A Hybrid of Artificial Bee Colony and Flux Balance Analysis for Identifying Optimum Knockout Strategies for Producing High Yields of Lactate in *Echerichia Coli*

Seet Sun Lee[1], Yee Wen Choon[1], Lian En Chai[1], Chuii Khim Chong[1], Safaai Deris[1], Rosli M. Illias[2], 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
{sslee6,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. The advent of genome-scale models of metabolism has laid the foundation for the development of computational procedures for suggesting genetic manipulations that lead to overproduction. Previously, for increasing the production of Lactate in *E. coli*, a traditional method of chemical synthesis was being used, this always lead the products are far below their theoretical maximums. This is not surprise as the cellular metabolism is always competing with the chemical overproduction. Besides, several optimization algorithms often get stuck at a local minimum in a multi-modal error. In this research, a hybrid of Artificial Bee Colony (ABC) and Flux Balance Analysis (FBA) is proposed for suggesting gene deletion strategies leading to the overproduction of Lactate in *E. coli*. In this work, the ABC is introduced as an optimization algorithm based on the intelligent behavior of honey bee swarm. As for the evaluation of fitness part, each mutant strain is evaluated by resorting to the simulation of its phenotype using the FBA, together with the premise that microorganisms have maximized their growth along natural evolution. This is the first research that successfully combined ABC and FBA for identifying optimum knockout strategies. The successfully created hybrid algorithm is applied to the *E. coli* model dataset.

Keywords: Artificial Bee Colony, Flux Balance Analysis, Lactate, Gene KnockOut, *Echerichia Coli*.

1 Introduction

There is a genetic technique called gene knockout where the one of the organism's genes is being made to inoperative, just like to knock out the specific gene from the

* Corresponding author.

A.S. Sidhu & S.K. Dhillon (Eds.): *Adv. in Biomedical Infrastructure 2013*, SCI 477, pp. 127–137.
DOI: 10.1007/978-3-642-37137-0_13

organism. This technique is a platform for human to learn about how a gene functions based on the sequenced gene. Researchers draw inferences from the difference between the knockout organism and normal individuals.

Besides, the term also refer as creating a new organism as “knocking out” a gene, this is essentially opposite of a gene knocking. Double knockout has the meaning of two genes being knocked out at the same time. The same meaning goes to triple knockout and quadruple knockout which describes the 3 and 4 genes being knocked out simultaneously.

Succinate and its derivatives have been used as common chemicals to synthesize polymers, as additives and flavoring agents in foods, supplements for pharmaceuticals, or surfactants. Currently, it is mostly produced through petrochemical processes that can be expensive and have significant environmental impacts. In fact, the knockout solutions that lead to an improved phenotype regarding the production of Succinates are not straightforward to identify since they involve a considerable number of interacting reactions.

Lactate and its derivatives have been used in a wide range of food-processing and industrial applications like meat preservation, cosmetics, oral and health care products and baked goods. Additionally, as lactate can be easily converted to readily biodegradable polyesters, it is emerging as a potential material for producing environmentally friendly plastics from sugars [1].

Several microorganisms have been used to commercially produce lactate [2], such as Lactobacillus strains. However, those bacteria also have undesirable traits, such as a requirement for amino acids and vitamins which complicates acid recovery. E. coli has many advantageous characteristics as a production host, such as rapid growth under aerobic and anaerobic conditions and simple nutritional requirements. Moreover, well-established protocols for genetic manipulation and a large knowledge on this microbe's physiology enable the development of E. coli as a host for production of optically pure D- or L-lactate by metabolic engineering [3].

The first approach to suggest gene deletion strategies was the OptKnock algorithm, where mixed integer linear programming (MILP) is used to reach an optimum solution. An alternative approach was proposed by the OptGene algorithm that considers the application of Evolutionary Algorithms (EAs), EAs are a meta-heuristic optimization method, and they are capable of providing solutions in a reasonable amount of time.

Unfortunately, for the above approaches, they may often get stuck at a local minimum in a multi-modal error. Based on this, above algorithms might not perform well in global and local optimization which will lead to local minimum and inefficiently used for multivariable and multimodal functions optimization [4]. Therefore, a combination of Artificial Bee Colony (ABC) and Flux Balance Analysis (FBA) has been looked into for identifying the gene knockout strategies for obtaining high yields of Succinate in *E. coli*. The developed algorithm is evaluated in term of the production of biochemical in *E. coli*.

