Genetics and Molecular Research

Genome sequencing and systems biology analysis of a lipase-producing bacterial strain

N. Li*, D.D. Li*, Y.Z. Zhang, Y.Z. Yuan, H. Geng, L. Xiong and D.L. Liu

School of Life Sciences, Central China Normal University, Wuhan, China

*These authors contributed equally to this study.
Corresponding author: D.L. Liu
E-mail: ldl@mail.ccnu.edu.cn

Genet. Mol. Res. 15 (1): gmr.15017331
Received September 14, 2015
Accepted December 28, 2015
Published March 18, 2016
DOI http://dx.doi.org/10.4238/gmr.15017331

ABSTRACT. Lipase-producing bacteria are naturally-occurring, industrially-relevant microorganisms that produce lipases, which can be used to synthesize biodiesel from waste oils. The efficiency of lipase expression varies between various microbial strains. Therefore, strains that can produce lipases with high efficiency must be screened, and the conditions of lipase metabolism and optimization of the production process in a given environment must be thoroughly studied. A high efficiency lipase-producing strain was isolated from the sediments of Jinsha River, identified by 16S rRNA sequence analysis as *Serratia marcescens*, and designated as HS-L5. A schematic diagram of the genome sequence was constructed by high-throughput genome sequencing. A series of genes related to lipid degradation were identified by functional gene annotation through sequence homology analysis. A genome-scale metabolic model of HS-ML5 was constructed using systems biology techniques. The model consisted of 1722 genes and 1567 metabolic reactions. The topological graph of the genome-scale metabolic model was compared to that of conventional metabolic pathways using a visualization software and KEGG database. The basic components and boundaries of the tributyrin degradation

subnetwork were determined, and its flux balance analyzed using Matlab and COBRA Toolbox to simulate the effects of different conditions on the catalytic efficiency of lipases produced by HS-ML5. We proved that the catalytic activity of microbial lipases was closely related to the carbon metabolic pathway. As production and catalytic efficiency of lipases varied greatly with the environment, the catalytic efficiency and environmental adaptability of microbial lipases can be improved by proper control of the production conditions.

Key words: Lipase-producing strain; Genome sequencing; Systems biology; Genome-scale metabolic model

INTRODUCTION

Lipase (triacylglycerol acylhydrolase; EC3.1.1.3) catalyzes the hydrolysis of long-chain fatty glycerides into long-chain fatty acids and glycerol, as well as the reverse reaction (Jaeger and Reetz, 1998). Microbial lipases are abundant in the environment, and possess advantages including high catalytic activity and diversity, as well as ease of industrial production, stability, low production cost, and short production cycle. In practice, these lipases are screened to obtain high-efficiency lipase-producing strains.

Fatty acid esters or biodiesels are produced with natural lipases as catalysts and biological grease and short-chain alcohol as raw materials using an enzymatic method (Ha et al., 2007). Biodiesel is a renewable energy source that has a broad range of advantages, including environmental protection and improved performance over ordinary diesel, and applications. The manufacture of biodiesel by the catalysis of waste oils is an effective means of dispensing with waste oils, in addition to providing an abundance of renewable energy. Therefore, the catalytic environment of lipases must be studied *in vivo*, and the biodiesel production improved by optimizing the catalytic conditions (Kamini and Iefuji, 2001).

A genome-scale metabolic model (GSMM) is an indispensable tool for the study of microbial metabolism; GSMM adopts a systems biology approach to integrate data from genomics, transcriptomics, proteomics, and metabolomics, and has been widely used to analyze the network properties of metabolism, predict and analyze the phenotypes of organism growth, interpret experimental data (Barrett et al., 2005), and in metabolic engineering. It is generally difficult to identify the key metabolic modules contributing to lipid physiology. Reconstruction GSMM can be used to systematically analyze the function of each gene, the metabolic reaction, and model the resultant effects using flux balance analysis (FBA). Specific pathways can be understood using a model of the whole metabolic network; moreover, strain design strategies can also be used to guide metabolic engineering experiments. GSMM studies provide a new approach to investigating complex lipid metabolism.

The genome sequence of the lipase-producing strain was drafted by high-throughput sequencing. A genome-scale metabolic model was constructed from the perspective of systems biology (Mahadevan and Schilling, 2003); that is, the catalytic process and conditions of microbial lipases were simulated. These findings could help formulate guidelines for *in vitro* biodiesel synthesis catalyzed by microbial lipases.

