005; Rajasulochana et al., 2010; Lechat et al., 1997).
Fermentation analysis

We estimated the annual bioethanol production potential from *K. alvarezii* based on Equation 1-20, where \( BPP \) (\( L \ ha^{-1} \ y^{-1} \)) is the estimated annual bioethanol production yield; \( \text{yield}_{\text{Biomass}} \) (\( kg \ \text{DW} \ ha^{-1} \ y^{-1} \)) is an is an average seaweed biomass yield; and \( \text{Efficiency}_{\text{Biomass} \rightarrow \text{Ethanol}} \) (\( L \ \text{ethanol} \ kg \ \text{DW}^{-1} \)) is the conversion efficiency of dry seaweed into bioethanol as predicted by BioLego.

\[
\text{BPP} = \text{Yield}_{\text{Biomass}} \times \text{Efficiency}_{\text{Biomass} \rightarrow \text{Ethanol}}
\]

As demonstrated in Table 1-13, we predict maximal bioethanol production rate in two-step fermentation setup with *S. cerevisiae* as first organism in the process. In such setup, we expect \( \sim 70\% \) product increase comparing to *S. cerevisiae* alone.

<table>
<thead>
<tr>
<th>Biomass</th>
<th>Configuration</th>
<th>Growth Rate ( E. \ coli ) [( h^{-1} )]</th>
<th>Growth Rate ( S. \ cerevisiae ) [( h^{-1} )]</th>
<th>Ethanol Production (min-max) [g/Kg]</th>
<th>Carbon utilization (min-max) [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original ( K. alvarezii )</td>
<td>( WT \ S. \ cerevisiae )</td>
<td>–</td>
<td>0.02</td>
<td>49.4 – 57.2</td>
<td>9.9 – 11.4</td>
</tr>
<tr>
<td></td>
<td>( E. \ coli )</td>
<td>0.05</td>
<td>–</td>
<td>69.9 – 72.2</td>
<td>14.0 – 14.4</td>
</tr>
<tr>
<td>( K. alvarezii ) after</td>
<td>( WT \ S. \ cerevisiae )</td>
<td>0.04</td>
<td>0.02</td>
<td>94.1 – 97.6</td>
<td>18.8 – 19.5</td>
</tr>
<tr>
<td>carrageenan extraction</td>
<td>followed by ( E. \ coli )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( 120 \ g/kg )</td>
<td>( E. \ coli )</td>
<td>0.05</td>
<td>0.0001</td>
<td>69.9 – 74.5</td>
<td>14.0 – 14.9</td>
</tr>
<tr>
<td></td>
<td>followed by ( WT \ S. \ cerevisiae )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( K. alvarezii ) after</td>
<td>( WT \ S. \ cerevisiae )</td>
<td>–</td>
<td>0.02</td>
<td>49.4 – 57.2</td>
<td>12.5 – 14.5</td>
</tr>
<tr>
<td>carrageenan extraction</td>
<td>( E. \ coli )</td>
<td>0.04</td>
<td>–</td>
<td>52.9 – 56.4</td>
<td>13.4 – 14.3</td>
</tr>
<tr>
<td>( 120 \ g/kg )</td>
<td>( WT \ S. \ cerevisiae )</td>
<td>0.03</td>
<td>0.02</td>
<td>77.6 – 81.7</td>
<td>19.6 – 20.7</td>
</tr>
<tr>
<td></td>
<td>followed by ( E. \ coli )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>followed by ( WT \ S. \ cerevisiae )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1-13: Simulation results of fermentation of \( K. alvarezii \) biomass

The predicted results for \( K. alvarezii \) fermentation using two step fermentation (94.1-97.6 g ethanol/kg biomass) correspond well to the experimentally observed results (81.9 g of ethanol /kg dry biomass in which authors used a special *S. cerevisiae* strain which
ferments galactose (Hargreaves et al., 2013)). Also, we show the potential production of ethanol from the residual biomass after 12% DW was extracted as carrageenan. We show that in this case of co-production, up to 81.7 g of ethanol/kg biomass can be produced in addition to 120 g/kg of carrageenan (Table 1-13).

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Growth Rate E. coli [h⁻¹]</th>
<th>Growth Rate S. cerevisiae [h⁻¹]</th>
<th>Ethanol Production (min-max) [g/Kg]</th>
<th>Carbon utilization (min-max) [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibers 36% Protein 9%</td>
<td>0.0017</td>
<td>0.0168</td>
<td>103.1 – 105.1</td>
<td>21.1– 23.5</td>
</tr>
<tr>
<td>Fibers 27% Protein 18%</td>
<td>0.0014</td>
<td>0.0145</td>
<td>82.7 – 96.8</td>
<td>17.7– 20.8</td>
</tr>
<tr>
<td>Fibers 18% Protein 27%</td>
<td>0.0012</td>
<td>0.0121</td>
<td>63.3 – 85.9</td>
<td>13.0– 17.7</td>
</tr>
<tr>
<td>Fibers 9% Protein 36%</td>
<td>0.001</td>
<td>0.0098</td>
<td>44.2 – 74.95</td>
<td>8.8– 14.9</td>
</tr>
</tbody>
</table>

Table 1-14: Sensitivity analysis of fermentation of K. alvarezii biomass in a serial two-step fermentation scenario with S. cerevisiae followed by E.coli

Biochemical analysis of K. alvarezii biomass show that total fibers and protein content vary as a function of season and nutrients availability (Periyasamy et al., 2014; Suresh Kumar et al., 2014; de Góes and Reis, 2012; Hung et al., 2009; Bindu and Levine, 2011; Masarin et al., 2016). The increase in the fibers content is compensated by the decrease of the protein content (Periyasamy et al., 2014; Suresh Kumar et al., 2014; de Góes and Reis, 2012; Hung et al., 2009; Bindu and Levine, 2011; Masarin et al., 2016).

We performed sensitivity analysis of the ethanol production in serial two-step fermentation scenario with S. cerevisiae as first organism in the process as a function of biomass chemical composition (Table 1-14 and Figure 1-10). To this end we simulated the variation of the fiber/protein ratio, keeping their total content constant (45%). Increasing the fiber content from 9% to 36% (fiber/protein ratio from 0.25 to 4) increased the predicted growth rate of S. cerevisiae from 0.0098 h⁻¹ to 0.0168 h⁻¹, E. coli from 0.001 h⁻¹ to 0.0017 h⁻¹; ethanol yield from 44.2-74.9 g kg⁻¹ to 103.1-105.1 g kg⁻¹ and carbon utilization yield from 8.8-14.9% to 21.1-23.5%.
The potential saturation of the increase indicates the limitations imposed by other metabolites on growth and ethanol production at higher fiber content (Figure 1-10).

**Figure 1-10:** Sensitivity analysis of fibers and protein content variation, expected with nutrients availability and seasonal changes, impact on (a) *S. cerevisiae* and *E.coli* growth rates (b). Ethanol yield and (c) Carbon Utilization yield based on the two-step fermentation scenario with *S. cerevisiae* as first organism.
1.4.5. Distributed flux balance analysis simulations of serial two-organism biomass fermentation.

In this study we present the distributed version of BioLego webservice leveraging Microsoft Azure Cloud. This paper addresses the details of the proposed architecture created to support heavy-load calculations. Several experiments were performed to assess the computational performance of the created BioLego platform. All the experimental prediction results and other relevant materials are available at http://wassist.cs.technion.ac.il/~edwardv/STORAGE/biolego2_data/.

This study is currently in preparation for submission to Bioinformatics journal.

Table 1-15 summarizes the runtimes and the following sections provide the details of the experiments performed in this study. All the implementation details were previously covered in Section 1.2 and 0Section 1.3.

<table>
<thead>
<tr>
<th>#</th>
<th>Research Setup</th>
<th>Number of Tasks</th>
<th>Computational Setup</th>
<th>Runtime</th>
</tr>
</thead>
</table>
| a. | Exploratory search for optimal biomass utilization setups for distinct types of corn biomasses: 1. *Corn Cobs* 2. *Corn Fiber Corn Stover* | • 240 model configurations 480 single fermentation task simulations | • 10 X Model Setup Constructors  
  • Standard A1v2 nodes (1 core, 2048MB)  
  • 2 threads  
  • 10 X Single Fermentation Task Estimators  
  • Standard A1v2 nodes (1 core, 2048MB) 3 threads | 21min |
| b. | (1) Sensitivity and (2) Gradient analyses for anaerobic two-step ethanol production from *Kappaphycus alvarezzi* | 2 X 37 single fermentation task simulations | • 5 X Single Fermentation Task Estimators  
  • Standard A1v2 nodes (1 core, 2048MB) 3 threads | 2 X 4min |
| c. | Analysis of two-step fermentation with knock-outs in each organism: 1. *S. cerevisiae* (2,280 reactions)  
  2. *E. coli* (2,914 reactions) | 6,649,115 single fermentation task simulations | • 45 X Single Fermentation Task Estimators  
  • Standard D3v2 nodes (4 core, 14336MB) 10 threads  
  LOCAL SERVER  
  • 25 X Task calculation request threads  
  10 X Task result processing threads | 126hr*  
  *Stronger Local Server can improve up to 50% |

Table 1-15: Runtime statistics for different computational experiments with distributed BioLego
a. Optimizing corn utilization

We performed exploratory search for optimal biomass utilization for three types of corn biomasses – corn cobs, corn fiber and corn stover. The exact biomass compositions for this experiment were derived from (Baldwin and Sniegowski, 1951; Mcaloon et al., 2000; van Eylen et al., 2011). Five potential fermentation targets were evaluated: Ethanol, Acetone, 1-Butanol, (R)-Propane-1,2-diol and (S)-Propane-1,2-diol.

