Gene Knockout Strategies Identification by Using a Hybrid of Bees Algorithm and Flux Balance Analysis for Optimizing Microbial Strains

Y. W. Choon**,** M. Mohd Saberi, D.Safaai, C. K. Chong, and L. E. Chai

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

Corresponding author's e-mail: saberi@utm.my

Abstract

By optimizing microbial strains it is possible to improve product yield or improve growth characteristics. Microbial strains can be optimized through genetic engineering. It is proven that through genetic engineering it is able to obtain the desirable phenotypes. However, it is difficult to predict the effects of genetic modifications on the resulting phenotype due to the complexity of the networks. Optimization algorithms are implemented in previous works in order to identify the effects of gene knockout on the results. Sadly, the previous works face the problem of falling into local minima. Thus, a hybrid of Bees Algorithm and Flux Balance Analysis (BAFBA) is proposed in this paper to solve the local minima problem and to predict optimal sets of gene deletion for maximizing the growth rate of certain metabolite. Lists of knockout genes and the growth rate after the deletion for improving the production of succinic acid, glycerol and vanillin as targets are the results from the experiments. Genome-scale model of the yeast Saccharomyces cerevisiae is the model organism in this paper. By comparing with the previous methods, BAFBA shows better results. The identified list suggests gene modifications over several pathways and may be useful in solving challenging genetic engineering problems.

Keywords: Bees Algorithm, Flux Balance Analysis, Gene Knockout, Microbial Strains, Metabolic Engineering

Introduction

Microbial strains are strains of microorganisms which have received considerable attention for genome-scale metabolic networks reconstructions in recent years [1]. Reconstructions of the metabolic networks are found to be very useful in health, environmental, energy issues [2] the identification of drug targets. The development of computational models for simulating the actual processes inside the cell has been expedited by a vast numbers of high-throughput experimental data. Constructing an efficient and accurate pathway models that may be useful in predicting cellular responses and providing better understanding of complex biological functions is one of the main goals in system biology.

There were many algorithms developed in order to identify the gene knockout strategies for obtaining improved phenotypes. The first rational modeling frameworks (named OptKnock) for introducing gene knockout leading to the overproduction of a desired metabolite was develop by Burgard *et al.* [3,4]. A set of gene (reaction) deletions to maximize the flux of a desired metabolite is identified by OptKnock without affecting the internal flux distribution such that growth is optimized.

OptKnock is very promising to find the global optimal solution due to the use of mixed integer linear programming (MILP) to formulate a bi-level linear optimization. OptGene is an extended approach of OptKnock which formulates the *in silico* design problem by using Genetic Algorithm (GA). These meta-heuristic methods are capable in producing near-optimal solutions with reasonable computation time, furthermore the objective function that can be optimized is flexible. SA is then implemented to allow the automatic finding of the best number of gene deletions for achieving a given productivity goal. However, SA faces the problem of falling into local minima far from the global optimum solution.

In this paper, a hybrid of Bees Algorithm and Flux Balance Analysis (BAFBA) is proposed to predict the gene knockout strategies. Bees Algorithm (BA) is a typical meta-heuristic optimization approach

