Identifying Minimal Genomes and Essential Genes in Metabolic Model Using Flux Balance Analysis

Abdul Hakim Mohamed Salleh[1], Mohd Saberi Mohamad[1],
Safaai Deris[1], and Rosli Md. Illias[2]

[1] Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
ahakim26@live.utm.my, {saberi,safaai}@utm.my
[2] Department of Bioprocess Engineering, Faculty of Chemical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
r-rosli@utm.my

Abstract. With the advancement in metabolic engineering technologies, reconstruction the genome of a host organism to achieve desired phenotypes for example, to optimize the production of metabolites can be made. However, due to the complexity and size of the genome scale metabolic network, significant components tend to be invisible. This research utilizes Flux Balance Analysis (FBA) to search the essential genes and obtain minimal functional genome. Different from traditional approaches, we identify essential genes by using single gene deletions and then we identify the significant pathway for the metabolite production using gene expression data. The experiment is conducted using genome scale metabolic model of Saccharomyces Cerevisiae for L-phenylalanine production. The result has shown the reliability of this approach to find essential genes for metabolites productions, reduce genome size and identify production pathway that can further optimize the production yield and can be applied in solving other genetic engineering problems.

Keywords: Metabolic engineering, minimal genome, essential genes, flux balance analysis, metabolites productions.

1 Introduction

Systems metabolic engineering has been recognized as a new paradigm for systematically designing novel strategies for improvement of microbial strain. This system level of understanding can be used to help researcher prioritize experimental projects to ensure efficiency in cost and time consumed. In silico metabolic engineering has enabled us to generate hypotheses and predictions systematically to ensure laboratory experiment can be conducted with prior knowledge for optimal results. The application of 'omics' data for metabolic analysis along with validation to experimental data can be used to evaluate the significance of the model [1].

Many approaches for optimizing microbial strains have been conducted using genome scale metabolic model of an organism. However these approaches did not

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 414–423, 2013.

utilize gene expression analysis to aid in their prediction. Numerous amount of research incorporate the genetic factors that contribute to the function of metabolic networks as proposed by Karp et al. [2] and Mlecnik et al. [3], but they can only identify groups of specified genes are important although only some genes within this known groups are contributing to the observe response. Probabilistic network models such as Markov Random Field [4] and Mixture Model on Graph [5] on the other hand able to confirm that the features to be logically connected within the metabolic network but an assumption has to be made that is the gene expression is discretely distributed. This may not correctly describe the underlying structure and mechanisms of the system.

In this research, we took vanillin production in S. cerevisiae as a case study to test our approach. S. cerevisiae is considered one of the backbones in metabolic engineering as it is widely used in many applications [6]. High worldwide consumptions of vanilla and laborious and time consuming process of harvesting the product has urge the researchers to find a better alternative of microbial host.

The next section of this paper will discuss about the methodology of this research which covers the processes involved in the approach and dataset used. Then it will be followed by experimental results obtained and discussions and finally conclusions which conclude the findings of this research.

2 Methodology

In this paper, we reduced the size of genomes by implementing gene deletion strategies which is not done by previous methods by assuming that smaller number of essential genes in genomes decreased the used of biochemical resources to produce metabolites thus a higher production of metabolites can be yield. In this research S. cerevisiae genome scale model (yeast.4.05.xml) [7] which consists of 1865 reactions and 1319 metabolites is used to show the enhancement of vanillin production.

Here we have chosen vanillin production to test our approach. The process of formation of vanillin is known as biotransformation of aromatic acids. The basic substrate that can be used to produce vanillin in S. cerevisiae is L-phenylalanine thus, we can say that the production of vanillin increased when the production of L-phenylalanine increased [8].

In the genome of an organism, essential genes are genes compulsory to be present and cannot be knockout as they would results in lower growth rates or exactly zero growth rates. One way to determine these genes is by conducting single gene knockout in Yeast metabolic model and determine based on the resulting growth rates of each knockouts. A series of single gene deletions were performed using the model in order to determine the essential metabolic genes. Minimize the genome size with the assumption that smaller genome size will have less competing production to vanillin.