For each tested biomass and fermentation target we performed simulations for both single-step and for two-step fermentations under either aerobic or anaerobic conditions for four organism models currently integrated in BioLego flow. In total, this experiment included 480 different simulations. Azure Cloud side for this experiment included 10 instances of model setup constructors each with 2 threads running on Standard A1v2 nodes with 1 core and 2048MB memory and 10 instances of single fermentation task estimating instances, running in 3 threads on Standard A1v2 nodes with 1 core and 2048MB memory. The calculations of all scenarios, which included construction of 240 model setup constructions and 480 single fermentation task simulations was completed in 21 mins (Table 1-15).

For ethanol production we predict corn cobs to be most promising among estimated corn medias, while maximal production yields (for all medias) were predicted for anaerobic two-step scenarios starting with S. cerevisiae.

For (S)-Propane-1,2-diol production we predict corn fibers to be most promising among estimated corn medias, while maximal production yields (for all medias) were predicted for two-step scenarios starting either with S. cerevisiae or with C. acetobutylicum. Interestingly, all the scenarios with positive production yields (including the aerobic one) include E. coli as one of the fermenting organisms.

Surprisingly, for other targets (acetone, 1-butanol and (R)-Propane-1,2-diol) we did not predict any positive production yields.
### b. Sensitivity and gradient analyses for ethanol production using fermentation of Kappaphycus alvarezii biomass

We analyzed the possible effects of changes in composition of *K. alvarezii* algal biomass (see 1.3.6) on the ethanol production yield in the anaerobic two-step process, in which first organism is *S. cerevisiae* and the second one is *E. coli*. As we show previously, this fermentation configuration is predicted to result in high ethanol production (1.4.4).

*K. alvarezii* media is composed of 37 different compounds (Table 1-12), testing each requires single fermentation estimation task (Equation 1-16, Equation 1-17). For this experiment Azure Cloud side included 5 instances of single fermentation task estimating instances, running in 3 threads on Standard A1v2 nodes with 1 core and 2048MB memory. Both sensitivity analysis and gradient estimations finished after 4 minutes.

Naturally, the ethanol production yields appeared to be most sensitive to the presence of Galactose and Glucose in media. Omission of these metabolites respectively decreased the ethanol production yields to 46-53% and 44-57% (range is minimal to maximal predicted productions). Surprisingly, omission of some media components (Glutamic acid, Glycine, Tyrosine and Leucine) slightly (up to 0.3%) increased predicted maximal ethanol production. These findings can be explained by decreased organism growth rate, which leaves more media components for ethanol production.

Surprisingly, gradient estimation results for minimal and maximal ethanol production yields differ from each other. First shows highest gradient for various monosaccharides media components – Mannose, Glucose, Galactose, Xylose, Arabinose; and lowest for Glutamic acid, Aspartic acid and Asparagine. While second receives higher values not only for sugars but also for some amino acids, like Threonine (highest) and Serine. Interestingly, Asparagine, Aspartic acid and Serine have positive gradient for maximal and negative gradient for minimal predicted ethanol production yields.
c. Two-step fermentation with knock-outs in each organism

We analyzed the anaerobic two-step fermentation process of *U. lactuca* into ethanol, in which first organism is *S. cerevisiae* and the second one is *E. coli*. Here we investigated the impact of knocking out single organism reaction either in one or in both organisms on estimated ethanol production yield (see Equation 1-18).

The *S. cerevisiae* model is composed of 2,280 metabolic reactions, and the *E. coli* model is composed of 2,914 metabolic reactions. In total, this leads to 6,649,115 possible single knockout scenarios.

![Figure 1-11: Heatmap of the expected minimal and maximal ethanol production yields per knocked pair of reactions, in [%] of WT. Here x-axis corresponds to number of minProd(\textit{act}) in Equation 1-18 and y-axis to number of maxProd(\textit{act}) giving the observed WT percentage.](image)

To handle a task of this scale, the local server side was configured to operate 25 task calculation-request-sending threads and 10 result-processing threads. The Azure Cloud side for this calculation included 45 single fermentation task estimating instances, each one running with 10 threads on Standard D3v2 nodes with 4 cores and 14336MB memory.
Running time for this experiment was 126 hours. Notably, this time can be significantly reduced (up to 50-70%), with the same Azure Cloud side configuration, by using a stronger local server machine.

The results describing the predicted impact of all knockouts on both minimal and maximal ethanol productions (Equation 1-18) are presented as heatmap in Figure 1-11. Minimal ethanol production yield is predicted to be increased at most to 146% of the WT yields (for effectively one knockout candidate pair); while for maximal ethanol production this number reaches 170% of WT (for 867 knockout candidate pairs).

There are 24 reaction pairs for which knockouts are predicted to increase both minimal and maximal ethanol productions by more than 130%. These 24 reaction pairs are combination of 2 reactions from *E. coli* and 12 reactions from *S. cerevisiae*.

The *E. coli* reactions consist, in fact, of one artificial and one real reaction working in a linear process of proton (H+) extractor. This means that in effect this result represents a single unique reaction in *E. coli*. This reaction pairs with 9 non-orphan *S. cerevisiae* reactions, this leading to 9 interesting combinations that can potentially be further explored.

The *E. coli* reaction can be encoded by one of the following *E. coli* genes: b0241, b0929, b1377, b2215. The 9 *S. cerevisiae* non-orphan reactions map to 35 known genes. Analyzing these genes with GOrilla (Eden et al., 2007, 2009) resulted mostly in proton-transporting and biosynthesis processes (p-value of 1E-15 and less).
d. Experimental validation of prediction quality

We compared the experimentally measured ethanol fermentation efficiency with the efficiency predicted by BioLego system in four different scenarios: two serial two-step fermentation scenarios (S. cerevisiae followed by E. coli and E. coli followed by S. cerevisiae) and 2 single step fermentations (S. cerevisiae and E. coli) with presence of small initial amount of oxygen. Specifically, organisms were grown for 24 hours in closed vials with small amount of air. In these experiments we measured ethanol production yield as a function of initial amount of sugars (Rhamnose, Galactose, Glucose, Xylose, Fructose and Glucuronic Acid) in U. lactuca biomass.

![Figure 1-12: Experimental validation of BioLego predictions](image)

During BioLego simulations to match the experimental conditions, we assumed that U. lactuca biomass is composed purely from the measured sugars (Rhamnose, Galactose, Glucose, Xylose, Fructose and Glucuronic Acid) and evaluated ethanol yields under a small allowed influx of oxygen (0-2mmol*gDW^{-1}*h^{-1}, see 1.3.7). Figure 1-12 demonstrates the similarity of the predicted results with the performed measurements.
Both scenarios which started with *S. cerevisiae* demonstrate good alignment (increasing with the decrease of oxygen influx) with actual measurements. In both scenarios started with *E. coli*, the BioLego predicts complete digestion of sugars by this organism, which does not happen in practice. We hypothesize that major reason for such discrepancy lays in simulation assumption of infinite growth time given to bacteria. For example, we measured the amount of Glucose in the initial media to be 56 mg/mg DW, while after the fermentation with *E. coli* to be 31 mg/mg DW. In infinity, assumed by BioLego simulations, we predict complete digestion of Glucose. Thus, in the setup when *S. cerevisiae* is grown as a second organism, it should not have any media components sufficient for ethanol production.

Another interesting hypothesis rises from the analysis of ethanol yield as a function of oxygen influx. Maximal similarity for *S. cerevisiae* predictions appears at low values of oxygen influx (maximum is in completely anaerobic setup), while for *E. coli* predictions best similarity needs at least 2mmol/h*gDW influx of oxygen. We hypothesize that oxygen-uptake transporters in these two organisms may work with different rates (approximately 1 order of magnitude) in the same conditions. This hypothesis is well aligned with the findings of Hagman et al, who report approximate *S. cerevisiae* oxygen uptake rate of 3.5mmol/h*gDW (Hagman *et al.*, 2014) and of Andersent et al, who report approximate *E. coli* oxygen uptake rate of 20mmol/h*gDW (Andersent and Meyenburgt, 1980).
Chapter 2. Improving genome-scale metabolic models with fitness data.

This project addresses the development of methodology for improvement of genome-scale metabolic models leveraging novel data type – gene fitness assays.

The manuscript describing this project (Vitkin et al., 2018) was published in BMC Bioinformatics: Vitkin, E., Solomon, O., Sultan, S., Yakhini, Z. “Genome-wide analysis of fitness data and its application to improve metabolic models” (October, 2018)

![Diagram](image)

**Figure 2-1:** The general flow for using fitness assays integrated together with additional data sources to improve the assignment of genes to orphan reactions

In Chapter 2 we present the proposed methodology, briefly depicted on Figure 2-1.
Section 2.1. Background

Progress in sequencing techniques has greatly improved our understanding of bacterial genomes (Riley et al., 2006; Zhou and Rudd, 2013). In parallel, technologies that support modifying the genomic sequences of living organisms, including bacteria (van Opijnen and Camilli, 2013; Bikard et al., 2014; Wetmore et al., 2015), enable targeting of known loci in the genome. The combination of these developments facilitates studying bacterial gene function by physically modifying related sequences in living genomes and measuring the phenotypic effects triggered by such modifications. An important example of this emerging technique is organism fitness profiles (Deutschbauer et al., 2011; Wetmore et al., 2015), where organism growth rates in different conditions and under different genomic modifications are measured. Progress in the quality and scope of synthetic DNA libraries and in applying them to studying regulation in living cells (Sharon et al., 2014; Weingarten-Gabbay et al., 2016; Melnikov et al., 2012; Sharon et al., 2012), as well as more affordable sequencing methods, support higher throughput approaches to phenotypic analysis of synthetically modified genomes.