MATERIAL AND METHODS

Strain selection and molecular characterization

Strain samples (10 g) were suspended in sterile water (20 mL), agitated, and made to stand for 10 min. The enrichment medium (yeast extract 2.0 g/L, disodium hydrogen phosphate 3.5 g/L, dipotassium phosphate 5.0 g/L, magnesium sulfate heptahydrate 0.5 g/L, sodium chloride 1.5 g/L, ammonium sulfate 2.0 g/L, olive oil 10.0 mL/L, pH 7.0) was inoculated with the supernatant (1 mL), and incubated at 28°C for 2 days. The medium was enriched thrice at 190 rpm. The bacterial suspension was diluted by gradient dilution method; the dilutions were inoculated onto tributyrin plates (tryptone 10.0 g/L, yeast extract 5.0 g/L, sodium chloride 10.0 g/L, agar 15.0 g/L, tributyrin 2.0 mL/L, pH naturally) and cultured at 28°C for 2 days. Colonies presenting obvious hydrolysis circles were picked, transferred to seed medium (glucose 20.0 g/L, ammonium sulfate 5.0 g/L, dipotassium phosphate 1.0 g/L, eEpsom salt 0.5 g/L, tryptone 25.0 g/L, olive oil 10.0 mL/L, pH 7.0), and cultured at 28°C for 1day. Genomic DNA was extracted from these cultures and used as template for the amplification of 16S rDNA. The primers used for PCR were: forward, F-5'-AGAGTTTGATCCTGGCTCAG-3' and reverse, R-5'-GGTTACCTTGTTACGACT-3'. The PCR system (50 μL) consisted of the DNA template (6 μL), forward and reverse primers (F and R; 2 μL each), enzyme mix (25 μL), and ddH_2O (15 μL). The PCR conditions were set as follows: denaturation at 94°C for 5 min; 30 cycles of denaturation at 94°C for 30 s, annealing at 55°C for 1 min, and extension at 72°C for 5 min; and a final extension at 72°C for 10 min. The PCR products were recovered using a standard PCR Purification Kit (Dongsheng Biotech Co., Ltd., Guangzhou, China) and sequenced by the GenScript Corporation [GenScript (Nanjing) Co., Ltd., Nanjing, China]. The obtained sequences were aligned with sequences obtained from GenBank using the Basic Local Alignment Search Tool (BLAST). The phylogenetic tree was built using the MEGA software.

Genome sequencing and strain identification

The whole genome of the strain was sequenced by Illumina Hiseq 2000; subsequently, an Illumina PE (~500-bp) library was constructed. The initial raw images were converted to sequences using the Base Calling program and stored as FASTQ files; these files contained the sequence information of reads and sequencing quality information. The raw data was trimmed: adaptor sequences were removed from reads; bases other than A, G, C, and T at the 5'-terminal were removed; terminal groups of reads with low sequencing quality (score <20) were removed; reads with N ~ 10% were removed; and small fragments containing adapters, and with length <25 bp after trimming were removed.

Multi-k-mer assemblies of the optimized sequences were constructed using SOAP *de novo* v2.04 (http://soap.genomics.org.cn/). The gaps in the optimized assemblies were filled and base corrections were performe dusing GapCloser v.1.12. The working principle of SOAP *de novo* is shown in Figure 1.

Multi-k-mer assemblies were evaluated based on the total length of assembly, number of scaffolds, and scaffold N50. Therefore, the optimal k-mer was chosen as the final assembly. The 16S rDNA sequences were aligned using BLASTn, and the strain was identified. The sequences were analyzed using Clustal Xv.1.8, and the phylogenetic tree built using the MEGA 6.0 software.