However, results derived from single gene knockout analysis would not sufficient for us to determine the effect of gene deletion in whole genome. This is because the numbers of genes are large and to perform combinations of multiple gene knockout would take a considerably huge amount of time to complete. By assuming that the genome consist of essential and non-essential genes for a particular process, we can deduce that the genome remain functional as long as the essential genes still exist and those that are not can be neglected or taken out of the system.

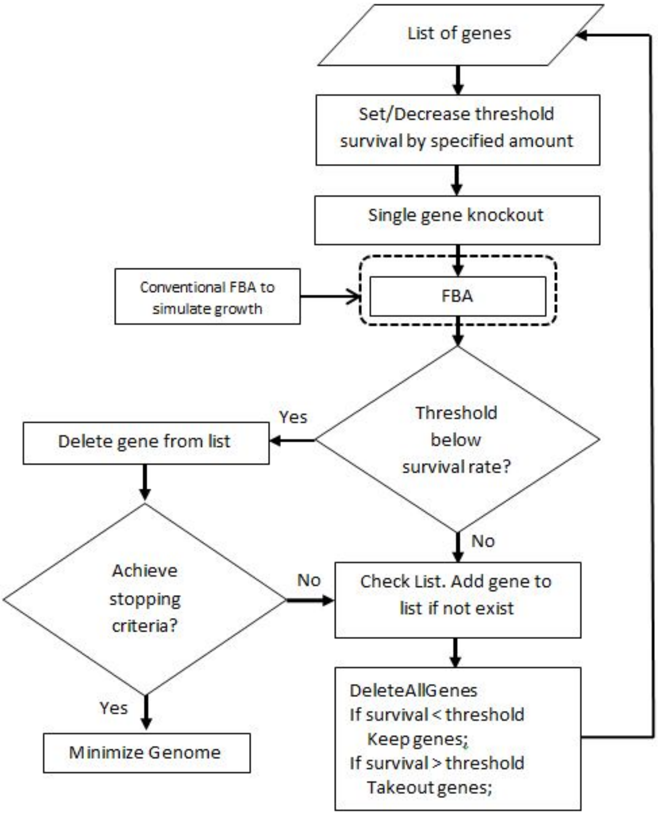
List of genes
Set/Decrease threshold survival by specified amount
Single gene knockout
Conventional FBA to simulate growth
FBA
Threshold below survival rate?
Yes
Delete gene from list
No
Achieve stopping criteria?
No
Check List. Add gene to list if not exist
Yes
Minimize Genome
DeleteAllGenes
If survival < threshold
Keep genes;
If survival > threshold
Takeout genes;

Fig. 1. General framework to minimize model

One of the strategies to address this is to reduce the size of genomes by finding the minimal number of components without reducing the significant functional capabilities. However, this reduction strategy is affected by the order of the genes that have been deleted. For instance, if a particular gene is deleted, it may have caused few other genes that are initially considered unessential to become essential and vice-versa. Another problem to address is the final growth rate. The resulting minimal genome will depend on percentage of final growth rate required, the larger the percentage, the size of minimal genome can afford to be quite large. Figure 1 shows the framework to obtain essential genes and minimal genomes.

It starts with a random model gene by deleting the gene and assessing the resulting growth rate. If the growth rate is considered acceptable (survival rate greater than threshold), the gene is permanently be deleted or otherwise it is placed back into the model. Then, new random gene is selected. When the resulting growth rate is calculated using FBA (dashed box in Figure 1), it is not only assessing the impact of this one gene deletion. It assesses the impact of all previously permanently deleted genes. The same process is repeated until all genes have been accounted. Next, the process is repeated again but with a lower threshold. The deletion cycle continue until the final growth rate is reached.

