

Prediction of Vanillin Production in Yeast Using a Hybrid of Continuous Bees Algorithm and Flux Balance Analysis (CBAFBA)

Leang Huat Yin¹, Yee Wen Choon¹, Lian En Chai¹, Chuii Khim Chong¹,
Safaai Deris¹, Rosli M. Illias², and Mohd Saberi Mohamad^{1,*}

¹ Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia

{lhyin2, ywchoon2, lechai2, ckchong2}@live.utm.my,
{safaai, saberi}@utm.my

² Department of Bioprocess Engineering, Faculty of Chemical Engineering,
Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia
r-rosli@utm.my

Abstract. Most food and beverage is containing artificial flavor compound. Creation of artificial flavors is not an easy step and it is hardly ever completely effective. In this paper, we introduce an *in silico* method in optimization of microbial strains of flavor compound synthesis. Previously, there are several algorithms such as Genetic Algorithm, Evolutionary Algorithm, OptKnock tool and other related techniques are widely used to predict the yield of target compound by suggesting the gene knockouts. The used of these algorithms or tools is able to predict the yield of production instead of using try and error method for gene deletions. Nowadays, without using *in silico* method, the direct experiment methods are not cost effective and time consumed. As we know, the cost of chemical is expensive and not all flavorist able to afford the cost. However, the main limitations of previous algorithms are it failed to optimize the prediction of the yield and suggesting unrealistic flux distribution. Therefore, this paper proposed a hybrid of continuous Bees algorithm and Flux Balance Analysis. The target compound in this research is vanillin. The aim of study is to identify optimum gene knockouts. The results in this paper are the prediction of the yield and the growth rate values of the model. The predictive results showed that the improvement in term of yield which may help in food flavorings.

Keywords: Bees Algorithm, Flux Balance Analysis, Yeast, Optimization.

1 Introduction

Vanillin is normally used as food ingredient or flavor compound. In order to get vanillin by traditional method of obtaining vanillin is from cured seed pods of the Vanilla

* Corresponding author.

planifolia (natural vanillin) and via chemical synthesis. As we know, yeast is considered to be a workhorse of the biotechnology industry for the production of many value-added chemical, alcoholic beverages and biofuels [1]. Thus, in this research, *S. cerevisiae* model is used in vanillin. Figure 1 showed the *de novo* biosynthetic pathway in *S. cerevisiae* for vanillin production. In order to meet the aim of prediction of the vanillin yield by *in silico* method, the algorithm introduced in this paper is a hybrid of Continuous Bees Algorithm and Flux Balance Analysis which able to predict a set of gene knockouts. The contributions of this paper are three fold. First, up to our best knowledge, this method is first used in prediction of biochemical production where no other researchers used this method before. Second, this prediction algorithm implemented in this research is able to predict the gene knockouts in the large number of reactions in *S. cerevisiae* model. Third, the experimental results shown that the prediction algorithm in this research had given a set of relatedness deletion whereby the experimental technique in wet lab can be avoided before the expected result is confirmed. This will contribute in term of cost efficiency where the materials for experiment are expensive.

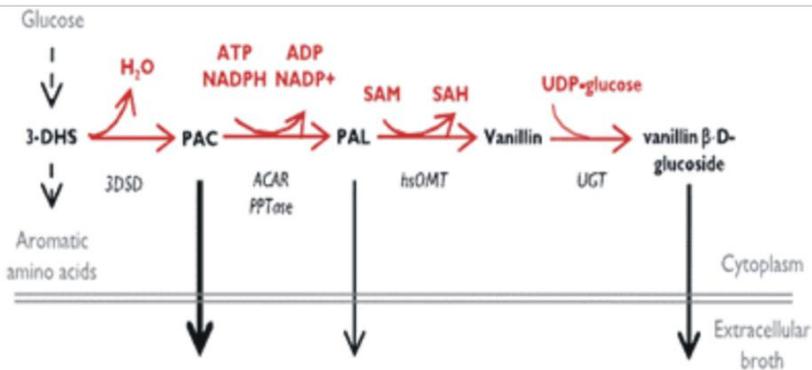


Fig. 1. The *de novo* biosynthetic pathway in *S. cerevisiae* for vanillin production.

Basically, the prediction of biochemical compounds is predicted by several algorithms such as Genetic Algorithm, Evolutionary Algorithm, OptKnock tool and OptGene which are widely used. Unfortunately, there is some limitation of those techniques. In this paper, the limitation of binOptGene is identified. In binOptGene, the representation of population is in binary variable where representation will form a set of “individual” representing a particular mutant. However, in this method the main problem is number of invalid individuals in population is larger and consequently negatively affects the convergence. It happened due to use of penalty functions after evaluation of individuals. Besides that, binOptGene also will suffer of several problems which causes by used of Genetic Algorithm in binOptGene or OptGene itself. One of the limitation of Genetic algorithm is stop criterion of the algorithm is not clear in every problem. In addition, it is tendency to converge towards local optima

rather than global optimum of the problem. In the measurement of the fitness in a single right/wrong problem, Genetic algorithm is failed to solve the problem efficiently.