The successfully created hybrid algorithm has contributed to the gene knockout field where it can design the experiment protocol so that biochemical production will be increased. Before this, there is no research is being carried out for the hybrid of

these two algorithms. Moreover, the newly formed hybrid algorithm is applied on the *E. coli* dataset.

2 Methods

2.1 Hybrid of Artificial Bee Colony and Flux Balance Analysis

In this section, we describe the details of the proposed ABCFBA in which ABC is newly combined with FBA to identify optimal gene knockout strategies. In essence, the proposed algorithm consists of five main steps:

1. Initialize population
2. Employed phase
3. Onlooker phase
4. Memorize the best
5. Scout phase

Figure 1 shows the flow of ABCFBA. Figure 2 shows the comparison of ABCFBA and ABC, rectangles that in red color showed the difference steps between the original ABC and ABCFBA. The flow chart on the left side indicates the original ABC algorithm while the flow chart on the right side is the ABCFBA. As compare with the original ABC, this study's method has integrated the FBA into ABC for the purpose of fitness calculation which main for identifying optimum knockout strategies in *E. coli* model.

Originally, the ABC is main for food foraging of honey bees, therefore, its fitness calculation is the nectar amounts calculation while ABCFBA is focusing on the gene knockout identification, so its fitness calculation step will be replaced by FBA.

Based on Edwards and Palsson [5], FBA was developed to analyze the metabolic capabilities of a cellular system based on the mass balance constraints. The mass balance constraints in a metabolic network can be represented mathematically by a matrix equation as follow:

$$S \cdot v = 0 \tag{1}$$

The matrix S is the mxn stoichiometric matrix, where m is the number of metabolites and n is the number of reaction in the network. The vector v represents all fluxes in the metabolic network, including the internal fluxes, transport fluxes and the growth flux.

For the *E.coli* metabolic network, the number of fluxes was greater than the number of mass balance constraints; thus, there were multiple feasible flux distributions that satisfied the mass balance constraints, and the solutions were confined to the null space of the matrix S. as follow:

$$\alpha i \leq vi \leq \beta i \tag{2}$$

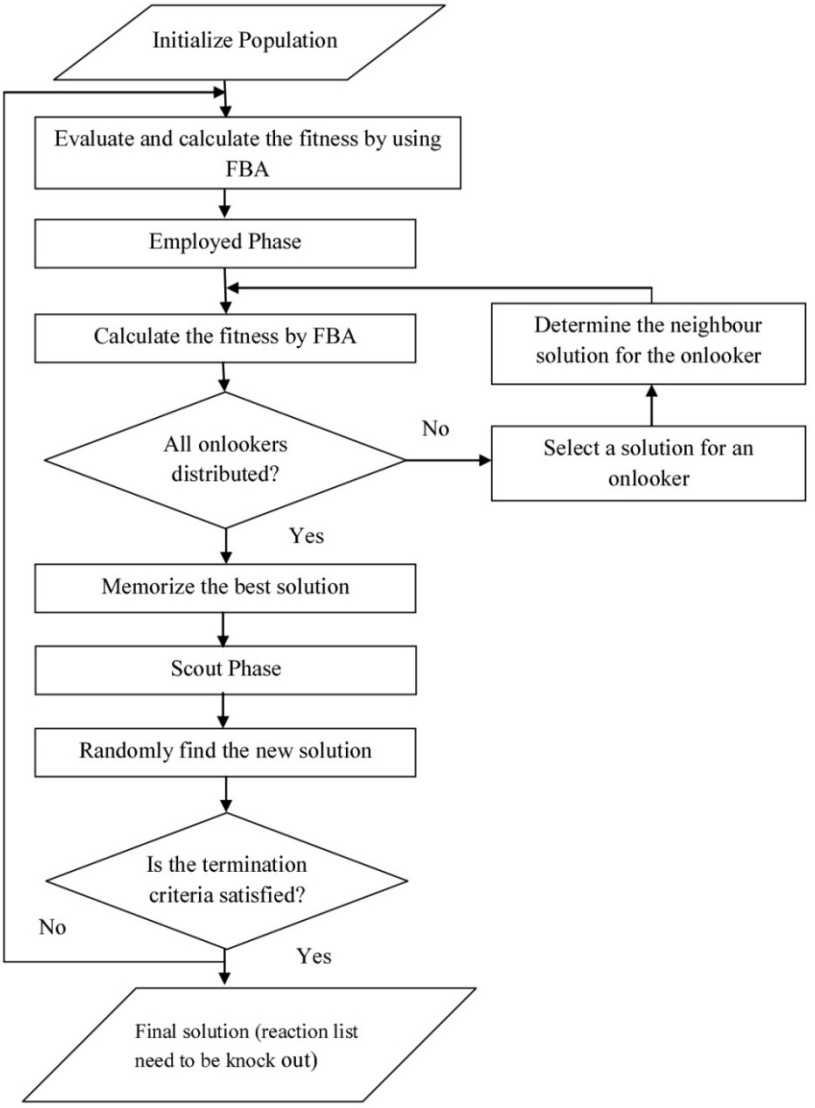

Fig. 1. Flow of ABCFBA

The linear inequality constraints were used to enforce the reversibility of each metabolic reaction and the maximal flux in the transport reactions. The reversibility constraints for each reaction are indicated online. The transport flux for inorganic phosphate, ammonia, carbon dioxide, sulfate, potassium, and sodium was unrestrained ($\alpha i = -\infty$ and $\beta i = \infty$).

The transport flux for the other metabolites, when available in the in silicon medium, was constrained between zero and the maximal level ($0 \leq v_i \leq v_i max$). The $v_i max$ values used in the simulations are noted for each simulation. When a

metabolite was not available in the medium, the transport flux was constrained to zero. The transport flux for metabolites capable of leaving the metabolic network was always unconstrained in the net outward direction.

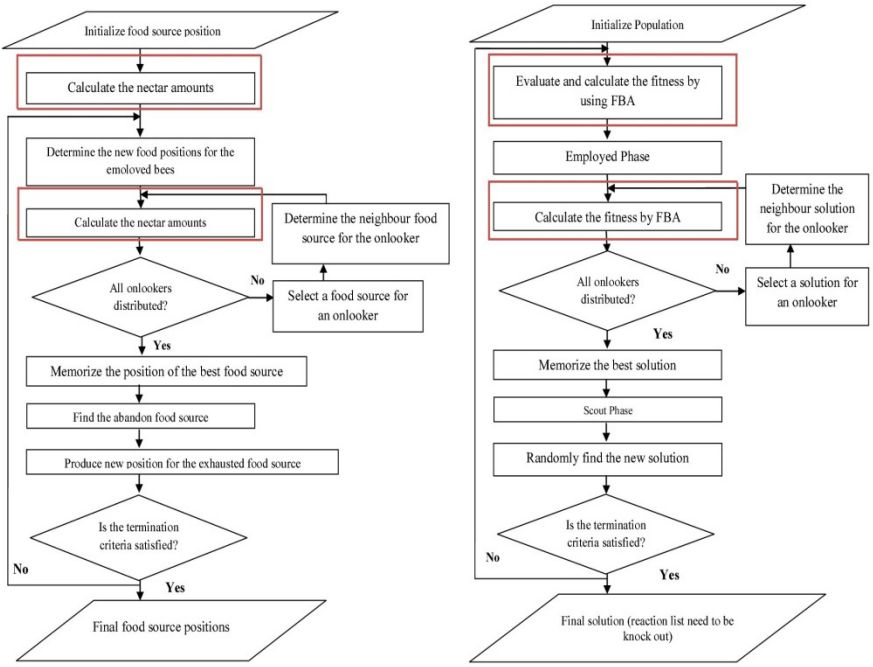

Fig. 2. Comparison of ABCFBA and ABC

The intersection of the nullspace and the region defined by the linear inequalities defined a region in flux space that we will refer to as the feasible set and the feasible set defined the capabilities of the metabolic network subject to the imposed cellular constraints. It should be noted that every vector v within the feasible set is not reachable by the cell under a given condition due to other constraints not considered in the analysis. The feasible set can be further reduced by imposing additional constraint and in the limiting condition where all constraints are known, the feasible set may reduce to a single point.