Next, we utilize KEGG database as our main reference for our pathways that is going to be extracted using microarray gene expression data. This experiment is based on the framework proposed by Hancock et al. [9] where more detailed explanation can be seen. In the initialization phase, the pathway structure is define where each gene is defined as node in the network and annotated by its gene code (*G*), reaction (*R*) and KEGG pathway membership (*P*). On the other hand, the edges that connect the nodes are identified as first substrate compound (C_F); the product compound of first reaction (C_M); final product compound (C_T) and the final KEGG pathway membership of C_T, (*P*) as in Eq. (1).

$$nodes = (G, R, P); edges = (C_F, C_M, C_T, P) \tag{1}$$

In Eq. (2), the probability of *y*, a binary response variable given that *X*, which is a binary matrix where the columns represent genes, the rows represent a pathway, and value of one indicates that the particular gene is included within specific path is defined that consist of two parts. First is the sum of probability π_m, which is the probability of each component with *y* given that *X* with βm parameter and second, product of $p(g_k, label_k | g_{k-1}; \theta_{km})$, which is the probability of path travers on edge $label_k$. g_k denotes the current gene and next gene in sequence, g_{k+1} where $label_k$ is the edge annotation. The result of this 3M (Markov Mixture Model) is *M* components defined by $\theta m = \{\theta_{sm}, [\theta_{2m},..., \theta_{tm},..., \theta_{Tm}]\}$. The θ_m is probabilities of each gene clustered within each component and indicate the importance of the genes. The parameters π_m, θ_{km} and βm are estimated simultaneously with an EM algorithm where more detailed explanation are discussed by Hancock and Mamitsuka [10].

$$p(y \mid X) = \sum_{m=1}^{m} \pi_m p(y \mid X, \beta_m) \prod_{k=2}^{k} p(g_k, label_k \mid g_{k-1}; \theta_{km}) \tag{2}$$

By using set of genes that involved in the particular pathway, *p-values* for each pathway are calculated using the hypergeometric distribution by summation of binomial coefficient. If the whole genome has a total of (*m*) genes, of which (*t*) are involved in the pathway under investigation, and the set of genes submitted for analysis has a total of (*n*) genes, of which (*r*) are involved in the same pathway, (*x*) is the number of pathway that have been chosen. Then the *p-value* can be calculated to evaluate enrichment significance for that pathway by Eq. 3:

$$p = 1 - \sum_{x=0}^{r-1} \frac{\binom{t}{x} \binom{m-t}{n-x}}{\binom{m}{n}} \tag{3}$$

FBA is a widely used and basic approach of constraints-based flux analyses and have shown to be successful for predicting growth, uptake rates and by-product secretion [11] without requiring the knowledge of metabolite concentration or the enzyme kinetics details of the system. FBA uses Linear Programming (LP) to maximize an objective function under different constraints. For example, to optimize an objective function denotes by *Z* at a particular period of time, *c* and *v* is reaction involved (*e.g* growth) typically the LP is formalized as in Eq. (4) and (5):

$$Z = \sum_{i=1}^{r} c_i v_i \tag{4}$$

$$S \,.\, v = 0,\ v_{min} \leq v \ \leq v_{max}. \tag{5}$$

Mass balance constraints are imposed by a system of linear equation, where stoichiometric *S* is an *m x n* matrix where *m* is the number of metabolites, and *n* is the number of reactions. *vmin* and *vmax* are set as lower and upper bounds on flux values that impose thermodynamic constraints that restrict directional flow of reaction, and capacity constraints. Using the minimal genome that consists of essential genes earlier and also the pathway membership, we initialized the stoichiometry matrix based on both of the results obtained. Using flux balance analysis we then calculate the optimization for our objectives function that is L-phenylalanine production and growth rate.

3 Experimental Results and Discussion

The experiment is conducted using glucose minimal media. The results obtained shows that the minimal genomes has an average size of 300 genes, approximately 30% of the original size. After running the experiment with threshold of original growth rates, there are about 130 genes of minimal genome are detected as essential genes out of the original 924 genes that produce growth rates of 1.3276 mmol gDW^{-1} hr^{-1}. Figure 2 shows the number of genes for 10 runs compared to the original number for L-phenylalanine production.