In order to solve limitations, the usage of Bees algorithm is as an optimization algorithm. The Bees algorithm is known as a new population-based search algorithm [2]. This algorithm is able to search optimum solutions in large search space. However, in Bees algorithm the representation method of population is difference from binOptGene where it represented in integer number. In this way, number of genes to be deleted can be directly imposed by changing the size of the individuals. Besides that, Flux Balance Analysis (FBA) is used as an approach that widely used for studying and analyzing biochemical networks, in particularly the genome-scale metabolic network constructions that have been built in the past decade [3]. Flux Balance Analysis also known as a constraint-based modeling approach in which the stoichiometry of the underlying biochemical network constrains the solution [4]. Constraints applied in Flux Balance Analysis are represented in two (2) ways: Firstly, as equations that balance reactions input and secondly output and as inequalities that impose bounds on the system. Basically, this approach is used a mathematical modeling approach for analyzing the flow of metabolites in metabolic network.

2 A Hybrid of Continuous Bees Algorithm and Flux Balance Analysis (CBAFBA)

In this section, we describe the details of the proposed a hybrid algorithm, hybrid of CBAFBA. In CBAFBA algorithm, there are 3 main parts is explained in next subsections: initialization, neighborhood search, and assignment of the remaining bees for random search and obtained the solution of CBAFBA. The Figure 2 shows the flow chart of CBAFBA.

In next subsection, we describe the dataset used in this proposed. Measurement of the evaluation of the result obtained is the optimization of metabolic production method used to determine the growth rate is included in Flux Balance Analysis. The function is defines as below:

$$\begin{aligned} &\text{Maximize } Z \\ &\text{Subject to} \quad \sum_{j=1}^N S_{ij} v_j = 0, \quad i = 1, \dots, M \end{aligned} \tag{1}$$

Thermodynamic and capacity constraints can be added as below:

$$\alpha_{-}(j) \leq v_j \leq \beta_{-} j, \quad j = 1, \dots, N \tag{2}$$

The Z is the linear objective function which to be minimize or maximize from particular metabolic engineering design objective to maximization of cellular growth of vanillin. The v_j corresponds to the rate of reaction j and S_{ij} is the stoichiometric coefficient. The different optimal solution obtained when different objective function is applied in optimization function.