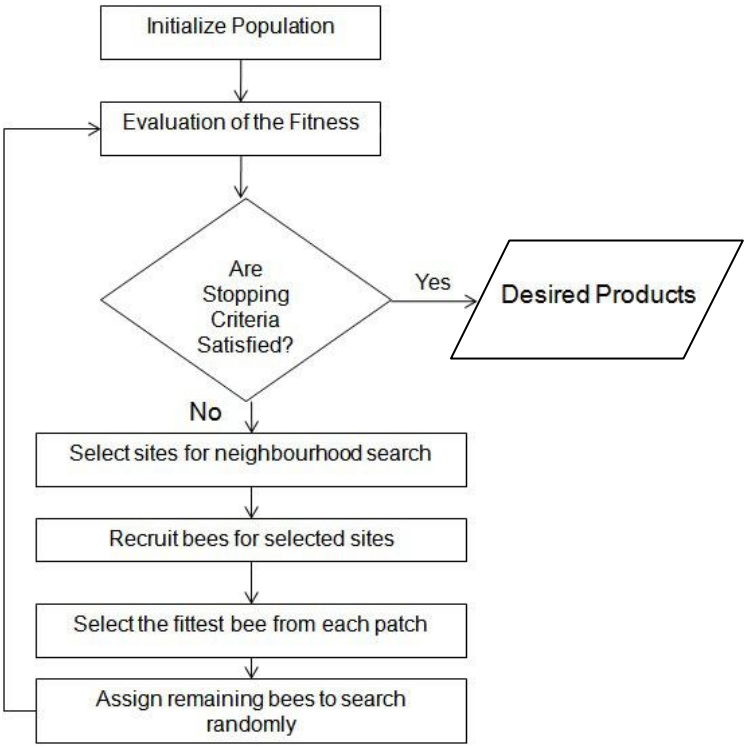
which was introduced by [5]. The search process of BA is based on the intelligent behaviors of honey bees. BA locates the most promising solutions, and selectively explores their neighbourhoods looking for the global maximum of the objective function. BA is proven to be efficient in solving optimization problems in the previous studies [5]. While the Flux Balance Analysis (FBA) approach which is used to calculate the fitness function is based on a steady state approximation to concentrations of the internal metabolites, which reduces the corresponding mass balances to a set of linear homogeneous equations. There are two advantages of BAFBA. First, BAFBA requires less computational time to solve larger size problems. Secondly, BA works out the local minima problem as it is capable of performing local and global search simultaneously. This paper presents the results obtained by BAFBA to three case studies where *S.Cerevisiae* is the target microorganism. This paper also evaluates the performance of BAFBA for identifying gene knockout strategies with existing tools and compares the performance of BA with the existing methods within experimental approaches.

Materials and Methods

In this paper, BAFBA is proposed to predict the gene knockout. Fig. 1 shows the flow of a basic BA. The flow of BAFBA is presented in Fig. 2. The important steps are explained in the following subsections.

Bee representation of metabolic genotype

One or more genes can be found in each reaction in the metabolic model. In this proposed method, each of those genes is represented by a binary variable indicating its absence or presence (0 or 1), these variables form a 'bee' representing a particular mutant that lacks some metabolic reactions when compared with the wild type (Fig. 3)



Note: Desired products represent the gene to be knockout.

Fig. 1: Flowchart of a basic BA.

Initialization of the population

Firstly, randomly initialize a population of *n* scout bees. Each bee is initialized as follows: assume that a reaction with *n* genes. Bees in the population can be initialized by assigning present or absent status to each gene randomly.

Scoring fitness of individuals

The fitness computation process for each site visited by a bee is evaluated through FBA (Fig. 4). Cellular growth is defined as the objective function Z, vector **c** is used to select a linear combination of metabolic fluxes to include in the objective function, **v** is the flux map and *i* is the index variable (1, 2, 3, …, *n*).

Maximize Z, where

$$Z = \sum_i c_i v_i = \mathbf{c}.\mathbf{v} \quad (1)$$

where *c* = a vector that defines the weights for of each flux.

Neighbourhood search

Neighbourhood searches is carried out in the selected sites, more bees are assigned to search near the best sites. The bees can be chosen directly according to their fitnesses associated with the sites they are visiting. Searches in the neighbourhood of the best sites which represent more promising solutions are further more detailed by recruiting more bees to follow them than other selected bees.

Randomly assigned and termination

Assigning the remaining bees in the population randomly around the search space scouting for new potential solutions. These steps are repeated until a stopping criteria is met. The stopping criteria are either the maximum loop value is met or the fitness function has converged. At the end of each iteration, the colony produces two parts to its new population – representatives from each selected patch and other scout bees assigned to conduct random searches.