The understanding of bacterial genomes enables the use of metabolic models for designing bacterial production systems and other synthetic biology devices. Genome-scale metabolic network models leverage the existing knowledge of organism biochemistry and genetics to construct a framework for simulating processes. The core of the metabolic model is the information about the stoichiometry of the metabolic reactions and the associations between protein coding genes, and the reactions that they catalyze (Edwards et al., 2002). iJO1366 (Orth et al., 2011), which is the latest model of Escherichia coli K-12 MG1655, contains information about 1366 genes, 1136 unique metabolites and 2251 metabolic reactions, out of which 128 reactions are orphan (70 metabolic and 58 transport), meaning that they are not associated with any gene.
An important part of the methodology developed here is the use of high throughput fitness data to infer genes that potentially encode for proteins catalyzing orphan reactions. Current approaches rely on the idea that genes and reactions in the local neighborhood have similar behavioral profiles. The exact definition of these profiles is deduced from the nature of the available biological data, such as sequence similarity (phylogenetic profile), sequence genomic context, gene-metabolome associations, gene expression data and others. For the best of our knowledge, none of the recent metabolic modeling studies proposes a method to improve the assignment of genes to reactions using fitness assays alone or incorporated with additional data sources (Chae et al., 2017; Saha et al., 2014; Rai and Saito, 2016; Chen and Vitkup, 2006; Green and Karp, 2007; Vitkin and Shlomi, 2012; Fuhrer et al., 2017).

The proposed mathematical framework is developed and tested over the iJO1366 *E. coli* model. Moreover, we propose candidate genes for 107 out of 128 orphan reactions and provide confidence estimations for each such candidate.
Section 2.2. Methods

2.2.1. Obtaining the fitness data

In the current study, we used Wetmore et al. (Wetmore et al., 2015) data as a comprehensive dataset for fitness in *E. coli* in 48 different media conditions. We analyzed both coding and non-coding regions. For coding sequence regions (CDS) we used fitness scores directly reported in Wetmore et al). For non-coding insertions (insertions in non-coding regions of the genome, i.e. promoters and un-translated regions, UTRs), which were not analyzed in the original paper, we calculated average fitness scores as explained below. Annotations of non-coding regions in *E. coli* were taken from RegulonDB and EcoCyc (Gama-Castro et al., 2016; Keseler et al., 2017). We denote an insertion in a genomic region position *p* by *Ins*(*p*). Considering a non-coding region *nc* and media condition *γ*, we compute (Equation 2-1):

Equation 2-1: Fitness score for non-coding region

\[
fitness\_score\,(nc,\gamma) = \frac{1}{|\pi\,(nc)|} \sum_{\pi\,(nc)} fitness\_score(Ins(p),\gamma)
\]

Where \( \pi\,(nc) = \{Ins(p) : p \in nc\} \), i.e. all insertions in genomic positions within the considered non-coding region. In other words, all the fitness scores related to genomic positions within a non-coding region are averaged to score this region. The value \( fitness\_score(Ins(p),\gamma) \) is taken from the raw insertion results table provided by Wetmore et al.

For each genomic element of interest (either coding or non-coding) we define its *fitness vector* as a vector combining its *fitness_score* over all available conditions *γ* in the dataset.
2.2.2. Fitness vectors for non-orphan reactions

A major challenge in constructing metabolic models is an assignment of genes to orphan reactions. The majority of automatic approaches addressing this issue (Green and Karp, 2004; Kharchenko et al., 2006, 2004; Vitkin and Shlomi, 2012) operate under the reasonable assumption that the behavior of each reaction (both orphan and non-orphan) is highly correlated with the behavior of the reaction neighborhood. Here, we also use this logic while addressing the reaction behavior in the context of the gene fitness data. To do so we need a definition of the fitness of non-orphan reactions based on the fitness measurements of its genes. Following the Boolean formulation of mapping between genes and reactions (see Equation 1-1: Gene-to-Reaction Boolean mapping), we define the organism fitness value at condition $\gamma$ after the knockout of a certain reaction to be the maximum over the fitness values of the disjunctive clauses, which, in turn, are the minimum over individual gene fitness values under the conjunction sign (Equation 2-2A). To continue the specific example described above in the context of Equation 1-1B, the fitness of $r_o$ in the condition $\gamma$ is described in Equation 2-2B.

**Equation 2-2: Fitness score for non-orphan reaction**

(A) $\text{fitness\_score}(r, \gamma) = \max_{p \in \text{Proteins}(r)} \left\{ \min_{g_j \in \text{Genes}(p)} \{ \text{fitness\_score}(g_j, \gamma) \} \right\}$

(B) $\text{fitness\_score}(r_o, \gamma) = \max \left\{ \frac{\text{fitness\_score}(p_1, \gamma)}{\text{fitness\_score}(p_2, \gamma)} \right\} = \max \left\{ \min \left\{ \frac{\text{fitness\_score}(g_1, \gamma)}{\text{fitness\_score}(g_2, \gamma)} \right\} \right\}$

Once created, reaction fitness values are Z-normalized per each condition as in Equation 2-3, where $ZFS(r, \gamma)$ denotes Z-normalized fitness score of reaction $r$ in condition $\gamma$.

**Equation 2-3: Z-normalized fitness score for non-orphan reaction**

$ZFS(r, \gamma) = \frac{\text{fitness\_score}(r, \gamma) - \text{mean} \{ \text{fitness\_score}(r, \gamma) \} }{\text{STD} \{ \text{fitness\_score}(r, \gamma) \} }$

A fitness vector for a given reaction $r$, denoted as $ZFV(r)$, is now constructed using $ZFS(r, \gamma)$, for all conditions $\gamma$ in the dataset.
2.2.3. Fitness vectors for genes and promoters

A fitness vector for a gene $g$ is directly taken from all measurements reported for $g$ in Wetmore et al (Wetmore et al., 2015), running across all conditions $\gamma$ in the dataset and using $Z$-normalized values (alike to Equation 2-3). We denote this vector by $ZFV(g)$.

Assignment of genes to orphan reactions may also benefit from the information available for the gene promoters. Clearly, if a candidate gene for some orphan reaction is highly correlated with the reaction neighborhood, but its known promoter does not confirm such correlation – this information should be incorporated in the total ranking of this gene with respect to other candidates. Thus, we estimate the fitness scores of each promoter according to Equation 2-1 and then construct $ZFV(p)$, the vector of $Z$-normalized fitness scores (estimated as in Equation 2-3) for the promoter $p$, running across all conditions.

2.2.4. Reaction neighborhood

Metabolic reactions connected to each other by shared metabolites are called neighboring reactions. Such neighboring reactions may have very similar activity patterns, i.e. there is a high chance that a reaction is active when its neighbor is active. Such activity similarities are most evident on linear metabolic pathways, when products of one reaction are transferred as input substrates to another.

We define a reaction to be adequate for neighborhood analysis (in short – adequate) if it has two or more non-orphan neighbors (with assigned fitness vectors).

Above-mentioned behavioral similarity between two neighboring reactions will not exist for all shared metabolites. Metabolites like H2O, H+ and others appear in extremely high numbers of reactions, thus taking them into account for the neighborhood assessment is almost always noisy. We call them high-frequency metabolites.
Figure 2-2: Greedy removal of high-frequency metabolites.

To correctly assess the neighborhood of a given reaction $r$, we greedily (from high to low frequencies) remove from its equation all high-frequency (frequency higher than 11) metabolites unless such removal leads to less than two non-orphan neighbors remaining for $r$ or for some other reaction. For example, as depicted at Figure 2-2, the high-frequency metabolite $h_f M$ will be removed from reaction $R1$, since after the removal it still has two non-orphan neighbors. However, it won’t be removed from reactions $R2$ and $R3$, since after the removal number of non-orphan neighbors will be 1 for each.

The process of removing high-frequency metabolites is greedy. It starts from the metabolite with the highest frequency and iterates over its reactions starting from the one with the most neighbors. The considered metabolite is omitted from the reaction unless this reaction or any of its neighbors become not adequate, i.e. unless the number of non-orphan reaction neighbors becomes less than 2. After the reaction update is finished the process continues in the same manner to the next reaction and/or metabolite. Notice, that final definition of reaction neighbors may still depend on the order of operations in this process.
2.2.5. Fitness vectors for orphan reactions. Gene–to-Reaction assignment based on reaction neighborhoods

For the orphan reactions, i.e. reactions without any known associated gene, the method above (see 2.2.2) for construction of fitness vectors is obviously not appropriate. However, we can guess that certain candidate gene candG is an encoding gene for the given orphan reaction Ro. If this guess is correct, it is reasonable to expect high similarity of the fitness behavior of candG with the fitness behavior of the genes on the non-orphan reactions in its neighborhood. We measure this similarity using a Spearman correlation test.

For each adequate orphan reaction Ro we define the Association Likelihood Score (in short – ALS) of assigning the candidate gene candG to Ro as the mean Spearman correlation of the fitness vector of candG to the 2 most-correlated fitness vectors of non-
orphan neighbors of \( Ro \) (Equation 2-4). The number “2” was selected as the minimal value, which allows solving single orphan reaction gap in a linear pathway.

**Equation 2-4: Association Likelihood Score for a given Reaction-Gene pair**

\[
ALS(Ro, candG) = \frac{1}{2} \left[ \text{Spearman}(ZFV(candG), ZFV(neighb_1(Ro))) \right] + \left[ \text{Spearman}(ZFV(candG), ZFV(neighb_2(Ro))) \right]
\]

Here \( neighb_1(Ro) \) and \( neighb_2(Ro) \) are the two neighbor reactions of \( Ro \) with fitness vectors most Spearman correlated to \( ZFV(candG) \).

**2.2.6. Combination of gene and promoter fitness**

Gene-to-reaction Spearman-based assignment scores can benefit from associated promoter fitness data. Indeed, this is valid for genes with existing mapping to promoters, as acquired from RegulonDB and EcoCyc (Keseler et al., 2017; Gama-Castro et al., 2016). If certain gene has high ALS score, but its known promoter does not support it (according to ALS score calculated as in Equation 2-5) – this should be taken into account.