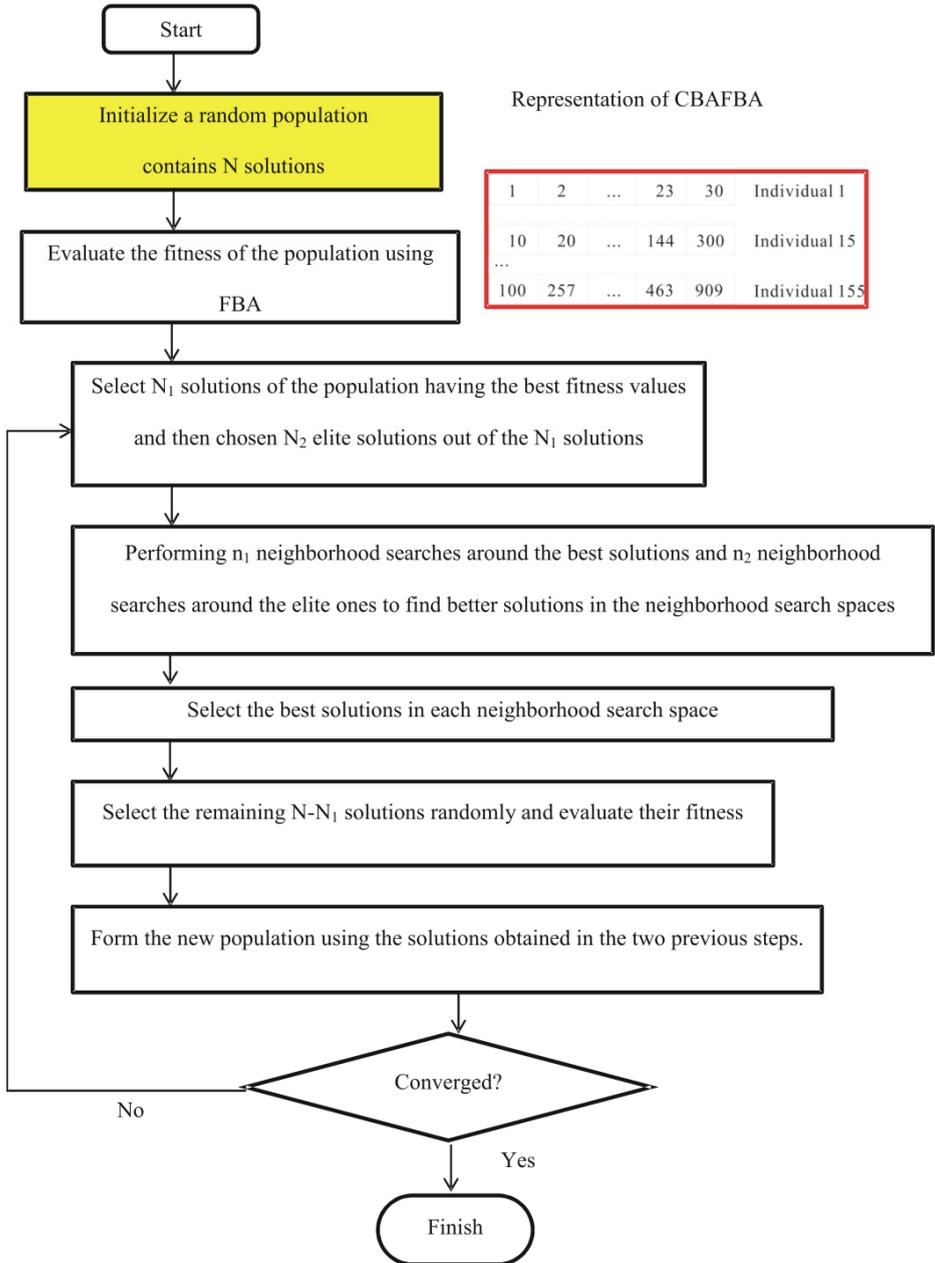


Fig. 2. The flow chart of CBAFBA

The difference of CBAFBA compare to binOptGene [5] and BAFBA [6] is the representation strategies in initialization of population. In binOptGene, the represented in binary which may cause the population become larger and consequently negatively affect the convergence. In CBAFBA, the population is represented in integer number [7]. The representation is showed in next section. Figure 3 shows the overall of OptGene algorithm which the representation of population is in binary variable. Besides that, the CBAFBA is able to search in global population where it formed a new population of each iteration had been completed. However, in binOptGene which applied Genetic algorithm is tendency to convert into local optimal rather than search for global optimal of the problem [8].

2.1 Initialization

Initialization of CBAFBA is the first step which used to create a random set of list which represent as population. The size of population can be set according to the size of dimension or search space needed. The bigger search space will cause the computational time increases. Therefore the search space is reducing by model pre-processing phase. In this initialization of population, the representation of individual is in integer number. The individuals are composed of integer numbers representing only the genes to be deleted. Therefore, it is based on the relative order in of metabolic model.

2.2 Neighborhood Search

In neighborhood search stage, there are 3 steps of Bees algorithm is executed. The selection of sites which has higher fitnesses, recruitment of bees for selected sites and selection of fittest bee from each search are the step in Bees algorithm. In order to select the sites with higher fitnesses, the sorting function is created whereby it sorts the fitnesses and positions of the population. The sorted list is in descending order.

After the population is sorted according to it fitness, the recruitment stage is begin. This recruitment stage is sending the bees around the fittest site and evaluated the fitness. The fitness of this research is based on the Flux Balance Analysis. In order to prevent very small values of the production at set growth rate, there is a comparison between minimum productions of population with a fixed minimum production.

2.3 Assignment of the Remaining Bees for Random Search and Obtained the Solution of CBAFBA

In this stage, CBAFBA is assigned the remaining bees for random search. The remaining bees are used to find the potential new solutions. The searched of potential new solutions is done around the search space. Again, in this stage the fitness by Flux Balance Analysis is used which will used the minimum production compare with fixed value. After the calculation of the fitness, the list of production is sorted in order to identify the best production. At final of each run, the prediction of the gene knock-out list will be generated after the CBAFBA algorithm is completely computed.