Logically, since only single gene deletion is performed we could not entirely conclude that the genome will survive and operates with only these genes because the number is considerably small compared to the original number of genes. Furthermore, the knockout process is dependent on just a single gene and the sequence of the genes. For example, if single gene A or B is knocked out, the cell may still survive but what if both of them are knocked out. Hence, the numbers of combinations are big.

Therefore, the size and contents of the minimal genomes might be varied depending on which genes are deleted first. Another factor to consider is the final growth rate. If the desired growth rate is 90% of it was originally, the resulting minimal genome may be quite large. If much smaller growth rate is desired, then the minimal genome can afford to become much smaller. Another factor is the growth medium. Highly enriched media will aid in achieving a smaller minimal genome.

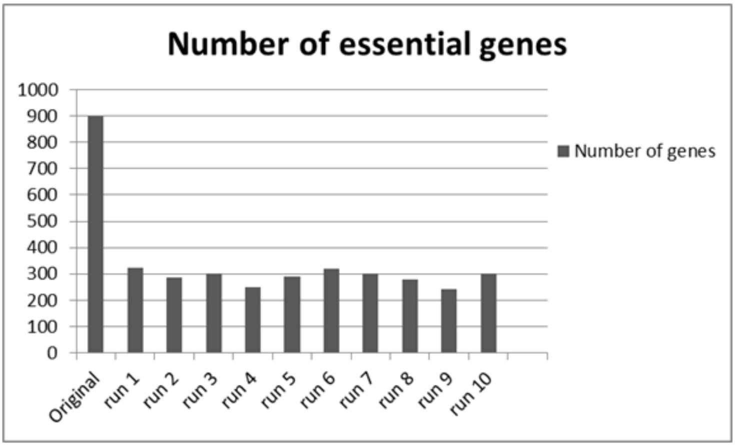

Fig. 2. Number of genes in minimize genomes

Existence of a large number of flux routes or pathways in genome-scale metabolic models requires the use of optimization or computational methods to predict the alternative routes consist of essential genes and deletion of genes in which help to improve the production. FBA is guaranteed to produce optimal results but not necessarily unique due to the existence of a large number of pathways involves [12]. With the essential genes obtained we extract significant pathways that lead to the production of L-phenylalanine using microarray gene expression data. Figure 3 shows the pathway extracted that consist of compound names as the nodes and KEGG reaction numbers for L-phenylalanine.

It is clear from set of compounds that made up the pathway, the highest path probability would be the transition and conversion alpha-D-glucose as the source moving towards the whole glycolysis pathway to produce pyruvate, CO_2 and Urea. Pyruvate plays a critical role in balancing between fermentation and respiration and also a potential intracellular indicator for limitation of glucose [13]. Then NH_3 is formed which leads to production of L-tyrosine and L-phenylalanine as final products.

Using set of genes that involve in the pathway we calculate the p-value for the whole genome to investigate furthermore which metabolism are actually contributing to the metabolite production. The *p-value* obtained can be used to measure the gene membership in the pathway. Table 1 shows the top 5 pathways correspond to that particular set of genes.

From the table it is obvious that Phenylalanine metabolism pathway has the lowest *p-value* with the highest gene ratio indicating the significant of the pathway with the gene set produce by the experiment. The pathways are considered to be highly statistically significant if having $p < 0.01$. This observation is probably caused by the production of L-phenylalanine and vanillin itself is a part of the component of the metabolism system therefore more number of genes is detected within this particular metabolism.

Figure 4 shows the result of glucose uptake rate effect towards the growth rate for both, new model (solid line) and original (dashed line). At the initial stage, with glucose uptake of 0 mmol gDW^{-1} hr^{-1} the maximum possible growth rate is 0 hr^{-1}. At approximately 18 to 20 mmol gDW^{-1} hr^{-1} which is the biologically realistic uptake rate [14] we can see the the production of L- phenylalanine is slightly higher with 1.3276 mmol gDW^{-1} hr^{-1} compared to the original 1.1596 mmol gDW^{-1} hr^{-1}.