**Equation 2-5: Association Likelihood Score for a given Reaction-Promoter pair**

\[
ALS(Ro, candP) = \frac{1}{2} \left[ \text{Spearman}(ZFV(candP), ZFV(neighb_1(Ro))) \right] + \left[ \text{Spearman}(ZFV(candP), ZFV(neighb_2(Ro))) \right]
\]

We define the adjusted (gene, promoter)-to-reaction ALS score based on fitness data as follows (Equation 2-6):

**Equation 2-6: Association Likelihood Score for a Reaction-(Gene&Promoters) pair**

\[
ALS_{fitness}(Ro, candG) = ALS(Ro, candG) + \alpha \sum_i ALS(Ro, candP_i)
\]

Where \( \{ candP_i \} \) is the set of all the promoters associated with \( candG \) (an empty set in case there is no such information) and the \( ALS(Ro,candP) \) is estimated according to Equation 2-5. We have tested different values of \( \alpha \) and the best performance was obtained at \( \alpha=0.21 \).
2.2.7. Uniform gene-to-reaction assignment null models

Two baselines strawman gene-to-reaction assignment scores were used to represent uniform assignment of genes to reactions as a basis for comparison.

First, under \textit{WEAK U-dist}, we assume that random selection of top $G$ genes for $R$ reactions will successfully identify $\frac{RG}{TotalGenes}$ of true reaction genes. For example, random selection of top 10 out of 3646 candidate genes for 1556 non-orphan reactions will identify true genes for $\frac{10*1556}{3646} \approx 4.3$ reactions.

Second, under \textit{STRONG U-dist}, we assume that the predicted genes are sampled only from the set of true reaction genes only (i.e. each sampled gene is a true encoding gene for some non-orphan reaction). Moreover, since some reactions are activated by several genes, at each step average number of genes per reaction will be identified. That is, selection of top $G$ genes will identify $G \times \text{Avg(genes per rxn)}$ of true reaction genes. For example, selection of top 10 candidate genes with average of 3.126 true genes per reaction in iJO1366 model will identify true genes for $10*3.126 \approx 31.26$ reactions.

2.2.8. Integration of fitness- and expression-based scores

Here we explain how to integrate ALS scores estimated based on fitness assays and on gene expression profiles.

Two gene expression datasets were used to associate reactions to expression vectors. The first is GSE32561 from Goh et al, which includes 11 gene expression microarray measurements (Goh et al., 2012). The second is GSE58806 from Keating et al, which includes 36 gene expression measurements (Keating et al., 2014). As a preprocessing step, we omitted expression data for genes which were not covered in Wetmore and colleagues’ fitness analysis (Wetmore et al., 2015) (Figure 2-4).
Figure 2-4: iJO1366 model genes covered by fitness and expression datasets.

First, the expression values were Z-normalized per each condition separately (as in Equation 2-3). Second, reaction expression vectors (ZEV) were constructed in a manner like that used for reaction fitness vectors (ZFV). Finally, using these normalized reaction- and gene-expression vectors we estimated $ALS_{expression}$ score for each reaction-gene pair according to Equation 2-4.

We defined the combined ALS score based on fitness and expression data as an average between the score resulted from each data source separately (Equation 2-7):

**Equation 2-7: Combined Association Likelihood Score for a given Reaction-Gene pair**

$$ALS_{combined}(Ro, candG) = \frac{1}{2}ALS_{fitness}(Ro, candG) + \frac{1}{2}ALS_{expression}(Ro, candG)$$
2.2.9. Confidence of gene to reaction assignment

We define confidence of each gene-to-reaction assignment as probability of receiving an assignment score $\sigma$ for true gene-reaction pair in both Bayesian (Equation 2-8A) and unbiased (Equation 2-8B) approaches.

Equation 2-8: (A) Bayesian and (B) Unbiased confidence of $ALS(\text{reaction, gene})$

**(A)** $BConf[ALS(r, g) = \sigma] = \frac{P(ALS(r, g) \geq \sigma | g \in genes[r]) * P(g \in genes[r])}{P(ALS(r, g) \geq \sigma)}$

**(B)** $UConf[ALS(r, g) = \sigma] = \frac{P(ALS(r, g) \geq \sigma | g \in genes[r])}{P(ALS(r, g) \geq \sigma | g \in genes[r]) + P(ALS(r, g) \geq \sigma | g \in genes[r])}$

Where $P(\{g \in genes(r)\}) \approx 0.044\%$ is the general probability of correct gene-reaction assignment, as calculated on the set of non-orphan reactions and the genes measured by Wetmore et al (Wetmore et al., 2015). And the distribution of scores obtained for spurious pairs versus the scores obtained for known pairs is depicted on Figure 2-5.

![Figure 2-5: Distribution of gene-to-reaction ALS scores](chart.png)
Section 2.3. Results and Findings

2.3.1. Performance evaluation based on non-orphan reaction data

To assess the quality of assigning genes to reactions using our approach we cast non-orphan reactions as orphans, one reaction at a time, and test our ability to reconstruct the known hidden assignment of at least one original gene. For each such reaction, we calculate the rank of the true known gene or genes out of all candidate genes. In total, iJO1366 metabolic model of *E. coli* contains 1556 adequate non-orphan reactions, which were used for this validation purpose (Figure 2-6 and Table 2-1):

![Graph identifying true genes for 1556 nonorphan reactions](image)

**Figure 2-6**: Comparative accuracy of Association Likelihood Score values for first gene prediction in a leave-one-out validation on the 1556 adequate non-orphan reactions in iJO1366 metabolic model of *E. coli*

Table 2-1 clearly shows that all the scoring approaches perform significantly better than random reaction assignments. More importantly, assignments based on fitness data were significantly better than those based on expression data – 503 vs 404 (24.5% improvement) correctly identified gene-reaction pairs in top-20 candidates. Considering both types of data produced even better results of 614 correct predictions in top-20 candidates for 1556 validation reactions (52% increase). Finally, incorporation of
promoter data boosts the performance on average by 0.5-2%. The combined prediction accuracy based on fitness data together with promoter and expression data is 39.7%, when considering the top-20 candidates, and 19.7% accuracy at predicting at the top candidate, which is respectively 52.7% and 66.3% better than accuracy based solely on expression data.

<table>
<thead>
<tr>
<th>Score</th>
<th>1st candidate</th>
<th>In top-5</th>
<th>In top-10</th>
<th>In top-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRONG U-dist (null)</td>
<td>3.13 (0.20%)</td>
<td>15.63 (1.00%)</td>
<td>31.26 (2.01%)</td>
<td>62.53 (4.02%)</td>
</tr>
<tr>
<td>Expression (GSE32561)</td>
<td>106 (6.8%)</td>
<td>228 (14.7%)</td>
<td>262 (16.8%)</td>
<td>290 (18.6%)</td>
</tr>
<tr>
<td>Expression (GSE58806)</td>
<td>126 (8.1%)</td>
<td>239 (15.4%)</td>
<td>273 (17.5%)</td>
<td>310 (19.9%)</td>
</tr>
<tr>
<td>Expression (combined)</td>
<td>184 (11.8%)</td>
<td>334 (21.5%)</td>
<td>363 (23.3%)</td>
<td>404 (26.0%)</td>
</tr>
<tr>
<td>Fitness</td>
<td>176 (11.3%)</td>
<td>413 (26.5%)</td>
<td>464 (29.8%)</td>
<td>503 (32.3%)</td>
</tr>
<tr>
<td>Fitness + Promoters</td>
<td>190 (12.2%)</td>
<td>425 (27.3%)</td>
<td>478 (30.7%)</td>
<td>513 (33.0%)</td>
</tr>
<tr>
<td>Fitness + Expression</td>
<td>297 (19.1%)</td>
<td>504 (32.4%)</td>
<td>567 (36.4%)</td>
<td>614 (39.5%)</td>
</tr>
<tr>
<td><strong>Fitness + Promoters + Expression</strong></td>
<td><strong>306 (19.7%)</strong></td>
<td><strong>511 (32.8%)</strong></td>
<td><strong>575 (37.0%)</strong></td>
<td><strong>617 (39.7%)</strong></td>
</tr>
</tbody>
</table>

Table 2-1: Comparative accuracy of scores predicting first reaction gene assignments for 1556 adequate non-orphan reactions in iJO1366 metabolic model of *E. coli*

Similar leave-one-reaction-out analysis was performed for 565 validation reactions with at least two known genes (Figure 2-7 and Table 2-2). Here the random gene-to-reaction assignment baseline is far more constraining and still, the performance of the assignment based both on expression and on fitness datasets has much better quality.

<table>
<thead>
<tr>
<th>Score</th>
<th>2nd candidate</th>
<th>In top-5</th>
<th>In top-10</th>
<th>In top-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRONG U-dist (null)</td>
<td>4.45 (0.79%)</td>
<td>11.11 (1.97%)</td>
<td>22.23 (3.93%)</td>
<td>44.46 (7.87%)</td>
</tr>
<tr>
<td>Expression (GSE32561)</td>
<td>3 (0.5%)</td>
<td>14 (2.5%)</td>
<td>27 (4.8%)</td>
<td>34 (6.0%)</td>
</tr>
<tr>
<td>Expression (GSE58806)</td>
<td>4 (0.7%)</td>
<td>19 (3.4%)</td>
<td>25 (4.4%)</td>
<td>31 (5.5%)</td>
</tr>
<tr>
<td>Expression (combined)</td>
<td>11 (1.9%)</td>
<td>36 (6.4%)</td>
<td>41 (7.3%)</td>
<td>53 (9.4%)</td>
</tr>
<tr>
<td>Fitness</td>
<td>6 (1.1%)</td>
<td>31 (5.5%)</td>
<td>41 (7.3%)</td>
<td>56 (9.9%)</td>
</tr>
<tr>
<td>Fitness + Promoters</td>
<td>6 (1.1%)</td>
<td>30 (5.3%)</td>
<td>44 (7.8%)</td>
<td>62 (11.0%)</td>
</tr>
<tr>
<td>Fitness + Expression</td>
<td>51 (9.0%)</td>
<td>85 (15.0%)</td>
<td>94 (16.6%)</td>
<td>122 (21.6%)</td>
</tr>
<tr>
<td><strong>Fitness + Promoters + Expression</strong></td>
<td><strong>53 (9.4%)</strong></td>
<td><strong>85 (15.0%)</strong></td>
<td><strong>92 (16.3%)</strong></td>
<td><strong>122 (21.6%)</strong></td>
</tr>
</tbody>
</table>

Table 2-2: Comparative accuracy of scores predicting second reaction gene assignments for 565 adequate non-orphan reactions in iJO1366 metabolic model of *E. coli* with at least 2 encoding genes.
Like in the case of the prediction of the first true reaction gene, the best accuracy was obtained when combining fitness and expression datasets together and with the incorporation of promoter fitness data. This combination led to accuracy of 53 (9.4%) for correctly identifying a 2nd candidate gene (meaning that the first candidate gene is also correctly identified) and of 122 (21.6%) at the cutoff of top-20 candidates, that is an improvement of respectively 382% and 130% compared to using solely gene expression data.

