Database for predicting a metabolic flux distribution within a cell

NOORLIN MOHD ALI[†1], KENTARO INOUE[†2], HIROYUKI KURATA[†1]

Abstract: The knowledge of metabolic components and their interactions is essential for modeling metabolic networks. This study develops the database for predicting metabolic flux distributions under different conditions and investigates the mechanism of how the flux distributions for various networks are accurately predicted. Two different sizes of metabolic networks, small-scale and medium-scale network, are tested to predict the flux distribution of *Escherichia coli* (*E.coli*) and *Saccharomyces cerevisiae* (*S.cerevisiae*).

Keywords: Database, metabolic flux

1. Introduction

Through many studies on microbes including *E.coli*, thousands of original literatures are published to produce lots of knowledge on metabolic systems. In understanding their functions, a metabolic network map becomes feasible to visualize the interaction among many biochemical reactions.

A metabolism can be regarded as a network interaction that generates energy (e.g.: ATP) and other building molecules for a cell to grow and produce metabolite compounds [1]. This involves integration of genome-scale knowledge at different levels, from genes to proteins and fluxes, as it may provide the fundamental mechanism of how individual components in the network system interact and affect the overall cell [2].

A network model denoted in a set of chemical reactions affect the prediction accuracy [1] and is further used to represent the qualitative information such as stoichiometry and elementary modes (EM) calculation.

This study aims to develop a database that will be used to predict the flux distributions under different conditions. In preparing a reliable dataset, various network models is implemented to investigate how these models accurately predict their flux distributions. In this study, two different sizes of metabolic network i.e. small scale and medium scale network are tested on several dataset of *Escherichia coli* (*E.coli*) and *Saccharomyces cerevisiae* (*S. cerevisiae*), by using Genetic Modification of Flux (GMF) [3, 4].

2. Method

2.1 The Metabolic Network model preparation

The network models are derived from the original publications, as shown in Table 1. The reversibility of each model is also set based on the original publication. The metabolic network is defined as:

- **Small scale model:**

The metabolic networks for *E.coli* and *S.cerevisiae* are reconstructed and analyzed by CADLIVE 2.75 [9] and CellNetAnalyzer (CNA) [8]. The reactions and metabolites are available in [4, 6]. The reactions of biomass formation are cited from references for *E.coli* [10] and *S.cerevisiae* are from [11].

E.coli small scale model is accomplished by most frequently encountered pathways: Glycolysis (14 reactions), pentose phosphate (8 reactions), Entner-Doudoroff (ED) (2 reactions), TCA Cycle (12 reactions), Pyruvate (4 reactions), Respiration (13 reactions). This model comprises of 53 reactions number.

S.cerevisiae small scale model: Glycolysis (8 reactions), pentose phosphate (8 reactions), TCA Cycle (10 reactions), Pyruvate metabolism (10 reactions), Serine Biosynthesis (3 reactions), Glycerol (2 reactions), Respiration (2 reactions). The network comprises of 43 reactions number.

- **Medium scale model:**

The detail network model is available in [7]. The metabolic network models of *E. coli* are revised from the model registered in CNA. There are 149 reactions and 30,579 EMs in *E.coli* metabolic network. While 106 reactions and 136,086 EMs in *S. cerevisiae* the metabolic network, including central carbon metabolism with amino acid syntheses.

E.coli medium-scale model is accomplished by: Glycolysis (17 reactions), pentose phosphate (8 reactions), Entner-Doudoroff (ED) (2 reactions), TCA Cycle (13 reactions), Pyruvate (5 reactions), Respiration (10 reactions), Serine Biosynthesis (6 reactions), Cysteine Biosynthesis (2 reactions), Glutamate Biosynthesis (3), Branched Chain Amino Acid Biosynthesis (13), Arginine Biosynthesis (10), Threonine, Lysine, Methionine Biosynthesis (10), Aromatic Amino Acids (17), Histidine Biosynthesis (9), Valine, Leucine and Isoleucine Biosynthesis (2), Purine (3), Methylglyoxal (4), Alanne, Asparatate and Glutamate (1), Glycogen Biosynthesis (2), UDP-N-acetyl-D-glucosamine biosynthesis (2), UDP-D-glucuronate biosynthesis (1), CMP-KDO Biosynthesis (1), CDP-diacylglycerol (1), D-Glutamine and D-Glutamate (7).

S.cerevisiae medium-scale model comprises of: Glycolysis (11), *pentose phosphate* (8), TCA Cycle (13), Pyruvate metabolism (51), Respiration (6), Serine Biosynthesis (3), Pentose and glucuronate interconversions (12), Glycerol (2).

2.2 Data preparation

A number of *E. coli* and *S. cerevisiae* original publications are collected and pre-processed as in Table 1.