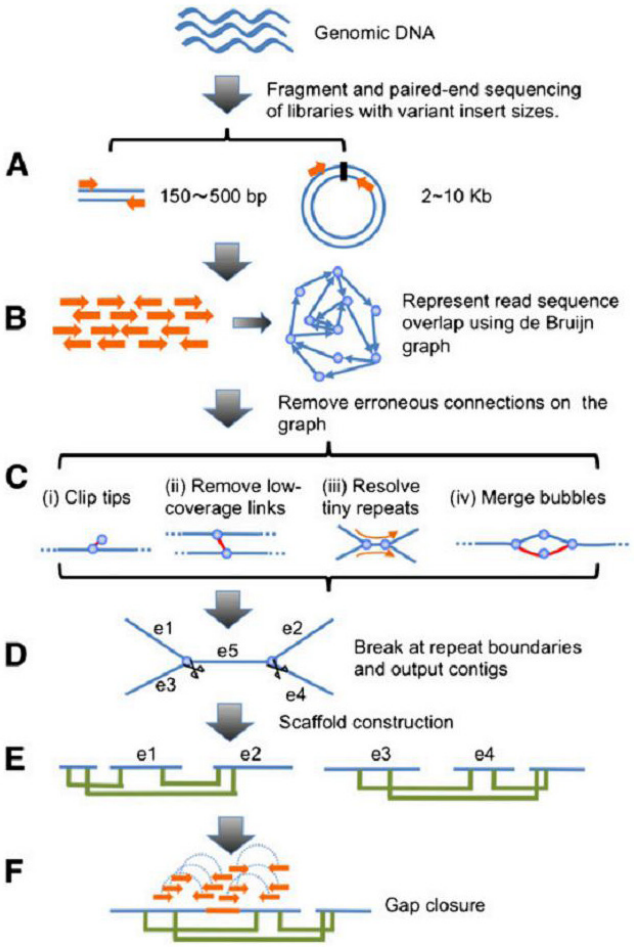

Figure 1. Schematic diagram of the principle of SOAP *de novo* (Li et al., 2010).

Functional genome annotation

Genome annotation was performed using RAST (http://rast.nmpdr.org/) (Aziz et al. 2008).

Construction, simulation, and revision of genome-scale metabolic network

A genome-scale metabolic network was constructed, simulated, and revised. The initial metabolic draft containing complete Gene-Protein-Reaction (GPR) associations was generated using MODEL SEED. The draft was later revised according to metabolic pathway data from Kyoto Encyclopedia of Genes and Genomes (KEGG) and BioCyc database. All reaction information obtained from KEGG was checked using a method described by Ma and Zeng (2003). Eleven rules, determining the direction in which the reaction would proceed, were proposed based on physiological knowledge. The genome-scale metabolic models of the two strains were constructed.

Visualization and topological analysis of the metabolic network

The metabolic network was visualized using the Cytoscape v.2.8.2 software (Shannon et al., 2003) and topological analysis was carried out using Network Analyzer. The levels of the network were analyzed, and the modules were established with MCODE.

Flux analysis of functional modules

Constraint-Based Reconstruction and Analysis (COBRA) Toolbox in Matlab was used for flux analysis (Orth et al., 2010). FBA was performed using optimize CbModel, which allowed for the prediction of the growth rate and product synthesis rate of the strain.

RESULTS

Strain selection and identification

The screening process revealed a high-efficiency lipase-producing strain (HS-L5). The samples from soil from the suburban area of Wuhan City (cultivated with castor plant) and sediments from the Jinsha River were enriched and screened twice (primary and secondary). As a result, 32 strains that utilized tributyrin were isolated, with 15 strains showing obvious hydrolysis circles on the plate. The plate for preliminary screening is shown in Figure 2.

Ten strains with a large ratio of diameter of transparent circle to colony diameter, numbered HS-L1 to HS-L10, were selected for the second screening. The ratios of diameter of transparent circle to colony diameter are summarized in Table 1.

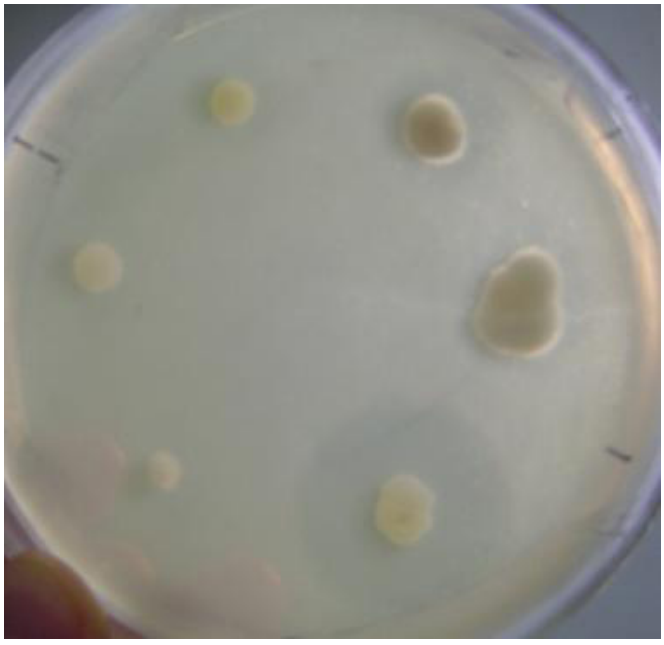

Figure 2. Preliminary screening of lipase-producing strains.

Table 1. Transparent circle method of screening for lipase-producing strains.