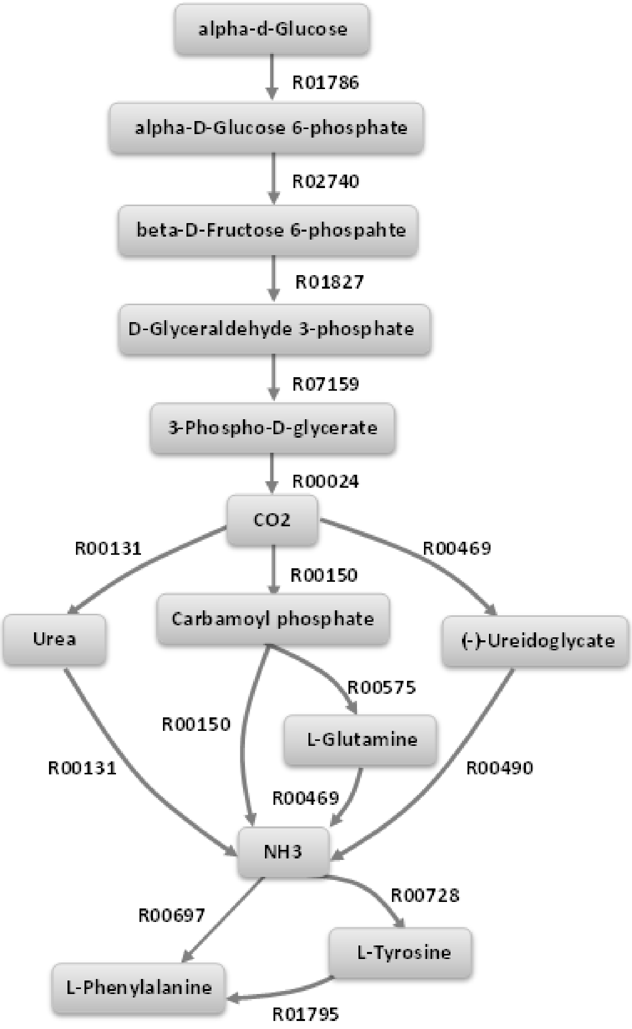

Fig. 3. Significant metabolic pathway for L-phenylalanine production based on KEGG

Table 1. The pathway membership for L-phenylalanine based on KEGG pathways

PATH	PATHWAY NAME	GENE RATIO	BACKGROUND RATIO	P-VALUE
00360	Phenylalanine metabolism	41/221	161/4377	1.11E-16
00010	Glycolysis / Gluconeogenesis	35/221	79/4377	4.44E-16
00020	TCA Cycle	21/221	58/4377	6.22E-15
00250	Alanine, aspartate and glutamate metabolism	14/221	35/4377	1.63E-14
00062	Arginine and proline metabolism	13/221	29/4377	1.89E-14

Normally, the biochemical production would increase along with cellular growth rate [15] hence, this indicate that the model able to survive and produce the desired products at optimal rate. Then the growth start to increase rapidly when enough glucose is available in the system meaning that the amount of ATP produce for growth has meet its requirement. After a certain period, the growth starts to increase less rapidly at one point until the end. This is due to the fact that at that particular point glucose is no longer the limiting factor for growth but instead its oxygen. In this condition the access glucose produce cannot be fully oxidize thus changing the flux to the production pathways.

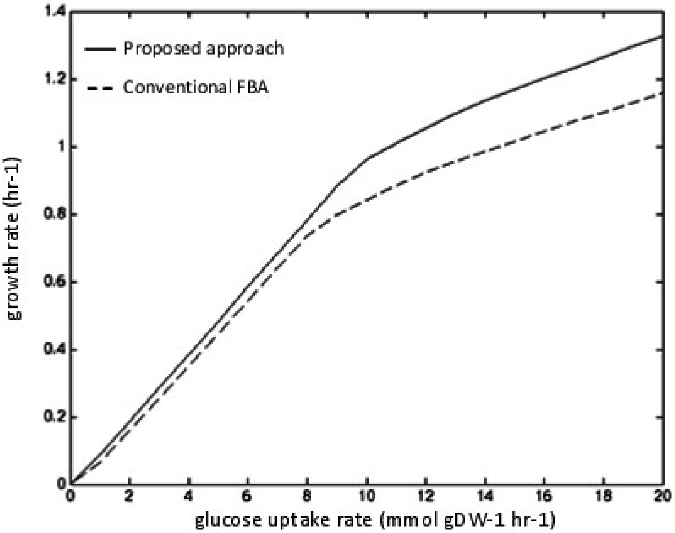

Fig. 4. Glucose uptake rate effect towards the growth rate for L-phenylalanine production