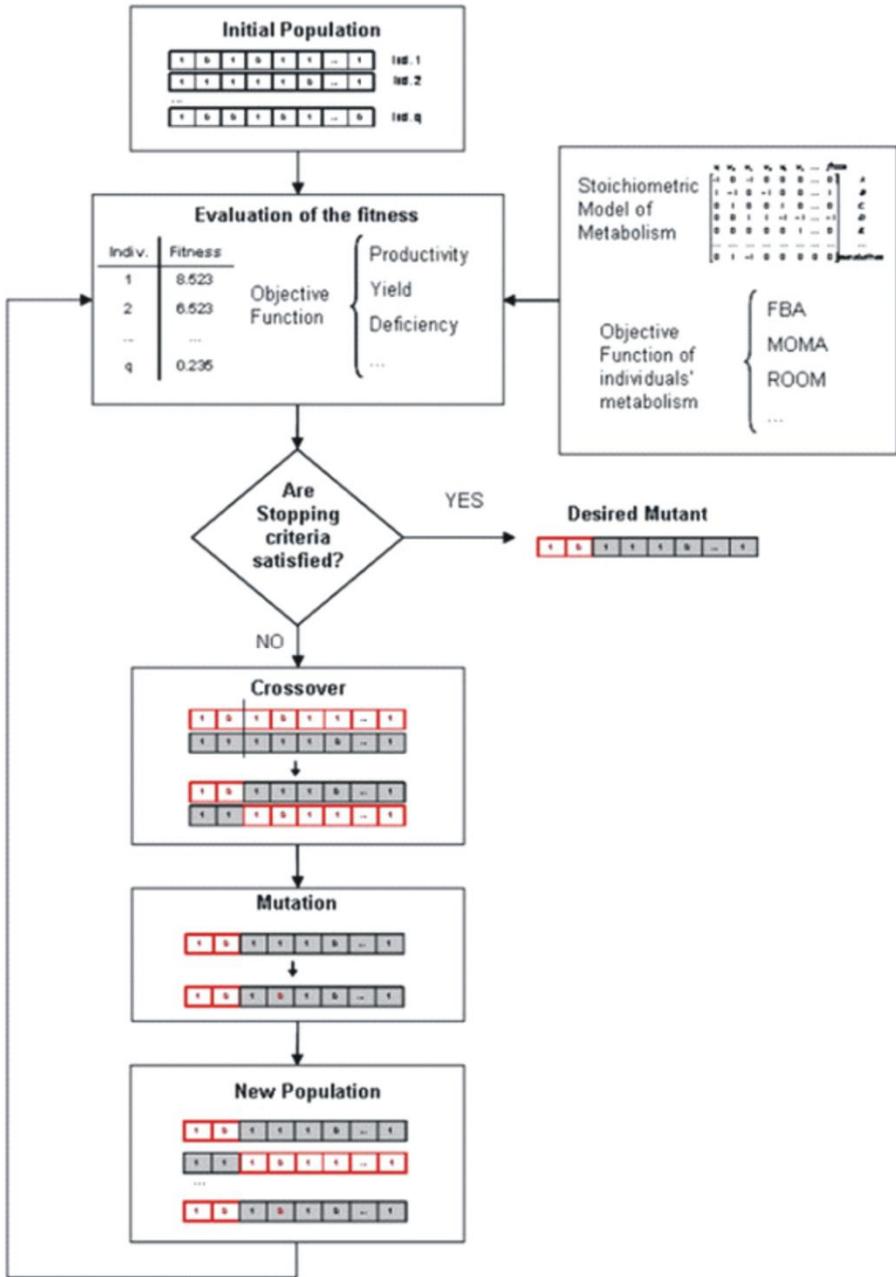


Fig. 3. The overall of OptGene algorithm

3 Results

3.1 Dataset

A model of *S. cerevisiae* dataset is used in the computational algorithms. Basically, *S. cerevisiae* is a type of baker's yeast. This dataset is originally obtained from Kyoto Encyclopedia of Gene and Genomes (KEGG). Yeast dataset from KEGG is then converted into System Biology Markup Language (SBML) format. In the model of *S. cerevisiae* dataset, all the pathways in Baker's yeast are included. From the abundant of pathways included in the dataset, several pathways is excluded or removed to reduce unrelated pathways and minimized the computational time during prediction process in on going. Thus, the model pre-processing is needed. The model pre-processing is include reduce dead-end reactions whereby problem size considerably small compare to initial model.

In this paper, the stoichiometric simulations provide an estimation of possible range of flux values for every reaction in the network. Due to the existence of a large number of alternatives flux routes or path-ways in genome-scale metabolic models require the use of optimization or computational methods to predict the alternative deletion of genes which will help to improve the production. The used of FBA are guaranteed to be optimal, but not necessarily unique due to the existence of a large number of pathways involves [8]. In this paper, the vanillin is the product to be predicted.

3.2 Vanillin Production

Selection of Target Reaction in Vanillin Prediction by Bees Algorithm and Flux Balance Analysis

Here, a core substrate which can contribute to vanillin production is L-phenylalanine which shows in Figure 4. In Figure 4, the formation or production of vanillin is known as biotransformation of aromatic acids. Generally, the production of vanillin increased when the production of L-phenylalanine increased. In the key reaction, phenylalanine is deaminated to transcinnamic acid, which catalyzed by phenylalanine ammonia lysase [9]. The transcinnamic acid then undergoes a chain reaction until reached the vanillin production which show in Figure 4.

In addition, the L-phenylalanine used as precursor to vanillin production in biosynthesis of alkaloid de-rived from shikimate pathway. The L-phenylalanine is synthesis from prehenate where prephenate dehydratase is an enzyme involved in the process. Next, the L-phenylalanine is involved in the synthesis of caffeoyl-CoA. There are series of process and enzymes involved in the synthesis of vanillin from caffeoyl-CoA. One of the enzymes involved is caffeoyl-CoA O-methyltransferase which catalyze the conversion of caffeoyl-CoA into feruloyl-CoA.

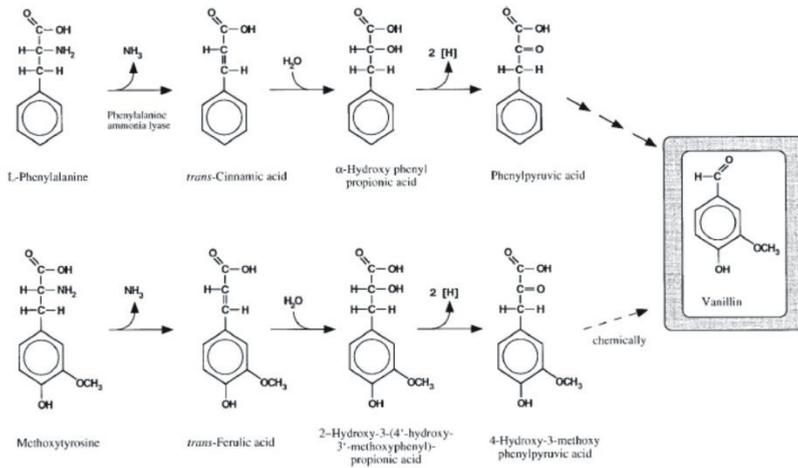


Fig. 4. Main contribution of vanillin production is L-phenylalanine

Therefore, the selection of synthesis of L-phenylalanine is contributed to vanillin production. In order to select the substrate and target reaction, glucose and prephenate dehydratase had been selected, respectively. The purpose prephenate dehydratase been chosen is due to this enzyme will affect the production of L-phenylalanine, a substrate to produce vanillin. Besides that, the purpose of glucose reaction been chose is due to the main substance of mostly biosynthesis in any pathways.