Strain	Diameter of strain (d) mm	Diameter of hydrolysis ring (D) mm	D/d
HS-L1	2.4	4.7	1.97
HS-L2	2.2	3.1	1.42
HS-L3	2.6	3.2	1.23
HS-L4	3.3	4.4	1.32
HS-L5	6.1	35.1	2.18
HS-L6	1.4	2.3	1.67
HS-L7	3.5	5.4	1.54
HS-L8	1.6	3.5	5.76
HS-L9	3.1	6.5	2.09
HS-L10	2.1	3.9	1.86

16S rDNA was amplified using genomic DNA of strain HS-L5 as the template and universal bacterial primers. The PCR products were purified and sequenced. The size of 16S rDNA of the HS-L5 strain was 1496 bp. This was followed by morphological characterization and physiological and biochemical identification. A phylogenetic tree was built based on alignment using the BLAST alignment tool. The homology between strain HS-L5 and *Serratia marcescens* was as high as 99%; therefore, the strain was preliminarily identified as *Serratia* sp. Sequence analysis using ClustalX 1.8 was used to build a phylogenetic tree using the MEGA 6.0 software package (Figure 3).

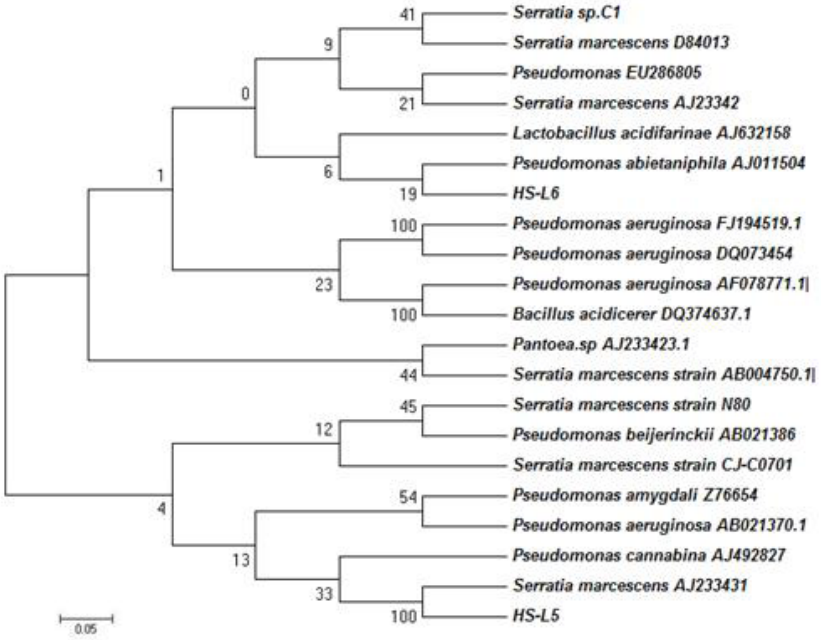

Figure 3. Phylogenetic tree showing the relationship between HS-L5 and other phylogenetically similar strains based on the 16s rDNA sequence (neighbor-joining method).

Genome sequencing

The indicators of final assemblies (Table 2) revealed that HS-L5 contained 190 scaffolds.

Table 2. Indicators of final assemblies.

	HS-L5
Number of all scaffolds	190
Bases in all scaffolds	10,583,491 bp
Number of large scaffolds (>1000 bp)	187
Bases in large scaffolds	10,581,762 bp
Largest length	427,419 bp
Scaffold N50	134,885 bp
Scaffold N90	333,55 bp
G+C content	54.196%
N rate	0.014%

Functional genome annotation

Genome annotation of HS-L5 was performed by RAST annotation (http://rast.nmpdr.org/); 4411 genes were annotated (Figure 4).