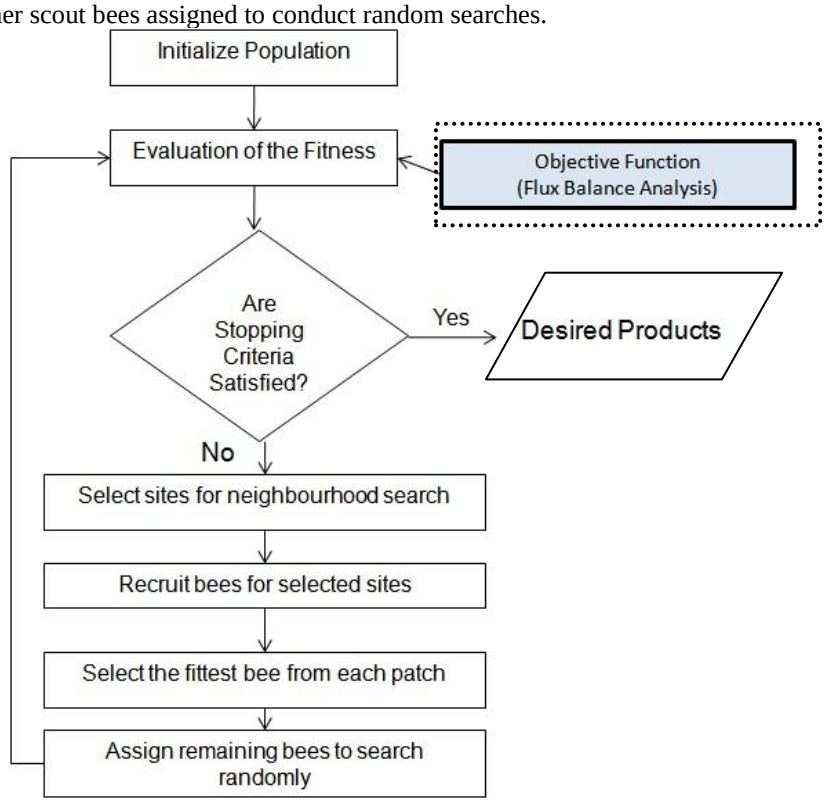
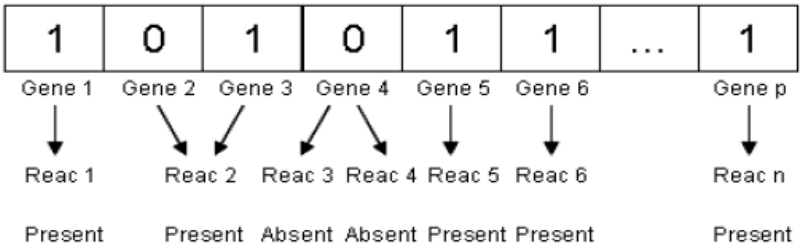


Fig. 2: The flow of BAFBA.



Note: Reac represents reaction.

Fig. 3: Bee representation of metabolic genotype

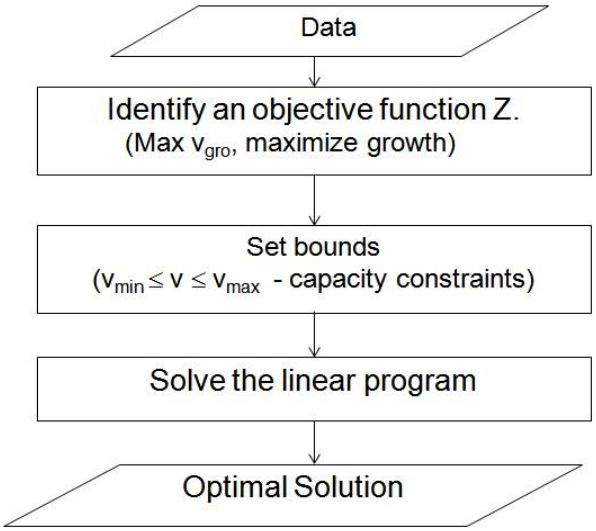


Fig. 4: Steps in FBA

Results and Discussion

In this paper, *S.Cerevisiae* is used as the dataset to test on the operation of BAFBA. The results obtained are compared to the previous works reported in the literature studies (6,7). Millimole (mmol) is the unit of concentration whereas millimoles per hour (mmol/hr) is used as the unit measurement in the experiments.

Table 1 and Table 2 summarize the results obtained from BAFBA. As shown from the results, this method has produced better results to the previous works. In this paper, potential reactions which can be removed are identified.

Firstly, BAFBA suggests the removal of five reactions from the network results in succinate growth rate reaching 1.7023 which is better than the other two methods. The list of knockout genes obtained is to eliminate the competing byproduct (i.e, pyruvate).

Next, BAFBA is applied to identify knockout strategy for producing glycerol. Table 2 shows the best result is obtained from this method is 1.7023. From the list of knockout genes, it can be concluded that this strategy focuses on inactivating PEP consuming reactions. BAFBA is also applied to produce vanillin in this paper. Table 3 shows the result of BAFBA compared with the other methods. The removal of three reactions from the network results in vanillin growth rate reaching 1.7023. BAFBA produced the best results in all cases, due to the advantage of BA performs local and global search simultaneously to avoid being trapped at locally optimal solutions. BA splits the search into exploration and exploitation, which are then executed parallely rather than serially like SA. Thus, BA performs better than SA where it solves the local minima problem faced by SA.