We also analyze the dependence of the performance of our approach on the number of conditions measured, and observe a clear benefit gained by increasing the number of measured conditions (Figure 2-8).
2.3.2. Predicting genes for orphan reactions

Gene-to-reaction assignment was performed using the combined fitness and expression scores with incorporation of promoter fitness data (Equation 2-7). Out of 128 orphan reactions in iJO1366 metabolic model of E. coli we identified 107 adequate assigned them genes.

We verified our predictions with EcoCyc (Keseler et al., 2017) data and further substantiated some of our findings based on this cross comparison. For example, for the orphan transporter reaction ALAt2rpp (L-alanine reversible transport via proton symport cytoplasm-periplasm) the first predicted candidate (Bayesian confidence of ~4% and Unbiased confidence of ~98.97%) is cycA (b4208), which can act as an L-alanine transporter (Schneider et al., 2004; Robbins and Oxender, 1973). Another example is the internal reaction CMPN (CMP nucleosidase: “\(\text{CMP} + \text{H}_2\text{O} \rightarrow \text{Cytosine} + \alpha\text{D-Ribose-5P}\)”, EC:3.2.2.10), for which the first predicted candidate (Bayesian confidence of 23% and unbiased confidence of ~99.85%), is rihC (b0030), known as “ribonucleoside hydrolase 3” and as “non-specific ribonucleoside hydrolase” (EC:3.2.-.-), and catalyzing, among others, the “Cytidine+H2O→D-ribofuranose + Cytosine” reaction (Petersen and Møller, 2001). This knowledge is not captured in the model and is revealed by our analysis.
Chapter 3. Analysis of steroid metabolism

This project addresses the analysis of steroid hormone metabolism with non-flux balance analysis-based approaches. It was performed in collaboration with researchers from Agilent Research Labs in Israel, Rambam Medical Center in Israel and Center of Child and Adolescent Medicine in Germany.

This study was presented on numerous conferences and received a significant interest from the community. It was published in Steroids journal in 2014 (Vitkin et al., 2014) and selected as a paper with significant impact for Diabetes and Obesity section of World Biomedical Frontiers in 2105 (Vitkin et al., 2015).

Section 3.1. Background

The analysis of a complex disorder such as childhood obesity by Chromatography–Mass Spectrometry (GC-MS) is associated with high throughput measurements. Such analysis requires a ladder of computational steps that include among others: (i) data normalization; (ii) pattern finding and (iii) result visualization methods. Here we present such a bioinformatics approach to generate insights for the studied cases. This includes proposal of novel data normalization and novel result visualization methods.

In a typical metabolomic dataset, the metabolites concentrations and enzymatic activities depend both on case-specific pathway properties and on factors unrelated to the disease process such as age and body composition (metadata). The identification and normalization for such confounding factors is non-trivial, yet an essential for the discovery of targeted dependencies. Current approaches to data normalization aim to reduce confounding factors by splitting the data to arbitrary age, gender or BMI groups (bins) and Z-normalizing each such bin separately. This approach suffers from three shortcomings: (i) The number of subjects in each bin is rapidly decreasing with the growing number of metadata dimensions, limiting the statistical power. (ii) It is not clear
a priory how to define the bins’ boundaries. (iii) The method will perform poorly on subjects close to the bin boundaries.

Here we present an alternative novel normalization approach, we call *peer-group normalization*, which overcomes these limitations and markedly reduces confounding meta-factors.

An ideal analysis visualization would be presented in the context of the steroidogenesis pathway and would incorporate all metabolomic data, including blood and urine levels of each metabolite, as well as levels of activity for each enzymatic reaction. However, such full pathway drawing is complex, and would hinder human comprehension even without considering quantitative information (Dutta *et al.*, 2012; Ogata *et al.*, 1999). On the other hand, the commonly used textbook scheme of the pathway that excludes secondary and intermediate compounds is oversimplifying and ignores quantitative information, such as the relative concentrations of metabolites and rates of enzymatic reactions (Weiser, 2008). Here, we propose a novel visualization approach that strikes a good balance between pathway simplicity and completeness; and displays quantitative information of interest.

As a validation of the proposed pipe, we initially applied it to find fingerprints of a well-studied steroid disorder – congenital adrenal hyperplasia due to 21-hydroxylase deficiency (CAH).

Finally, we apply this approach for a ‘steroid-related complex disorder’ such as childhood obesity and generate some novel insights.
Section 3.2. Methods and Subjects

3.2.1. Subjects

Three datasets were utilized: (i) normal subjects of matched age to, (ii) CAH patients and (iii) obese children age 2-18. The first dataset included 191 females and 133 males; they span a Body Mass Index (BMI) range from 13 to 29 kg/m². The CAH dataset included 27 subjects, (15 females) age 6-12, with a BMI range of 14-29 kg/m². The obesity dataset contained 70 subjects (35 males), age 4-17 and BMI 19-42 kg/m² (BMI SDS >2).

3.2.2. GC-MS Data

Steroid profiles were analyzed using quantitative data that were generated by GC-MS analysis (Wudy et al., 2007; Remer et al., 2005).

The 31 urinary metabolomic measurements were combined to infer 12 blood metabolites concentrations. Each blood metabolite concentration value was estimated as the sum of concentrations of its urinary end product. In addition, enzymatic rates for 13 pathway reactions were estimated using the product-to-substrate ratios. To convert the rates distribution to be Normal-like, we use the log-value of the estimated rate in further analyses.

In total, 56 evaluated measurements, including blood and urinary metabolomic concentrations and enzymatic activities are available for each subject.

3.2.3. Peer-Group Normalization

For each subject (either case or control) we choose a set of control subjects with similar metadata values. We call such set of control subjects – peer group. The peer group data are used as context to normalize subject variables, alike bins in the original approach. The process is as following.

First, we quantify the similarity between subjects. To this end we define a distance function that depends on the metadata values. Formally, let \( p \) and \( q \) be two subjects, and
let $d$ be some metadata dimension (e.g. age). We denote by $p_d$ the values of $p$ in the dimension $d$, transformed into the range $[0,1]$; $q_d$ is defined analogously. We define the distance between $p$ and $q$ to be the following weighted Euclidian distance (Equation 3-1):

**Equation 3-1: Distance in metadata space between two subjects**

\[
\text{dist}(p,q) = \sqrt{\sum_{d \in \text{metadata}} [w_d (p_d - q_d)^2]}
\]

Where $w_d$ is the relative weight for dimension $d$, defined as the number of variables (out of 56), whose values in the control population is highly correlated (p-value<0.01) with the values of $d$. For example, if the subject sex is correlated with 10 variables and subject age is correlated with 39 variables (Table 3-1), then we assign $w_{\text{sex}}=10$, and $w_{\text{age}}=39$. This definition captures the following intuition: the greater the correlation for a certain metadata dimension against variables of interest, the greater emphasis we want to give to distances in that dimension.

Second, for each peer subject, we define its peer group to be the set of controls closer than a pre-estimated threshold distance (Figure 3-1). The threshold distance for pre-selected size $K$ is estimated as a minimal distance for which maximal number of affected subjects has a peer-group of at least $K$ controls (Equation 3-2).

**Equation 3-2: Threshold distance estimation**

\[
\text{ThresholdDistance}(K) = \text{argmin}_{\text{Radius}} \{ \exists \text{peer } - \text{group } P \ \forall \text{affected subject} : |P| \geq K \}
\]

Finally, the subject variable values for each of the 56 measurement inferred variables are $Z$-normalized using corresponding variable values in the peer group.
3.2.4. Analysis of Metabolites and Reactions

We are interested in identifying metabolites and reactions that have a statistically significant difference comparing case and control measurement inferred variables. To this end we employed both parametric (Student T-Test) and non-parametric (Threshold Number of Misclassifications, TNoM) methods (Ben-Dor et al., 2001; Bittner et al., 2000). The Student T-Test tests the difference of means between cases and controls, while accounting for the populations variances. TNoM finds the optimal threshold value separating both populations and evaluates the probability of obtaining such separation at random.
3.2.5. Urine-to-Blood (U2B) Visualization

Our visualization approach is based on the customary schematic representation of the steroid pathway, as depicted in traditional medical literature. Such representation usually includes major circulating metabolites (graph nodes) and reactions connecting them (graph edges). However, it does not provide quantitation of either nodes or edges.

Our Urine-to-Blood (U2B) visualization enhances this representation by (i) adding the relevant urinary end-products as sub-nodes and by (ii) adding color-code for up- and down-regulation of both nodes and edges (Figure 3-2, Figure 3-3).

Both Figure 3-2 and Figure 3-3 show the U2B visualization of human steroidogenesis pathway with 15 nodes (normally presented on this pathway) and 31 sub-nodes corresponding to the 31 urinary products measured by GC-MS. Even though urinary end-products may derive from more than a single circulating metabolite, to simplify visualization we schematically related each urinary metabolite to its dominant substrate. Then, a gradient blue-to-yellow color-code was applied to denote up- or down-regulation of both nodes and edges.