A particular metabolic flux distribution within the feasible set was found using Linear Programming (LP). LP identified a solution that minimized a metabolic objective function, and was formulated as shown below:

$$\text{Minimize} - Z \tag{3}$$

where $Z = \sum ci\ vi = < c \bullet v >$

The vector c was used to select a linear combination of metabolic fluxes to indicate in the objective function. Herein, c was defined as the unit vector in the direction of

the growth flux, and the growth flux was defined in terms of the biosynthetic requirement.

$$\sum_{all\ m} dm \bullet Xm \xrightarrow{V_{grow}} \text{Biomass} \quad (4)$$

where dm is the biomass composition of metabolite Xm, and the growth flux was modeled as a single reaction that converts all the biosynthetic precursors into biomass.

As compare to other well-known metabolic modeling approaches, FBA is different in term of accuracy, this is because instead of predicting the metabolic behavior, it defines the 'best' the cell can do. FBA assumes that the regulation is such that metabolic behavior is optimal but not directly considers regulation or the regulatory constraints. Therefore, the results are generally consistent. However, it is only valid for a system that has evolved toward optimally.

In mutant strains, the regulation of the metabolic network has not evolved to operate in an optimal fashion. Because of this, it will cause a problem when coupling to highly parallel experimental programs, such as large-scale mutation studies.

FBA is an effective tool for the analysis of metabolic networks. FBA can complement the uncertainly and incompleteness of metabolic data, and thereby provide a better characterization of cellular phenotypes. Recent advances in FBA include the prediction of flux distribution of engineered cells, investigation of a cellular objective and the design of a mutant strain with desired properties.

Although the development of analytic techniques has facilitated the generation of dynamic profiles of metabolites, such data sets are not accurate enough for generating large-scale kinetic models. FBA has its pro and con in analyzing the biological network.

Initialize Population

The system start with create a population with the matrix of 95x500, since there are 95 reactions in the E. coli model and the dimension of the matrix where it must more than the number of reactions which is 500. This matrix was essentially create with all value 0's, then the value 1's were randomly distributed among them. The 1's represent those reaction that will be knockout while the 0's represent those reaction that cannot be knockout.

After the population has been created, each line of the columns, the population of the possibility of the reaction knockout, will be the inputs of the FBA for calculating the fitness. The system will return growth rate which determine whether the cell still survive after the deletion occurs where the value must more than 0.1. Another value that will return is the minimum production which represents the minimum production of biochemical after the deletion occurs where it must be more than -1e-3 to prevent the very small values from being considered as improvement.

Employed Phase

As for this stage, it is performing a job of randomly creating a new population where it is near the original population. For the 500 populations that created from the first

stage, the system will randomly create another 500 populations. Then, the greedy algorithm will being applied so that those with the smaller value of fitness will be abandoned. The new generated population will formed with the better fitness values.

The greedy algorithm is based on the evaluation of a pre-defined maximum number of solutions that are obtained in the neighborhood of the best ones found and by using exhaustive search when no local search can be performed.

There might have some original populations have the higher fitness value than newly formed possibilities. This showed the current solution cannot be improved. Therefore, the control parameters, trial, will increase by 1. Otherwise, it will remain as 0.

Onlooker Phase

The onlooker phase basically is randomly generating other neighborhoods, but it has a little bit different with employed phase. This phase started with calculating the probability value p for the fitness, fiti values by using the formula:

$$P_i = \frac{fit_i}{\sum_{i=1}^{\frac{CS}{2}} fit} \tag{5}$$

The highest values of that specify possible reaction knockout will as the input of this phase, the system will randomly generate another new population and compare with the old one. If newly formed has higher fitness than the older formed, it will replace the older, vice-versa situation happen on the other hands.

Then, the system will recalculate the value p to decide the next population that will be replaced or remained. This phase will iterate till 500 times. Those populations that cannot be replaced will increase the trial value by 1 while those have replaced will set the trial to 0.

This will result the good potential population will become better while the bad population will being abandoned forever as the p values of good populations will keep increasing while the bad populations' p values will keep decreasing since the fitness value is divided by sum of all the fitness values for every population.

Memorize the Best

After gone through the three phases, the 500 populations will be the input for this stage. The best population will be selected based on the fitness value by using Greedy Selection algorithm. Only one population will remain as result where it represents the best reactions knockout list in term of highest growth rate and highest yields of target biochemical productions.

Scout Phase

If the population cannot be improved where its predetermined number of trial has exceed the limit=100, the population considered exhausted, it will be abandoned. The

Employed bee will immediately transform become the scout bee, then it will randomly generate another new population, evaluate and calculate the new fitness.