Selection of Gene Knockouts List Based on Growth Rate Prediction Using Flux Balance Analysis

Prediction strategies described in this work are based on the assumption that microbial cells would evolve in higher growth rates and biochemical production. As we knows, knockout mutants that force the coupling between biomass and biochemical production allow researcher to use growth rate as a selective pressure and find adaptively evolved strains with improved growth rates and production capabilities [10].

Figure 5 shows Average of frequency for predicted result cases in each growth rate values from same biochemical product which is vanillin. Based on that result, the higher growth rate is 1.7023 which indicates the cellular cell is alive and able to produce the desired product at optimal rate. Basically, the bio-chemical production would increase along with cellular growth rate [10].

On the other hand in this research, the minimum growth rate is $-5.4019\text{e-}013$ where this set of genes knockout is eliminate. The reason of the elimination is due to the cellular cell is unable to live or cell is death. The elimination is also because of the deletion of lethal gene. This deletion may cause the cellular cell die and fail to produce desire biochemical production.

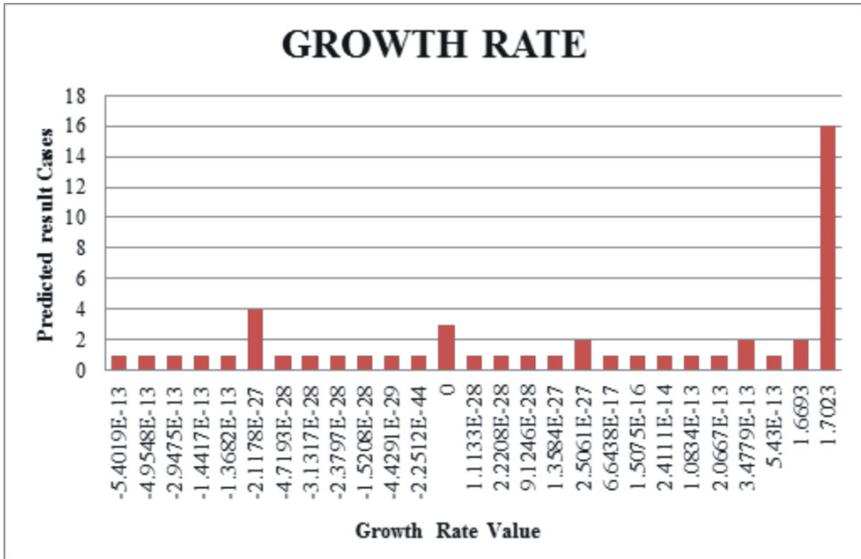


Fig. 5. Average of frequency for predicted result cases in each growth rate value

Selection of Genes Knockout List Based on Production Rate

In general, the knockout strategy consists of two approaches; reaction-based deletion and gene-based deletions. In this section, the gene-based deletion approaches are preferred compared to reaction-based deletion. As we know, the relationship between genes, proteins and reactions is not one-to-one. In other words, a metabolic reaction usually carries out by one or more enzymes where each of them can comprise to produce multiple gene products (proteins) [10]. The removal of multiple genes may affect the additional reactions by removal of additional reactions [10]. The worst scenario for removal of additional reactions is the reaction for desired product being removed incidentally.

After applying the proposed method in yeast model, the result shows that there are 3 mutants obtained. Table 1 summarizes three of the identified gene knockout strategies for L-phenylalanine (i.e. mutant A, B, and C). Based on the result obtained from CBAFBA, maximum yield is 0.19466 for three mutants in the phenylalanine target reaction. Here, the result for mutant A suggests that removal of shikimate pathway reaction from the network. In mutant A, ARO4 gene is being removed which is encoded into 3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase (DAHPS). Experimentally from web lab, DAHPS is used for condensation of erythrose-4-phosphate and phosphoenolpyruvate to 3-deoxy-d-arabino-heptulosonate-7-phosphate (DAHP) [11].

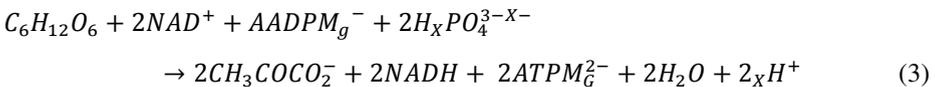
The removal in mutant A is less intuitive strategy which focuses on inactivating phosphoenolpyruvate (PEP) consuming reactions rather than eliminating competing

by product mechanism. With this strategy, some researches assuming that the maximum biomass yield could be attained. Note that the predicted yield in CBAFBA is assumed only by the theoretical maximum where further experiment from wet lab is needed for proofing purpose. Figure 6 reveals the flux distribution of the mutant A.