Table 1: Comparison between different methods for production of Succinate

Method	Growth Rate (mmol/hr)	List of knockout genes
BAFBA	1.7023	(R)-lactate:ferricytochrome-c 2-oxidoreductase, 2-dehydropantoate 2-reductase, 2-deoxy-D-arabino-heptulosonate 7-phosphate synthetase, 2-keto-4-methylthiobutyrate transamination, 2-oxobutanoate dehydrogenase
SA + FBA [7]	0.05398	PGII_I, PGII_2, FBPI, PDC6, ADH4, SDH3_2, AAHI_I, URHI_I, U30_, MET3, ALD4_2, GSHI, UI03_, YER053C, CTPI_I
OptGene [6]	0.39	SDH-complex, ZWF I, PDC6, UI33, U221

Note: The shaded column represents the best result.

Table 2: Comparison between different methods for production of Glycerol

Method	Growth Rate (mmol/hr)	List of knockout genes
BAFBA	1.0723	(R)-lactate:ferricytochrome-c 2-oxidoreductase, (R,R)-butanediol dehydrogenase, 2-deoxy-D-arabino-heptulosonate 7-phosphate synthetase, 2-keto-4-methylthiobutyrate transamination, 2-methylcitrate synthase
OptGene [6]	0.49	FBP1, Glyceraldehyde-3-phosphate dehydrogenase

Note: The shaded column represents the best result.

Method	Growth Rate (mmol/hr)	List of knockout genes
BAFBA	1.7023	(S)-lactate:ferricytochrome-c 2-oxidoreductase, 1-phosphatidylinositol-3-phosphate 5-kinase, 2-deoxy-D-arabino-heptulosonate 7-phosphate synthetase
OptGene [6]	0.57	Pyruvate decarboxylase, Glutamate dehydrogenase

Table 3: Comparison between different methods for production of Vanillin

Note: The shaded column represents the best result.

In addition, Table 4, Table 5 and Table 6 show the results of three of the identified gene knockout strategies for succinate, glycerol and vanillin overproduction.

Table 4 shows three of the identified gene knockout strategies (i.e., mutants A, B, and C). For the production of succinate, (R)-lactate:ferricytochrome-c 2-oxidoreductase which contributes to the phosphotransferase system for all three mutant A, B, and C is disabled, this causes the network to rely exclusively on glucokinase for glucose uptake.

Table 4: Result of different knockout strategies for production of Succinate

Mutants	Growth Rate (mmol/hr)	List of knockout genes
A	5.7285e-013	(R)-lactate:ferricytochrome-c 2-oxidoreductase, 2,5-diamino-6-ribitylamino-4(3H)-pyrimidinone 5'-phosphate deamin, 2-aceto-2-hydroxybutanoate synthase
B	0.57285	(R)-lactate:ferricytochrome-c 2-oxidoreductase, (S)-lactate:ferricytochrome-c 2-oxidoreductase, 2-aceto-2-hydroxybutanoate synthase, 2-dehydropantoate 2-reductase
C	1.7023	(R)-lactate:ferricytochrome-c 2-oxidoreductase, 2-dehydropantoate 2-reductase, 2-deoxy-D-arabino-heptulosonate 7-phosphate synthetase, 2-keto-4-methylthiobutyrate transamination, 2-oxobutanoate dehydrogenase

Table 5: Result of different knockout strategies for production of Glycerol

Mutants	Growth Rate (mmol/hr)	List of knockout genes
D	4.8295e-013	(R)-lactate:ferricytochrome-c 2-oxidoreductase, 2-aceto-2-hydroxybutanoate synthase, 2-dehydropantoate 2-reductase
E	5.4019e-013	(S)-lactate:ferricytochrome-c 2-oxidoreductase, 2-aceto-2-hydroxybutanoate synthase, 2-dehydropantoate 2-reductase, 2-deoxy-D-arabino-heptulosonate 7-phosphate synthetase
F	1.0723	(R)-lactate:ferricytochrome-c 2-oxidoreductase, (R,R)-butanediol dehydrogenase, 2-deoxy-D-arabino-heptulosonate 7-phosphate synthetase, 2-keto-4-methylthiobutyrate transamination, 2-methylcitrate synthase