Then the phase will loop back to the calculation of fitness, go through the employed phase, onlooker phase, and memorize the best phase again and again. The system will repeat the cycle until it satisfies the termination criteria which is the maxCycle more than 200.

Finally, the result obtained has the best fitness value where it is the best knock out reaction list in the model in term of local and global search in ABC algorithm.

3 Results and Discussion

3.1 Experimental Data

In this research, the *E. coli* K-12 stoichiometric model [5] dataset is being used. The dataset can be found in BioModels Database, KEGG and system biology research group. The datasets is in SBML format.

This model basically contained 26 fields. Since the research is mainly focus on the biochemical productions which results from deletion of certain pathways or reactions that happened inside the cell. Therefore, the fields that used in this research are rxns, lb, ub and rxnNames.

For rxns field, it contained of 95 reactions (see Figure 3) which are the input data set and based on the knockout number, the system will identify which reactions can be deleted as to result a high yield of biochemical productions.

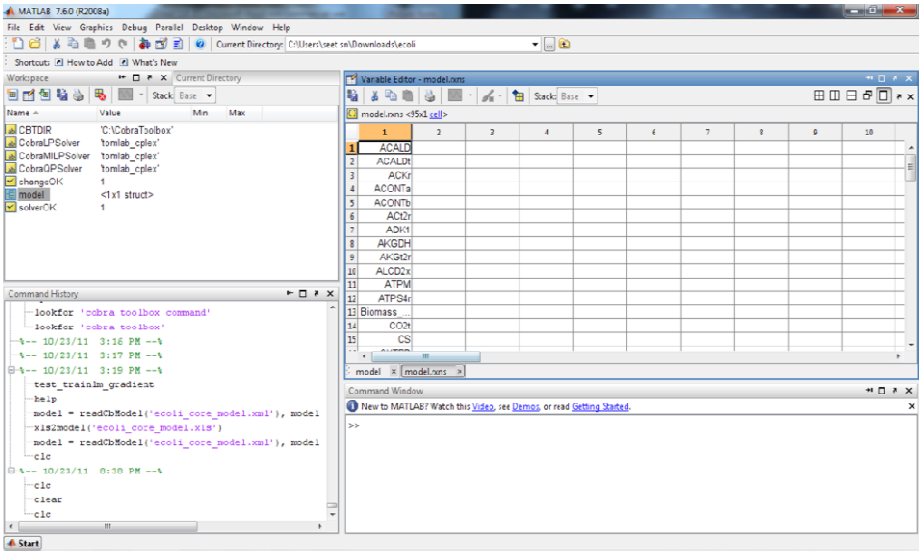

Fig. 3. 95 Reactions in the rxns field

While for lb and up, they stands for the lower boundary and upper boundary of the 95 reactions respectively. Last but not least, the rxnNames is the full names of the 95 reactions that can be understand by human language.

The target reaction in this research is lactate. Table 1 shows 3 sets of knockout list as result after 50 runs. The deletion of gene adhE which formed enzyme Alcohol dehydrogenase (ethanol) will increase the production of Lactate to 18.0738 mol per hour. The growth rate of the *E.coli* is 0.1186 indicates the cells still survived after the deletion. According to Q. Hua *et al*. [6], mutation in gene such adhE in the anaerobic environment, the lactate secretion will significantly increase. Gene adhE is catalyzing the reduction of acetyl-CoA to ethanol. After the deletion of adhE, the more highly reduced fermentation byproduct ethanol cannot be produced, NAD+ regeneration will mainly depend on the reduction pathway of pyruvate to lactate. Therefore the Lactate production will keep on increasing.

The deletion of genes ackA and adhE has the same lactate production as previous deletion, 18.0738 mol per hour. Although the lactate production remains the same but the growth rate for this deletion is higher than the previous deletion, which is 0.1253 if compare to 0.1186. L. Zhou. et al. [7]'s study stated in strain B0013, acetate is the main byproduct, the encoding gene (ackA) was initially deleted to reduce acetate yield and to increase lactate yield. Acetate kinase catalyzes the conversion of pyruvate via acetyl coenzymeA (CoA) and acetyl-phosphate to acetate. By deleting gene ackA, the main pathway for acetate production in E. coli has been restricted. Since the inhibitor of lactate production was disappear, then lactate will be produced significantly.