TABLE 1. Data collection from original publication

Enzyme	Number of files	
	Small Scale Model	Medium Scale Model
pykF	1	1
gnd	1	1
pgi	1	1
pck	3	3
lpdA	1	1
S. cerevisiae	3	3

† Department of Bioscience and Bioinformatics, Kyushu Institute of Technology
† Laboratory for Cellular Systems Modeling, RIKEN RCAI

ppc	1	1
zwf	3	3
sucA	1	1

B. Subtilis	1	1
anaerobic	10	10
aerobic	1	1
Total	31	31

2.3 Genetic Modification of Flux (GMF)

The network models are tested by Genetic Modification of Flux (GMF) [3] [4], elementary modes based algorithm. GMF is a combination of two algorithms; modified CEF (mCEF) and Enzyme Control Flux (ECF). The main purpose of this algorithm is to predict the flux distribution of genetically modifed mutants. To estimate the elementary mode coefficients (EMC), Maximum Entropy Principle (MEP) objective function is applied to GMF. The data and network prepared above is used to calculate EMs by CNA that embedded in GMF application. Figure 1 shows the GMF algorithm.

Input: Reactions and metabolic data model in Excel
1. Calculate EM using CellNetAnalyzer(CNA) [4]
2. GMF Module
  2.1 Optimize EMC using objective function
  2.2 Get mCEF ratio
    2.2.1 Calculate the mCEF of wild type
    2.2.2 Calculate the mCEF of mutant
  2.3 ECF calculates the flux distribution of mutant

Figure 1.
The GMF algorithm

3. Result and Discussion

We compared the performance of both models. Table 2 shows the prediction error of each dataset. The prediction error is evaluated by:

$$\text{Prediction error} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(v_{prediction_i} - v_{experimental_i}\right)^2} \quad \text{(i)}$$

TABLE 2. The prediction error of the employed models

Model	Condition	Prediction error	
		Small scale Model	Medium scale model
E.coli (*pykF* knockout)	Batch: 5H	6.61	3.14
	Batch: 6H	3.73	0.97
	Batch: 7H	2.54	0.93
S. cerevisiae	μ: 0.15h^{-1}	20.17	1.07
	μ: 0.30h^{-1}	4.27	2.04
	μ: 0.40h^{-1}	8.84	3.66

Table 2 summarized the performance of six (6) dataset of *E.coli* and *S.Cerevisiae*. The table shows a significant improvement of prediction performance when the data are tested on medium scale network model.

The medium scale network makes the prediction better than the small scale network. This result is consistent to a study conducted by [1] on highly connected nodes and lesser connectivity. The study found that, the highly connected nodes are more likely to receive a new connection than lesser connectivity. The highly connected nodes models also described able to make the key prediction because it is closely simulating the origin life.

In this study, small scale and medium scale network models are compared. On the initial analysis, it is shows that the involvement of amino acid reactions in medium scale network is more than in the small scale network. In addition, the co-factors involvement, i.e. nicotinamide-adenine dinucleotide (NADH) and nicotinamide-adenine dinucleotide phosphate (NADPH) are reduced in small scale network as compared to the large scale network. Furthermore, the information on growth and non-growth-dependent adenosine-5'-disphosphate (ATP) is relatively neglected in small scale network than the medium scale network.

4. Conclusion

A number of datasets that are originally collected from publications are pre-processed. These datasets are further investigated for the mechanism on how flux distributions are accurately predicted for various networks. Small and medium scale metabolic networks are tested to the *E.coli pykF* knockout mutant and *S.cerevisiae* using the GMF algorithm with MEP objection function. The prediction performance is improved when the medium scale network models is applied to the datasets.

References

1) Wagner, A., The Large-scale Structure of Metabolic Networks: A Glimpse at Life's Origin? Wiley Periodicals, Inc., 2003 8(1): p.15-19

2) Forster, J., et al., *Genome-scale reconstruction of the Saccharomycescerevisiae metabolic network.* Genome Research, 2003. **13**(2): p. 244-253.

3) Kurata, H., et al., *Integration of enzyme activities into metabolic flux distributions by elementary mode analysis.* BMC Systems Biology, 2007. **1**.

4) Zhao, Q. and H. Kurata, *Genetic modification of flux for flux prediction of mutants.* Bioinformatics, 2009. **25**(13): p. 1702-8.

5) Toya, Y., et al., *C-13-Metabolic Flux Analysis for Batch Culture of Escherichia coli and Its pyk and pgi Gene Knockout Mutants Based on Mass Isotopomer Distribution of Intracellular Metabolites.* Biotechnology Progress, 2010. **26**(4): p. 975-992.

6) Frick, O. and C. Wittmann, *Characterization of the metabolic shift between oxidative and fermentative growth in Saccharomyces cerevisiae by comparative C-13 flux analysis.* Microbial Cell Factories, 2005. **4**.

7) Zhao, Q.Y. and H. Kurata, *Use of maximum entropy principle with Lagrange multipliers extends the feasibility of elementary mode analysis.* Journal of Bioscience and Bioengineering, 2010. **110**(2): p. 254-261.

8) Klamt,S.:*Cell Net Analyzer,* http://www.mpi-magdeburg.mpg.de/projects/cna/cna.html/.

9) Kurata,H. *et al., Extended CADLIVE: a novel graphical notation for design of biochemical network maps and computational pathway analysis.* Nucleic Acids Res., 2007. **35**, e134.

10) Wiback,S.J. *et al., Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the Escherichia coli spectrum.* Biotechnol. Bioeng., 2004. **86**, 317–331.

11) Förster,J. *et al., A functional genomics approach using metabolomics and in silico pathway analysis*. Biotechnol. Bioeng., 2002. **79**, 703–712.