4 Conclusion

In this paper, we proposed an approach to identify the essential genes that able to form a minimal genome without degrading the biological function using FBA. Then, based on the essential genes obtained, we construct a metabolic pathway from gene expression data for a particular production of metabolites of interest. FBA is used to produce a fitness function with the assumption that the genome is in a steady state condition whereby optimization of the objective functions, in this case L-phenylalanine production can be conducted.

Based on the experiment conducted on S. cerevisiae for L-phenylalanine production, the results shown that the information provided by gene expression analysis has improve the prediction of constraint based analysis such as FBA and can potentially be extend. The integration of different data such as gene expression data, transcriptional regulatory and metabolic flux data has also shown to be successful in metabolic engineering for various purposes. Hence, the next big challenge would be integrating these models to a more biologically significant representation of these interrelated networks.

Acknowledgments. We would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grants (vot numbers: QJ130000.7123.00H67 and QJ130000.7107.01H29).

References

1. Edward, J.S., Ibarra, R.U., Palsson, B.O.: In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nature Biotechnology 19, 125–130 (2001)
2. Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., Caspi, R.: Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 11(1), 40–79 (2010)
3. Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F., Trajanoski, Z.: PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic Acids Research 33(1), 633–637 (2005)
4. Wei, Z., Li, H.: A markov random field model for network-based analysis of genomic data. Bioinformatics 23(12), 1537–1544 (2007)
5. Sanguinetti, G., Noirel, J., Wright, P.C.: Mmg: a probabilistic tool to identify submodules of metabolic pathways. Bioinformatics 24(8), 1078–1084 (2008)
6. Varges, F.A., Pizzarro, F., Perez-Correa, J.R., Agosin, E.: Expanding a dynamic flux balance model of yeast fermentaion to genome-scale. BMC Systems Biology 5, 75 (2011)
7. Mo, M.L., Palsson, B.Ø., Herrgård, M.J.: Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Systems Biology 3, 37–41 (2009)
8. Priefert, H., Rabenhorst, J., Steinbüchel, A.: Biotechnological production of vanillin. Appl. Microbiol. Biotechnol. 6, 296–314 (2001)
9. Hancock, T., Takigawa, I., Mamitsuka, H.: Mining metabolic pathways through gene expression. Gene Expression 26(17), 2128–2135 (2010)

10. Hancock, T., Mamitsuka, H.: A Markov Classification Model for Metabolic Pathways. In: Salzberg, S.L., Warnow, T. (eds.) WABI 2009. LNCS, vol. 5724, pp. 121–132. Springer, Heidelberg (2009)
11. Reed, J.L., Palsson, B.O.: Thirteen Years of Building Constraint-Based InSilico Models of Escherichia coli. J. Bacteriol. 185(9), 2692–2699 (2003)
12. Brochado, A.R., Matos, C., Moller, B.L., Hansen, J., Mortensen, U.H., Patil, K.R.: Improved vanillin production in baker's yeast through in silico design. Microbial Cell Factories 9, 84 (2010)
13. Boer, V.M., Crutchfield, C.A., Bradley, P.H., Botstein, D., Rabinowitz, J.D.: Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations. Mol. Biol. Cell 21(1), 198–211 (2010)
14. Orth, J.D., Thiele, I., Palsson, B.Ø.: What is flux balance analysis? Nature Computational Biology 28, 245–248 (2010)
15. Kim, J., Reed, J.: OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. BMC Bioinformatics 4(53), 1–19 (2010)