Table 1. KnockOuts list for the target reaction of Lactate in *E.coli*

KnockOuts	Enzyme	Lactate (gram-glucose.hour)$^{-1}$	Growth Rate (h^{-1})
1 NAD + 1 ETOH <==> 1 NADH + 1 H + 1 ACALD	Alcohol dehydrogenase (ethanol)	18.0738	0.1186
ACTP + ADP <== > AC + ATP 1 NAD + 1 ETOH <==> 1 NADH + 1 H + 1 ACALD	Acetate kinase Alcohol dehydrogenase (ethanol)	18.0738	0.1253
FADH2 + Fumarate <==> FAD + SUCC	Fumarate reductase	18.0738	0.1253

The deletion of gene frdA, Fumarate reductase, has generated 18.0738 mol of lactate per hour, and the growth rate is 0.1253. Both biomass and growth rate achieved the same amount with the second deletion even through both deletions are not the same. Based on the study of Y. Zhu et al. [8], accumulation of succinate was prevented by knockout of gene frdA. During the anaerobic respiration, menaquinol-fumarate

oxidoreductase (QFR) is used for succinate production. Since the production of succinate being prevented, the lactate production increase significantly.

For the obtained results, three of them are having the same Lactate production 18.0738 with different growth rate. All the obtained results have proved with the wet laboratory results that the predicted knockout list has increased the biochemical production in the industry. This also proved that the newly formed hybrid algorithm has good performance in identifying the gene knockout list.

Overall, the obtained results are consistent. This is because ABC algorithm has the advantages of simple, high robustness, fast convergence, high flexibility and fewer control parameters. Hence, it solved the multidimensional and multimodal optimization problems.

4 Conclusions

As a conclusion, our proposed hybrid algorithm showed a better performance than the previous gene knockout tools such as OptKnock and OptGene in term of the gene knockout identification for producing high yields of succinate and lactate in *E,coli*. ABC algorithm has the advantages of simple, high robustness, fast convergence, high flexibility and fewer control parameters. In the future work, another new data set was suggested to put in as to test the feasibility of this newly develop algorithm. Besides, other intelligent optimization algorithms like ants colony, particle swarm optimization (PSO) were encouraged to replace the artificial bee colony algorithm, so that by comparing these algorithms, a better algorithm will be found.

Acknowledgements. 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. Hofvendahl, K., Hahn-Hagerdal, B.: Factors Affecting the Fermentative Lactic Acid Production from Renewable Resources. Enzyme and Microbial Technology 26(2-4), 87–107 (2000)
2. John, R.P., Nampoothiri, K.M., Pandey, A.: Production of L(+) Lactic Acid from Cassava Starch Hydrolyzate by Immobilized *Lactobacillus delbrueckii*. J. Basic Microbiol. 47(1), 25–30 (2007)
3. Chang, D.E., Jung, H.C., Rhee, J.S., Pan, J.G.: Homofermentative Production of D- or L-Lactate in Metabolically Engineered *Escherichia coli* RR1. Appl. Environ. Microbiol. 65(4), 1384–1389 (1999)
4. Karaboga, D., Basturk, B.: A Powerful and Efficient Algorithm for Numerical Function Optimization: Artificial Bee Colony (ABC) Algorithm. Journal of Global Optimization 39(3), 459–471 (2007)

5. Edwards, J.S., Palsson, B.O.: Metabolic Flux Balance Analysis and The In Silico Analysis of *Escheichia coli* K-12 Gene Deletions. BMC Bioinformatics 1, 1 (2000)
6. Hua, Q., Joyce, A.R., Fong, S.S., Palsson, B.O.: Metabolic Analysis of Adaptive Evolution for In Silico-Designed Lactate-Producing Strains. Biotechnology and Bioengineering 95(5), 992–1002 (2006)
7. Zhou, L., Zuo, Z.R., Chen, X.Z., Niu, D.D., Tian, K.M., Bernard, A.P., et al.: Evaluation of Genetic Manipulation Strategies on D-Lactate Production By *Escherichia Coli*. Current Microbiology 62(3), 981–989 (2010)
8. Zhu, Y., Eiteman, M.A., DeWitt, K., Altman, E.: Homolactate Fermentation by Metabolically Engineered *Escherichia coli* Strain. Applied and Environmental Microbiology 73(2), 456–464 (2006)