Blood metabolite (left) and urinary end-products (right) of the expanded steroid pathway. Triangles correspond to unidirectional blood enzymatic reactions. Rhombs correspond to bidirectional blood enzymatic reaction, while small triangle in it indicates the direction used for estimation of product-substrate ratio. Color saturation corresponds to analysis p-values showing down regulation in CAH as compared to control (blue), and up-regulation (yellow); white - to insignificant difference; gray - non-evaluated metabolites and reactions.

The additional major advantage of the proposed U2B approach is its ability to represent both population-based results (i.e. comparison of target population vs. control group) as well as subject-based results (i.e. comparison of certain subject with its peer-group).
Section 3.3. Results and Findings

3.3.1. Evaluation of peer-group normalization

The effect of the proposed peer-group normalization method was validated in the population of control patients.

In the raw data, out of total 56 measurement inferred variables for each subject (31 measured urine metabolites, 12 estimated blood metabolites and 13 estimated reactions) 10 were correlated (Pearson p-value<0.01) with sex, and ~40 were correlated with age, weight, height and BMI (Table 3-1).

<table>
<thead>
<tr>
<th>Control Dataset</th>
<th>gender</th>
<th>age</th>
<th>weight</th>
<th>height</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAW data</td>
<td>10</td>
<td>39</td>
<td>38</td>
<td>41</td>
<td>36</td>
</tr>
<tr>
<td>Normalization by sex and constant age ranges: [2.5-4.5), [4.5-6.5), [6.5-8.5), [8.5-10.5), [10.5-12.5), [12.5-14.5), [14.5-16.5) and [16.5-18.5)</td>
<td>27</td>
<td>30</td>
<td>26</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>Peer-group of size 10, metadata includes: age and gender</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Peer-group of size 10, metadata includes: age and gender, BMI, weight, height</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3-1: Counts of highly-correlated measurement inferred variables for each metadata dimension in the control population. These counts represent number of variables with correlation providing p-value <0.01 out of total 56 (31 urinal metabolites, 12 blood metabolites and 13 reaction activities).

Traditional normalization, performed by partitioning the population into bins based on age and sex and Z-normalizing each received bin, leads to a small reduction in correlating parameters, bringing the number of highly correlated data variables to 24-30 for all 5 meta-dimensions (Table 3-1).

On the other hand, peer-group normalization completely eliminated the strong impact of confounding factors (Table 3-1).
3.3.2. Data normalization analysis and visualization in CAH patients

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Variable Name</th>
<th>TTEST p-value</th>
<th>Corrected TTEST p-value</th>
<th>TNoM p-value</th>
<th>Corrected TNoM p-value</th>
<th>Sign (relative to peer-group)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urine</td>
<td>Po-5β,3α</td>
<td>5.63e-06</td>
<td>3.15e-04</td>
<td>1.10e-27</td>
<td>6.16e-26</td>
<td>Up</td>
</tr>
<tr>
<td>Urine</td>
<td>11-O-Pt</td>
<td>6.75e-10</td>
<td>3.78e-08</td>
<td>1.10e-27</td>
<td>6.16e-26</td>
<td>Up</td>
</tr>
<tr>
<td>Urine</td>
<td>PT</td>
<td>9.86e-05</td>
<td>5.52e-03</td>
<td>1.24e-25</td>
<td>6.94e-24</td>
<td>Up</td>
</tr>
<tr>
<td>Urine</td>
<td>Po-5α,3α</td>
<td>5.42e-06</td>
<td>3.04e-04</td>
<td>1.24e-25</td>
<td>6.94e-24</td>
<td>Up</td>
</tr>
<tr>
<td>Blood</td>
<td>17-OH Progesterone</td>
<td>1.16e-05</td>
<td>6.50e-04</td>
<td>1.10e-27</td>
<td>6.16e-26</td>
<td>Up</td>
</tr>
<tr>
<td>Reaction</td>
<td>21-Hydroxylase β</td>
<td>1.75e-09</td>
<td>9.80e-08</td>
<td>1.10e-27</td>
<td>6.16e-26</td>
<td>Down</td>
</tr>
</tbody>
</table>

Table 3-2: Most significant differences between control and CAH patients for the estimated values after normalization for gender, age, BMI, weight and height. Results are sorted according to TNoM p-values. P-value correction was done by Bonferroni correction.

CAH data analysis was performed after peer-group normalization in 5-dimensional space of sex, age, height, weight and BMI. The results confirmed the common knowledge about CAH (Figure 3-2); the most significantly up-regulated node is the derived 17-hydroxyprogesterone and its sub-nodes, while the most significantly reduced edge are cortisol and the enzyme 21-hydroxylase activity (Table 3-2).

**Figure 3-2:** (A) TNoM and (B) Student-TTest comparison between CAH-affected and control children.

Blood metabolite (*left*) and urinary end-products (*right*) of the expanded steroid pathway. Triangles correspond to unidirectional blood enzymatic reactions. Rhombs correspond to bidirectional blood enzymatic reaction, while small triangle in it indicates the direction used for estimation of product-substrate ratio. Color saturation corresponds to analysis p-values showing down regulation in CAH as compared to control (*blue*), and up-regulation (*yellow*); white - to insignificant difference; gray - non-evaluated metabolites and reactions.
3.3.3. Data normalization analysis and visualization in childhood Obesity

Childhood obesity data analysis was performed after peer-group normalization for both control and cases on a 2-dimensional space of sex and age. Weight, height and BMI were omitted being part of the entity definition.

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>TTEST p-value</th>
<th>Corrected TTEST p-value</th>
<th>TNoM p-value</th>
<th>Corrected TNoM p-value</th>
<th>Sign (relative to peer-group)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urine</td>
<td>Cortisol</td>
<td>1.01e-22</td>
<td>5.66e-21</td>
<td>7.22e-12</td>
<td>4.04e-10</td>
<td>Down</td>
</tr>
<tr>
<td>Urine</td>
<td>16α-OH-DHEA</td>
<td>3.82e-02</td>
<td>2.14e+00</td>
<td>8.93e-04</td>
<td>5.00e-02</td>
<td>Up</td>
</tr>
<tr>
<td>Urine</td>
<td>6β-OH-F</td>
<td>1.66e-14</td>
<td>9.30e-13</td>
<td>4.33e-03</td>
<td>2.42e-01</td>
<td>Down</td>
</tr>
<tr>
<td>Urine</td>
<td>DHEA</td>
<td>7.52e-02</td>
<td>4.21e+00</td>
<td>9.43e-02</td>
<td>5.28e+00</td>
<td>Up</td>
</tr>
<tr>
<td>Urine</td>
<td>20α-DHF</td>
<td>5.58e-05</td>
<td>3.12e-03</td>
<td>9.43e-02</td>
<td>5.28e+00</td>
<td>Down</td>
</tr>
<tr>
<td>Urine</td>
<td>PD</td>
<td>1.07e-04</td>
<td>5.99e-03</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Urine</td>
<td>THS</td>
<td>1.24e-04</td>
<td>6.94e-03</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Urine</td>
<td>αTHB</td>
<td>1.89e-03</td>
<td>1.06e-01</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Urine</td>
<td>11-O-Pt</td>
<td>2.02e-03</td>
<td>1.13e-01</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Urine</td>
<td>THF</td>
<td>4.78e-03</td>
<td>2.68e-01</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Urine</td>
<td>A5-3β,17β</td>
<td>5.23e-03</td>
<td>2.93e-01</td>
<td></td>
<td></td>
<td>Up</td>
</tr>
<tr>
<td>Urine</td>
<td>THB</td>
<td>5.49e-03</td>
<td>3.07e-01</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Blood</td>
<td>DHEA</td>
<td>3.69e-02</td>
<td>2.07e+00</td>
<td>4.33e-03</td>
<td>2.42e-01</td>
<td>Up</td>
</tr>
<tr>
<td>Blood</td>
<td>Cortisol</td>
<td>8.98e-03</td>
<td>5.03e-01</td>
<td>9.43e-02</td>
<td>5.28e+00</td>
<td>Down</td>
</tr>
<tr>
<td>Blood</td>
<td>Progesterone</td>
<td>1.07e-04</td>
<td>5.99e-03</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Blood</td>
<td>11-Deoxycorticosterone</td>
<td>1.24e-04</td>
<td>6.94e-03</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Blood</td>
<td>11-Deoxycorticosterone</td>
<td>8.59e-04</td>
<td>4.81e-02</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Blood</td>
<td>Corticosterone</td>
<td>1.94e-03</td>
<td>1.09e-01</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Reaction</td>
<td>HSD3β β</td>
<td>7.44e-04</td>
<td>4.17e-02</td>
<td>8.93e-04</td>
<td>5.00e-02</td>
<td>Up</td>
</tr>
<tr>
<td>Reaction</td>
<td>17-Hydroxylase β</td>
<td>9.62e-07</td>
<td>5.39e-05</td>
<td>2.05e-02</td>
<td>1.15e+00</td>
<td>Up</td>
</tr>
<tr>
<td>Reaction</td>
<td>11β-HSD1 (B→DOC·A)</td>
<td>3.23e-08</td>
<td>1.81e-06</td>
<td>2.05e-02</td>
<td>1.15e+00</td>
<td>Up</td>
</tr>
<tr>
<td>Reaction</td>
<td>11β-HSD1 (F→E)</td>
<td>1.70e-05</td>
<td>9.52e-04</td>
<td>9.43e-02</td>
<td>5.28e+00</td>
<td>Up</td>
</tr>
<tr>
<td>Reaction</td>
<td>21-Hydroxylase β</td>
<td>3.21e-06</td>
<td>1.80e-04</td>
<td></td>
<td></td>
<td>Down</td>
</tr>
<tr>
<td>Reaction</td>
<td>17,20-Lyase α</td>
<td>1.10e-05</td>
<td>6.16e-04</td>
<td></td>
<td></td>
<td>Up</td>
</tr>
<tr>
<td>Reaction</td>
<td>17-Hydroxylase α</td>
<td>4.31e-03</td>
<td>2.41e-01</td>
<td></td>
<td></td>
<td>Up</td>
</tr>
</tbody>
</table>

Table 3-3: Most significant differences between control and obese subjects after normalization for gender and age. Results are sorted according to TNoM p-values. P-value correction was done by Bonferroni correction.
As expected of a complex disorder, analysis of obese against control children provided less dramatic results (Table 3-3, Figure 3-3). Still, several interesting patterns emerged.