Table 1. List knockout for vanillin case study which control by L-phenylalanine compound

<i>L-Phenylalanine</i>			
Mutant	Gene	Knockouts	Enzyme
A	ARO4	PEP + D-erythrose 4-phosphate + H ₂ O = DHAP + phosphate	3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase
B	BDH1	(R,R)-Butane-2,3-diol + NAD ⁺ = (R)-Acetoin + NADH + H ⁺	(R,R)-butanediol dehydrogenase
C	ARO10	Ehrlich pathway	2-isopropylmalate synthase

Second gene deletion is removal of butanoate metabolism pathway whereby CBAFBA suggested deletion of (R,R)-butanediol dehydrogenase. The gene involved in encoding (R,R)-butanediol dehydrogenase in *S. cerevisiae* is BDH1. This enzyme is involved in the formation of (R)-acetoin from (R,R)-butane-2,3-diol. In order to obtain the (R)-acetoin in this reaction, the (R,R)-butanediol dehydrogenase is dependent on NAD⁺ which is the co-enzyme for butanoate metabolism [12]. Theoretically, this reaction is assumed to be affected by the glycolysis pathway in order to produce PEP. In the glycolysis pathway, NAD⁺ is used as a co-enzyme in the conversion of glyceraldehyde-3-phosphate into 3-phospho-D-glyceroyl-phosphate. The following balanced equation shows that the oxidation of glucose to pyruvate [13].



The strategy assumed by CBAFBA in mutant B shows that the production yield is 0.19466 and the growth rate is 1.7023 respectively after the BDH1 gene is deleted. Figure 7 shows the glycolysis pathway in *S. cerevisiae*.

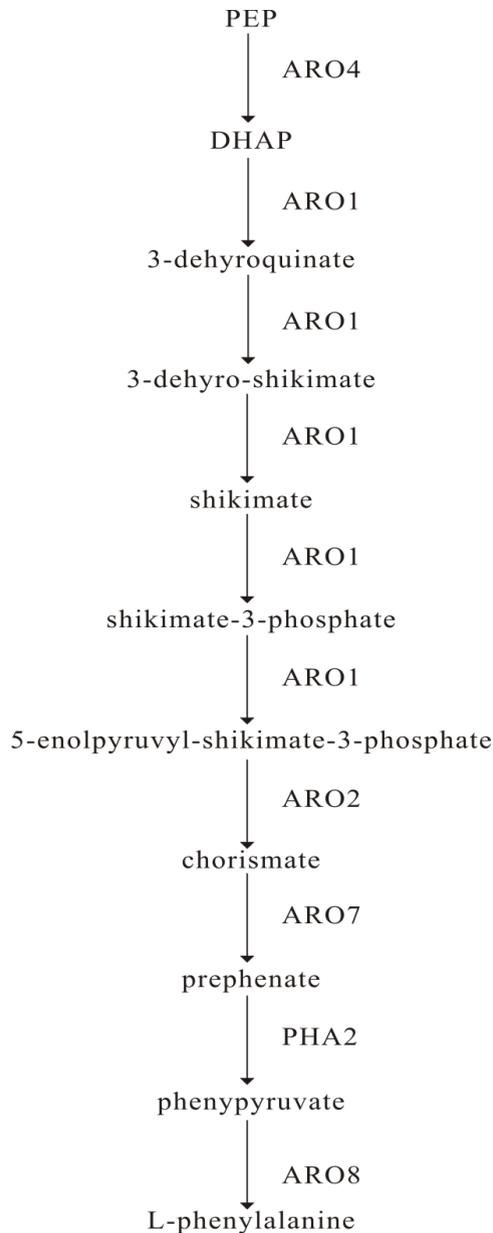


Fig. 6. The flux distribution of the mutant A

Third mutant (mutant C) suggests that deletion of ARO10 gene. The deletion of ARO10 gene is directly remove Ehrlich pathway of *S. cerevisiae*. The Ehrlich pathway is chemical reactions and pathways involved in the catabolism of amino acids to produce alcohols with one carbon less than the starting amino acid. The catabolism

process is involving the breaking down of molecules (amino acids) into smaller units (fusel alcohols). In *S. cerevisiae*, amino acids that assimilated by the Ehrlich pathway is taken up slowly throughout the fermentation time [14]. The amino acids usually taken up from Ehrlich pathway are valine, leucine, isoleucine, methionine, and phenylalanine. This will affect the production of the L- phenylalanine, where the yield will decrease if the Ehrlich pathway fails to be removed.