Table 6 Comparison between different methods for production of Vanillin

Mutants	Growth Rate (mmol/hr)	List of knockout genes
G	1.7023	(S)-lactate:ferricytochrome-c 2-oxidoreductase,1-phosphatidylinositol-3-phosphate 5-kinase,2-deoxy-D-arabino-heptulosonate 7-phosphate synthetase
H	1.8127e-014	(S)-lactate:ferricytochrome-c 2-oxidoreductase,1-pyrroline-5-carboxylate dehydrogenase,2-aceto-2-hydroxybutanoate synthase,2-dehydropantoate 2-reductase
I	9.6778e-017	1-phosphatidylinositol-3-phosphate 5-kinase,1-pyrroline-5-carboxylate dehydrogenase,2-hydroxybutyrate:NAD+ oxidoreductase,2-isopropylmalate hydratase,2-oxo-4-methyl-3-carboxypentanoate decarboxylation

Table 5 shows the result of different knockout strategies for the production of glycerol, phosphotransferase system for all three mutant D, E and F are disabled. Lastly, Table 6 shows the different knockout strategies obtained by BAFBA in producing vanillin, for mutant G and H, (S)-lactate:ferricytochrome-c 2-oxidoreductase which contributes to the phosphotransferase system is disabled, this causes the network to rely exclusively on glucokinase for glucose uptake. For mutant I, deletion of 1-pyrroline-5-carboxylate dehydrogenase which belongs to the family of oxidoreductases, results in an increased availability of NADPH needed for vanillin biosynthesis. In conclusion, the phosphotransferase system affect greatly to the production of succinate, glycerol and vanillin.

Conclusion and Future Works.

BAFBA is proposed to predict optimal sets of gene deletion in order to maximize the production of certain metabolite in this paper. This method is based on BA, where the local minima problem faced by SA is worked out as BA is capable of performing local and global search simultaneously. The FBA approach is used as a fitness function whereby it is based on a steady state approximation to concentrations of the internal metabolites, which reduces the corresponding mass balances to a set of linear homogeneous equations.

Experimental results on *S.Cerevisiae* model dataset obtained from literature [6] showed that BAFBA is a useful tool in Metabolic Engineering as it is effective in generating optimal solutions to the gene knockout prediction.

The performance of BAFBA can be further improved by applying an automated pre-processing operation in BAFBA to simplify the genome-scale metabolic model. The development of multi-objective optimization algorithms in a single run to achieve two goals, for example, maximizing the biomass and the desired product, is another interesting feature which can be implemented in the standard BAFBA. Lastly, as BA employs many tunable parameters which are difficult for the user to select, it is important to find ways to help the user choose appropriate parameters, for example, parameter tuning.

Acknowledgement

This work is financed by Institutional Scholarship MyPhD provided by the Ministry of Higher Education of Malaysia. We also would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grants (vot number: Q.J130000.7123.00H67 and Q.J130000.7107.01H29).

References

[1] A.M. Feist, M.J. Herrgård , I. Thiele, J.L. Reed, B.O. Palsson: *Reconstruction of biochemical networks in microorganisms*, Nat Rev Microbiol, Vol. 72 (2009), p.129-143.
[2] D. Chandran, W.B. Copeland, S.C. Sleight, H.M. Sauro: *Mathematical modeling and synthetic biology*, Drug Discovery Today Disease Models, Vol. 5, No.4 (2008), p. 299-309.
[3] A.P. Burgard, P, Pharkya, C.D. Maranas: *OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strains optimization*, Biotechnol Bioeng, Vol. 84 (2003), p. 647-657.

[4] P. Pharkya, A.P. Burgard, C.D. Maranas: *OptStrain: a computational framework redesign of microbial production systems*, Genome Res, Vol.14 (2004), p. 2367-2376.
[5] D.T. Pham, A. Ghanbarzadeh, E. Koç, S. Otri, M. Zaidi: *The bees algorithm – a novel tool for complex optimization problems*, Proceedings of the Second International Virtual Conference on Intelligent Production Machines and Systems, (2006), p. 454-461.
[6] K.R. Patil, I. Rocha, J. FÖrster, J. Nielsen: Evolutionary *programming as a platform for in silico metabolic engineering*, BMC Bioinformatics, Vol.6 (2005), p. 308.
[7] M. Rocha, P. Maia, R. Mendes, J.P. Pinto, E.C. Ferreira, J. Nielsen, K.R. Patil, I. Rocha: *Natural computation meta-heuristics for the in silico optimization of microbial strains*, BMC Bioinformatics, Vol. 9 (2008), p. 499.