We found an intriguing general decrease in the level of urinary glucocorticoid metabolites in a condition previously suggested to show high (Mårin et al., 1992) or normal circulating cortisol (Abraham et al., 2013). This discrepancy may generate a hypothesis to be tested of reduced cortisol breakdown in obesity, related to the suppression of cortisol metabolizing enzymes, as recently shown for critical illness (Boonen et al., 2013).

We also found down regulation of mineralocorticoids and their urinary metabolites, which may be related to the changing renin-angiotensin-aldosterone system in obesity (Thethi et al., 2012) and the enhanced mineralocorticoid receptor activity in obesity and metabolic syndrome, as reviewed (Briet and Schiffrin, 2011). The results suggest that mineralocorticoid relative insufficiency may be related to enhanced conversion to androgens by the 17-hydroxylase/17,20-lyase-δ4 complex, which may contribute to the well-known androgenization in obese women (Hirsch et al., 2012). The novel observation of relative 21-OHase insufficiency has been suggested in PCO syndrome (Pall et al., 2010), and turns out to be part of the obesity component.

Figure 3-3: (A) TNoM and (B) Student-TTest comparison between obese and control children.

Blood metabolite (left) and urinary end-products (right) of the expanded steroid pathway. Triangles correspond to unidirectional blood enzymatic reactions. Rhombs correspond to bidirectional blood enzymatic reaction, while small triangle in it indicates the direction used for estimation of product-substrate ratio. Color saturation corresponds to analysis p-values showing down regulation in CAH as compared to control (blue), and up-regulation (yellow); white - to insignificant difference; gray - non-evaluated metabolites and reactions.
Chapter 4. Discussion and Future directions

Despite the desire to completely embrace the presented topics, many directions remain open. Here we briefly describe potential extensions of the projects in this dissertation.

Section 4.1. Extending BioLego project

The modular approach proposed and implemented in BioLego project results in notably simplifying the construction of different fermentation scenarios as well as the evaluation of different participant organisms within the scope of existing scenarios.

There are several directions for potential extension of BioLego project. Among others, probably the most important one is the development of better integration of additional metabolic models into the framework. Currently, the model of each new organism is to be inserted in a semi-automatic manner as an offline preprocessing before running the simulations. This can be improved by providing an interface to receiving models externally mapped to our envelope metabolites. Such interface will open the opportunity for evaluation of fermentation potential using additional, custom and proprietary organism models.

Another important improvement direction in the abovementioned context is an integration of automatic model reconstruction algorithms, such as MIRAGE (Vitkin and Shlomi, 2012), as part of the BioLego framework. As a vision, biorefinery designers should be able to select any organism with available sequence data and perform the simulations.

A third important direction for extending BioLego framework is in providing the interface to create more complex fermentation scenarios. This can be achieved either in the form of drag-and-drop UI and/or in the form of purposefully designed description language, which will allow to specify the relative fermentation order of the selected organisms. One particularly interesting aspect of this direction is the extension of the proposed approach to support construction of fermentation pipes with ratios between the number of containers with first organism and the number of containers with the second
being different from 1:1 (e.g. three containers with Organism1 are followed by single container with Organism2, leading to ratio 3:1).

The support for this ratio aspect is particularly interesting even in the contexts of the 1:1 ratio scenarios, in cases where the first organism utilizes most of the media, leading to negligible and numerically unstable growth and production rates. For the serial two-step scenario, we can address this issue by changing the stoichiometry of the Organism1-to-Organism2 metabolic flow (Rule 1.2 and Rule 1.3 in Table 1-1) from 1:1 to the pre-selected amplification ratio $A1:A2$ and by changing the stoichiometry of the Organism2-to-TotalProduct (Rule 1.7 in Table 1-1) to the opposite ratio of $A2:A1$.

Forth, it will be useful to incorporate knockout-optimization algorithms, such as OptKnock (Burgard et al., 2003) and RobustKnock (Tepper and Shlomi, 2010). Currently, BioLego project does not directly address knockout optimization, rather provides an option to evaluate the system performance under preselected set of knockouts of interest (including the brute-force screening of the knock-out space).

Finally, addition extension of interest is to improve the quality of the predictions provided by BioLego framework. This can be achieved by introducing the aspect of time, which is particularly important in the industrial processes. From our perspective this means shifting from steady-state FBA optimizations performed now to the dynamic FBA optimizations, leveraging the framework proposed by Mahadevan (Mahadevan et al., 2002).


Section 4.2. Leveraging fitness assays for metabolic modeling

There are several potential directions that can leverage our understanding of fitness data and the methodology we propose here. The most obvious one is to extend the proposed methodology to include (together with gene expression profiles) additional types of data. Applying our method on combined fitness and expression ALS scores outperformed the applying the method based on each of the data types individually. Thus, we expect additional improvement from the integration of other types of data, such as protein-protein interactions, phylogenetic knowledge, gene-metabolome associations and others. Principle foundations for such integration of distinct types of data were demonstrated here on the example of gene-expression profiles.

Another important direction is to leverage the proposed methodology for a gap-filling in metabolic networks, thus addressing a major challenge in automatic construction of metabolic networks. Principle foundations for such an approach were proposed in the MIRAGE (Vitkin and Shlomi, 2012) algorithm, developed during my MSc research (Vitkin, 2011).

Section 4.3. Steroidogenesis analysis

The proposed urine-to-blood visualization approach should be tested in the context of additional metabolic pathways. This can be a beneficial general tool for studying metabolomic profiles for various medical conditions. The proposed peer-group normalization approach is potentially useful not only in the context of steroidogenesis data, as the correct elimination of confounding factors is vital in almost every data-driven study.
References


GLPK - GNU Project - Free Software Foundation (FSF).


Henry, C.S. et al. (2010) High-throughput generation, optimization and analysis of genome-scale
Home - LEGO.com.
Márín, P. et al. (1992) Cortisol secretion in relation to body fat distribution in obese


119


Appendix A. Abstracts of published manuscripts

Vitkin, E., Gillis, A., Polikovsky, M., Bender, B., Golberg, A., Yakhini, Z. “Distributed flux balance analysis simulations of serial two-organism biomass fermentation”. (Under preparation for Bioinformatics)

Motivation: Efficient and sustainable conversion of biomass into valuable products is a major challenge for science, governments and businesses worldwide. The composition of the feedstock biomass and the ability of microorganisms to efficiently ferment it are two most critical factors influencing the efficiency of biorefineries, which are facilities integrating all the aspects of biomass processing flow. Intelligent design that address both these factors greatly benefits from dedicated tools for computational simulation and computer-assisted optimization of the fermentation processes.

Results: We present BioLego Vn2.0 framework based on Microsoft Azure Cloud, which allows high-scale simulations of the efficiency of biomass fermentation in two-step processes. BioLego provides capabilities for simultaneous analysis of multiple fermentation scenarios, for comparison of fermentation potential of multiple feedstock compositions and for the resource intensive analysis of media and organism modifications. Our biological results include (1) exploratory search for optimal biomass utilization setup for three different types of corn biomasses – corn cobs, corn fiber and corn stover; (2) analysis of possible effects of changes in composition of K. alvarezi algal biomass on the ethanol production yield in the anaerobic two-step process (S. cerevisiae followed by E. coli); (3) analysis of impact on the estimated ethanol production yield of knocking out single organism reactions either in one or in both organisms in anaerobic two-step fermentation process of Ulva sp. into ethanol (S. cerevisiae followed by E. coli); and (4) comparison of the experimental measurements of ethanol fermentation efficiency with the efficiency predicted by BioLego system.

Availability: All the source code, installation instructions, simulation results and other Supplementary materials are freely available at: http://wassist.cs.technion.ac.il/~edwardv/STORAGE/biolego2_data/.

Background: Synthetic biology and related techniques enable genome scale high-throughput investigation of the effect on organism fitness of different gene knock-downs/outs and of other modifications of genomic sequence.

Results: We develop statistical and computational pipelines and frameworks for analyzing high throughput fitness data over a genome scale set of sequence variants. Analyzing data from a high-throughput knock-down/knock-out bacterial study, we investigate differences and determinants of the effect on fitness in different conditions. Comparing fitness vectors of genes, across tens of conditions, we observe that fitness consequences strongly depend on genomic location and more weakly depend on gene sequence similarity and on functional relationships. In analyzing promoter sequences, we identified motifs associated with conditions studied in bacterial media such as casaminos, D-glucose, Sucrose, and other sugars and amino-acid sources.

We also use fitness data to infer genes associated with orphan metabolic reactions in the iJO1366 E. coli metabolic model. To do this, we developed a new computational method that integrates gene fitness and gene expression profiles within a given reaction network neighborhood to associate this reaction with a set of genes that potentially encode the catalyzing proteins. We then apply this approach to predict candidate genes for 107 orphan reactions in iJO1366. Furthermore – we validate our methodology with known reactions using a leave-one-out approach. Specifically, using top-20 candidates selected based on combined fitness and expression datasets, we correctly reconstruct 39.7% of the reactions, as compared to 33% based on fitness and to 26% based on expression separately, and to 4.02% as a random baseline. Our model improvement results include a novel association of a gene to an orphan cytosine nucleosidation reaction.