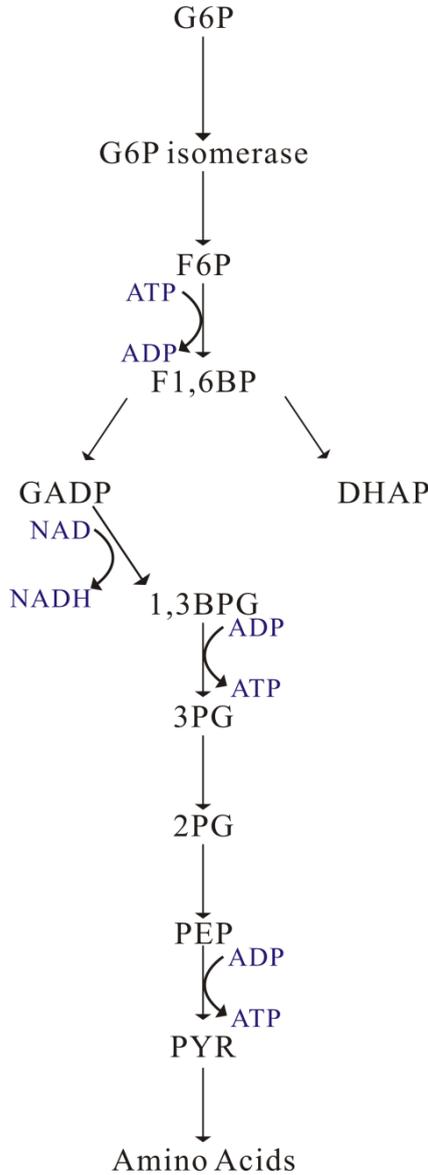


Fig. 7. The glycolysis pathway in *S. cerevisiae*

Besides that, in aerobic glucose-limited chemostat, phenylalanine is used as sole nitrogen sources, where phenylalanine is converted predominantly to fusel acids and only very low concentrations of fusel alcohols are formed [14]. However, without glucose-limited chemostat of *S. cerevisiae*, growth is pre-dominantly fermentative, and when phenylalanine is the sole nitrogen source, it is converted into a mix-ture of phenylethanol and phenylacetate [14]. This proves that the deletion of ARO10 is assumed to be increasing the phenylalanine yield. Figure 8 shows overall of the Ehrlich pathway.

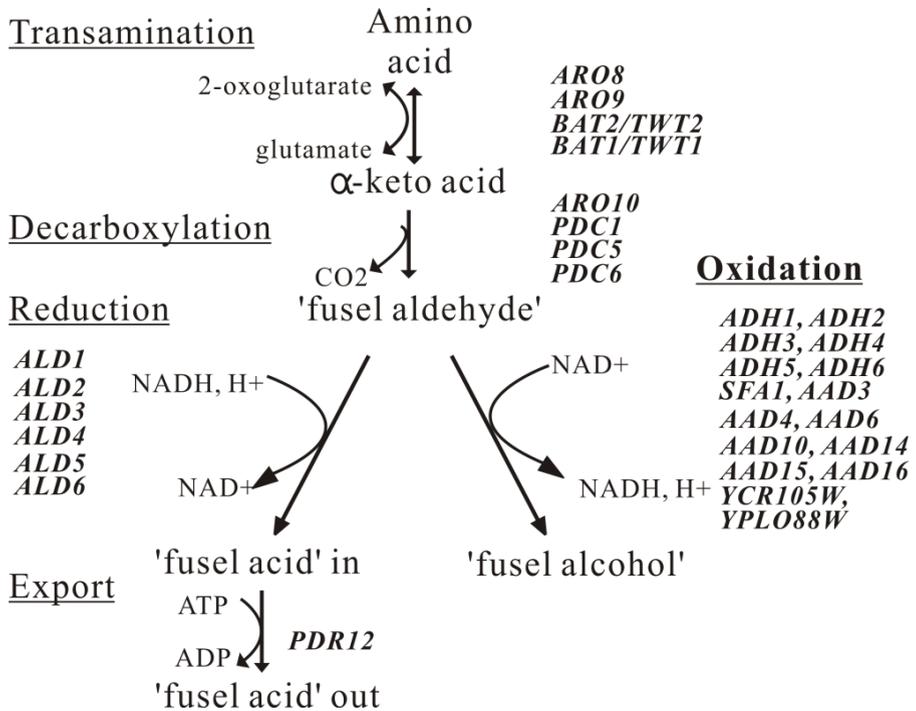


Fig. 8. Overall of the Ehrlich pathway

Table 2 shows the biomass overall yield of vanillin in batch cultivation after simulated by OptGene tool. The production of vanillin β -D-glucoside (VG) in table 2 shows the minimum yield (μ_{max}) is 0.10 (VG1) and the maximum yield is 0.2 (VG2). All stains (VG0, VG1, VG2, VG3, and VG4) are involved in either the removal or overexpression of pyruvate decarboxylase (PDC) and glutamate dehydrogenase (GDH). Based on Brochado and his colleagues, they believe that by in silico analysis, PDC was found as a target to increase formation of VG considering both respiratory and respire-fermentation reference flux distribution. By comparing with the proposed method in this research, the formation of vanillin is acceptable. The μ_{max} of this research is reached 0.19466 where slightly less than μ_{max} for VG2 in Table 2. However, the μ_{max} of this research is higher than μ_{max} of VG0, VG1, VG3,

and VG4 in table 2. Based on Brochado and his colleagues [8], they agreed that VG4 strain showed significantly improved in cellular fitness compare VG2 strain. This is due to the yield of biomass on substrate (glucose) is higher compare other strains.