Conclusion: Our pipeline for metabolic modeling shows a clear benefit of using fitness data for predicting genes of orphan reactions. Along with the analysis pipelines we developed, it can be used to analyze similar high-throughput data.

Displacing fossil fuels and their derivatives with renewables, and increasing sustainable food production are among the major challenges facing the world in the coming decades. A possible, sustainable direction for addressing this challenge is the production of biomass and the conversion of this biomass to the required products through a complex system coined biorefinery. Terrestrial biomass and microalgae are possible sources; however, concerns over net energy balance, potable water use, environmental hazards, and uncertainty in the processing technologies raise questions regarding their actual potential to meet the anticipated food, feed, and energy challenges in a sustainable way. Alternative sustainable sources for biorefineries are macroalgae grown and processed offshore. However, implementation of the offshore biorefineries requires detailed analysis of their technological, economic, and environmental performance. In this chapter, the basic principles of marine biorefineris design are shown. The methods to integrate thermodynamic efficiency, investment, and environmental aspects are discussed. The performance improvement by development of new cultivation methods that fit macroalgae physiology and development of new fermentation methods that address macroalgae unique chemical composition is shown.


Marine macroalgae are potential sustainable feedstock for biorefinery. However, this use of macroalgae is limited today mostly because macroalgae farming takes place in rural areas in medium- and low-income countries, where technologies to convert this biomass to chemicals and biofuels are not available. The goal of this work is to develop models to enable optimization of material and exergy flows in macroalgal biorefineries. We developed models for the currently widely cultivated red macroalgae Kappaphycus alvarezii being biorefined for the production of bioethanol, carrageenan, fertilizer, and biogas. Using flux balance analysis, we developed a computational model that allows the prediction of various fermentation scenarios and the identification of the most efficient conversion of K. alvarezii to bioethanol. Furthermore, we propose the potential implementation of these models in rural farms that currently cultivate Kappaphycus in Philippines and in India.
Understanding the impact of all process parameters on the efficiency of biomass hydrolysis and on the final yield of products is critical to biorefinery design. Using Taguchi orthogonal arrays experimental design and Partial Least Square Regression, we investigated the impact of change and the comparative significance of thermochemical process temperature, treatment time, %Acid and %Solid load on carbohydrates release from green macroalgae from Ulva genus, a promising biorefinery feedstock. The average density of hydrolysate was determined using a new microelectromechanical optical resonator mass sensor. In addition, using Flux Balance Analysis techniques, we compared the potential fermentation yields of these hydrolysate products using metabolic models of Escherichia coli, Saccharomyces cerevisiae wild type, Saccharomyces cerevisiae RN1016 with xylose isomerase and Clostridium acetobutylicum. We found that %Acid plays the most significant role and treatment time the least significant role in affecting the monosaccharides released from Ulva biomass. We also found that within the tested range of parameters, hydrolysis with 121 °C, 30 min 2% Acid, 15% Solids could lead to the highest yields of conversion: 54.134–57.500 gr ethanol kg\(^{-1}\) Ulva dry weight by S. cerevisiae RN1016 with xylose isomerase. Our results support optimized marine algae utilization process design and will enable smart energy harvesting by thermochemical hydrolysis.

Biomass to fuel programs are under research and development worldwide. The largest biomass programs are underway in the USA and in Europe. In the coming decades, however, developing countries will be responsible for the majority of increases in transportation fuel demand. Although the lack of existing large-scale infrastructure and primary resources preclude oil refining in the majority of developing countries, this provides an opportunity for the rapid implementation of small scale distributed
biorefineries to serve communities locally. Currently, most of bioenergy production is based on the terrestrial biomass, which, however, has a major limitation of land availability. In parallel vein, recent results indicate the potential of the marine biomass, seaweeds, potential as a next generation bioenergy crop that does not require arable land, drinking water and fertilizers used in the terrestrial agriculture systems. The use of seaweeds as a feedstock for bioenergy production is very limited today mostly because seaweed farming takes place in the remote rural area in the medium and low-income countries, where the processing technology is not available. Moreover, the efficiency of seaweed conversion into transportation biofuels is not optimized and thus the estimations of the productivity are not as well understood as in the case of the first and second generation bioenergy crops. In the first part of this paper, we report on the high-level exergy model of the biorefinery. This model allows to find the parts of the biorefinery that can be optimized in a way the whole system gets the largest benefit. In the second part, we report on the metabolic model of seaweed fermentation into bioethanol, a platform chemical that can be used for a chemical and biofuel industries. The models allow for a rapid prediction of various fermentation scenarios to identify the most efficient conversion process, given the chemical composition of the biomass. In the third part, we analyze the potential of bioethanol production from off-shore cultivated macroalgae from the *Kappaphycus* family. We show that using a two-step fermentation first by yeast, *Saccharomyces cerevisiae*, and second by a bacterium, *Escherichia coli*, it is possible to generate additional revenue streams to the rural farms that are involved in the seaweed cultivation in Philippines. We also show that using currently available cultivation and fermentation methods it is possible to supply 100% of Philippines current demand on the transportation bioethanol from the seaweed industry, if the cultivation area is increased only 3 times from its current size.


The composition of feedstock biomass and the selection of fermenting microorganisms are critical factors in biorefinery design. Feedstock biomass composition is constrained by local supply materials, but microorganism selection affords considerable flexibility. Once biomass feedstock is identified, biorefinery designers need to select optimal fermenting organisms. While fermentation by microorganism communities can increase the range of digested biomass compounds and can be more resistant to infections, it has intrinsic problems in the context of species competition, process design and modeling —
issues related to insufficient process control. Using a serial fermentation approach, we offset some of these issues to allow maximal process control, while benefiting from organism diversity to maximize feedstock conversion rates. Here, we describe BioLEGO, a freely available web-based application that enables computer-assisted a single and two-step multiorganism fermentation process design. BioLEGO is based on a modular modeling approach, enabling the generation of different fermentation configurations consisting of independent organism modules. BioLEGO supports the evaluation of possible biomass-to-product yields for biomass mixes or general media and recommends media changes to increase the process efficacy.


Biomass to fuel programs are under research and development worldwide. The largest biomass programs are underway in industrialized countries. In the coming decades, however, developing countries will be responsible for the major increase in transportation fuel demand. Although the lack of existing large-scale infrastructure and primary resources preclude oil refining in developing countries, this provides an opportunity for the rapid implementation of small-scale distributed biorefineries to serve multiple communities locally. The principles for biorefinery design, however, are still in their infancy. This review sets a precedent in combining thermodynamic, metabolic, and sustainability analyses for biorefinery design. We exemplify this approach through the design and optimization of a marine biorefinery for an average town in rural India. In this combined model, we include sustainability and legislation factors, intensive macroalgae Ulva farming, and metabolic modeling of the biological two-step conversion of Ulva feedstock by a yeast (Saccharomyces cerevisiae), and then by a bacterium (Escherichia coli), into bioethanol. We hope that the model presented here will be useful in considering practical aspects of biorefinery design.

Also published as a selected paper with significant impact in Diabetes and Obesity section of World Biomedical Frontiers. (2015).

Traditional interpretation of GC–MS output involved the semi-quantitative estimation of outstanding low or high specific metabolites and the ratio between metabolites. Here, we utilize a systems biology approach to steroid metabolomics of a complex steroid-related disorder, using an all-inclusive analysis of the steroidal pathway in the form of a subject steroidal fingerprint and disease signature, providing novel methods of normalization and visualization.

The study compares 324 normal children to pure enzymatic deficiency in 27 untreated 21-hydroxylase CAH patients and to complex disease in 70 children with obesity. Steroid profiles were created by quantitative data generated by GC–MS analyses. A novel peer-group normalization method defined each individual subject’s control group in a multi-dimensional space of metadata parameters. Classical steroid pathway visualization was enhanced by adding urinary end-product sub-nodes and by color coding of semi-quantitative metabolic concentrations and enzymatic activities.

Unbiased automated data analysis confirmed the common knowledge for CAH – the inferred 17-hydroxyprogesterone was up-regulated and the inferred 21-hydroxylase enzyme activity was down-regulated. In childhood obesity, we observe a general decrease of both glucocorticoid and mineralocorticoid metabolites, increased androgens, up-regulation of 17,20-lyase, 17-OHase and 11β-HSD1 activity and down-regulation of 21-OHase enzymatic activity.

Our study proved novel normalization and visualization techniques are to be useful in identifying subject fingerprint and disease signature in enzymatic deficiency and insufficiency, while demonstrating hypothesis generation in a complex disease such as childhood obesity.
Technion - Computer Science Department - Ph.D. Thesis  PHD-2018-12 - 2018
The thesis provides an innovative approach to the BioLego framework, which is based on Microsoft Azure. The enhancement includes the development of a new model for the hotel industry, which is then applied to the hotel industry. The approach is based on the concept of a virtual hotel, which is designed to improve the customer experience. The thesis includes a detailed analysis of the current hotel industry, as well as a comparison with other industries. The thesis concludes with a discussion of the future of the hotel industry, as well as recommendations for further research.
הmaktור בוצע בהנהייתו של ד"ר ז正確 יוני
בפסקולות lemמי הקטשוב בסקסיים.

אני מודע לתכניות של התמיינה הנדיבה בשושתם
משתתפת.
היבטיםحلولים
בתחילה MEP-בניית
מיטול, אנליזה ורישום

תורור על מחקר

לשם מחמוד חלקלק שרדורישות ל렙ולחטאור
ודקאורי לפלדופאיה

 зрואד ריש poate

הלוגס לפוסט תכנית – מכון טכנולוגי לישראל
نصر transparent 2018

ודס

מלק החשיפה

Technion - Computer Science Department - Ph.D. Thesis PHD-2018-12 - 2018
היבטים יישוביים
בהכללות מתכלה
מיזוג, אנליזה ויישומים

אדריאד יישנא