Table 2. Biomass Overall Yield of Vanillin in Batch Cultivation [7]

Strains	Engi- neered Genotype	μ_{\max}	Y_{Sx}	$Y_{S\text{ EtOH}}$	$Y_{S\text{ gly}}$
VG0		0.14	0.10	0.23	0.05
VG1	gdh1 Δ	0.10	0.07	0.25	0.03
VG2	pdh1 Δ	0.20	0.14	0.23	0.07
VG3	pdh1 Δ gdh1 Δ	0.11	0.10	0.27	0.05
VG4	pdh1 Δ gdh1 Δ \uparrow GDH2	0.17	0.17	0.25	0.07

Overall, suggested genes deletion by proposed method are included ARO4, BDH1 and ARO10 genes which can contribute to formation of L-phenylalanine, precursor for biotransformation of aromatic amino acids to produce vanillin. In theory, the more L-phenylalanine compound is produced, the more vanillin.

4 Conclusion

As a conclusion, our proposed CBAFBA which predicts the gene knockouts by *in silico* method showed to perform better in terms of time and cost-effective. The strategies applied in CBAFBA could lead to chemical production in *S. cerevisiae*. This is done by ensuring that the drain towards the metabolites/compounds necessary for growth resources such as carbons and energy must be accompanied. However, it should be noted that our proposed is deal with the reactions in the model not the real experiment. Therefore, the experiments are needed to carrier out to validate the deletion technique suggested by *in silico* technique. Specifically, CBAFBA is pinpoints which reactions needed to remove from a metabolic network, which can realized and contribute to yield the product by gene deletions where it associated with the identified the functionality. Reminder, it is important to note that the suggested gene

deletion strategies must be interpreted carefully. For instance, in many cases the deletion of gene in one branch of a branched pathway is equivalent to the significant up regulation in the other [15]. Lastly, the suggested set of gene(s) deletions is not always uniquely specified. Thus, the technique of identification of most economical gene set accounting for enzyme and multifunctional enzyme needs to be made.

Acknowledgments. We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by UTM GUP research grant that was sponsored by Universiti Teknologi Malaysia.

References

1. Vargas, F.A., Pizzarro, F., Perez-Correa, J.R., Agosin, E.: Expanding a dynamic flux balance model of yeast fermentation to genome-scale. *BMC Systems Biology* 5, 75 (2011)
2. Pham, D.T., Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S., Zaidi, M.: The Bees Algorithm-A novel tool for complex optimisation problems. *Intelligent Production Machine and Systems* (2006)
3. Orth, J.D., Thiele, I., Palsson, B.O.: What is Flux Balance Analysis. *Nat. Biotechnol.* 28(3) (2010)
4. Kauffman, K.J., Prakash, P., Edwards, J.S.: Advances in Flux Balance Analysis, vol. 14, pp. 491–496. Elsevier (2003)
5. Patil, K.R., Rocha, I., Forster, J., Nielsen, J.: Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* 6, 308 (2005)
6. Choon, Y.W., Mohamad, M.S., Deris, S., Chong, C.K., Chai, L.E., Ibrahim, Z., Omatu, S.: Identifying Gene Knockout Strategies Using a Hybrid of Bees Algorithm and Flux Balance Analysis for in silico Optimization of Microbial Strains. In: *The 9th International Symposium on Distributed Computing and Artificial Intelligence (DCAI 2012)*. University of Salamanca, Spain (2012)
7. Chaturvedi, D.K.: *Soft Computing Techniques and its Applications in Electrical Engineering*. SCI, vol. 103, pp. 363–381 (2008)
8. 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)
9. Priefert, H., Rabenhorst, J., Steinbüchel, A.: Biotechnological production of vanillin. *Appl. Microbiol. Biotechnol.* 56, 296–314 (2001)
10. Kim, J.H., Reed, J.: OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Systems Biology* 4, 53 (2010)
11. Helmstaedt, K., Strittmatter, A., Lipscomb, W.M., Braus, G.H.: Evolution of 3-deoxy-d-arabino-heptulosonate-7-phosphate synthase-encoding genes in the yeast *Saccharomyces cerevisiae* (2005)
12. González, E., Fernandez, M.R., Marco, D., Calam, E., Sumoy, E., Parés, X., Dequin, S., Biosca, J.A.: Role of *Saccharomyces cerevisiae* Oxidoreductases Bdh1p and Ara1p in the Metabolism of Acetoin and 2,3-Butanediol. *Applied and Environmental Microbiology*, 670–679 (2010)

13. Lane, A.N., Fan, T.W.M., Higashi, R.M.: Metabolic acidosis and the importance of balanced equation. *Metabolomics* 5, 163–165 (2009)
14. Hazelwood, L.A., Daran, J.M., van Maris, A.J.A., Pronk, J.T., Dickinson, J.R.: The Ehrlich Pathway for Fusel Alcohol Production: a Century of Research on *Saccharomyces cerevisiae* Metabolism. *Applied and Environmental Microbiology*, 2259–2266 (2008)
15. Burgard, A.P., Pharkya, P., Maranas, C.D.: OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering* 85(7) (2003)