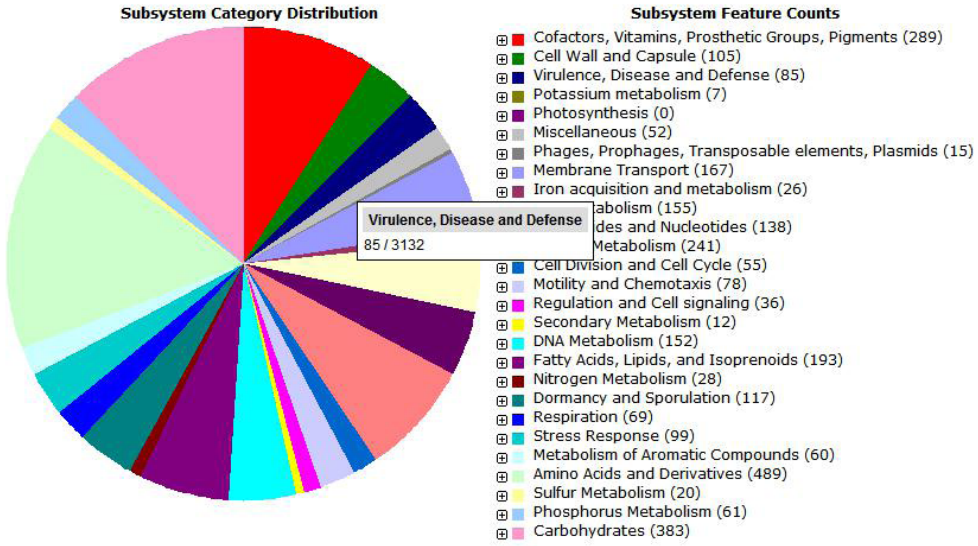

Figure 4. Results of genome annotation of HS-L5.

Construction of metabolic network

A genome-scale metabolic network was constructed, simulated, and revised. The initial metabolic draft containing complete GPR associations was generated using MODEL SEED. The draft was later revised based on metabolic pathway data from KEGG and BioCyc. All reaction information in KEGG was checked using a method described by Ma and Zeng (2003). Eleven rules determining the direction in which the reaction would proceed were proposed based on physiological knowledge. A genome-scale metabolic model was constructed for the strain HS-ML5 (Table 3). The gene coverage of the metabolic model was 19.89%.

Table 3. Parameters of gene network model of HS-L5 and HS-L8.

Strain	Total number of genes annotated	Number of reactions covered	Number of genes covered
HS-ML5	5335	1130	1061

Flow analysis for the metabolic model

The reactions related to biomass synthesis were supplemented using the reactions of *Shewanella oneidensis* MR-1. (6.9E-5) 12dag3p[c] + (0.123001) 12dgr[c] +(0.05) 5mthf[c]+ (5.0E-5) accoa[c] + (0.010152) agpe[c] + (4.21E-4) agpg[c] + (0.001) amp[c] + (220.22) atp[c] + (6.0E-6) coa[c]+(0.001608)dna [c]+(5.0E-5) fad[c] + (0.472431) glycogen[c] + (220.22) h2o[c] + (0.01114) lps[c] + (0.00215) nad[c] + (5.0E-5) nadh[c] + (1.3E-4) nadp[c] + (4.0E-4) nadph[c] + (0.069393) pe[c] + (0.027781) peptx[e] + (0.020076) pgly[c] + (4.87E-4) protein_aerobic[c]+ (0.035) ptrc[c] + (0.002818) rna [c] + (0.007) spmd[c] + (3.0E-6) succoa[c] + (0.003) udpg[c] ---> (220.22) adp[c] + (220.22) h[c] + (220.22) pi[c].

The “biomass Precursor Check” model (COBRA) in Matlab was used to check if the precursor substances for biomass synthesis in the metabolic network could be synthesized. If not, “gap Analysis” was used to identify the metabolic gaps; these were then filled by supplementing reactions from, or by a Demand Reaction search of, databases and literature. The precursor substances were thus synthesized. The upper and lower boundaries of flux were configured using exchange reactions. The lower and upper boundaries of flux were set as “-1000.0” and “1000.0”, respectively. Therefore, the absorption and secretion of substances were without limit. The maximal growth rate of strain HS-ML5 was calculated using “optimize CbModel.” If the cells could not grow, gaps in the metabolic network were identified and filled by a search of the KEGG, BRENDA, and Biocyc databases, and relevant literature. “None gene reactions” were supplemented with none GPRs, if required.

Visualization and topological analysis of the metabolic network

The metabolic model of HS-ML5 contained 2774 nodes and 3955 sides (Figure 5). Network Analyzer revealed that the average length of the path of the network was 8.068. The average length of path is an important indicator of the speed of information communication between nodes. Since metabolic networks are typical scale-free networks, they demonstrate a “small world” property (Watts and Strogatz, 1998). Most metabolites in the two metabolic networks are mutually convertible by reactions at less than 9 steps. Therefore, the changes in metabolite concentration can rapidly spread to the entire network.

The scale-free property of the complex biological network renders a high amount of robustness to the system. In other words, if nodes other than hubs are removed, the network still exists and remains connected. However, scale-free networks are vulnerable to attack by the hubs. Because of this feature, we must adhere to one principle in enhancing the model objective through biological operations: the genes or proteins to be operated on should not be hubs, while keeping as close to the hubs as possible. We calculated the degree of connection of nodes in the network (Ravasz et al., 2002; Opsahl et al., 2010). Figure 6 shows that limited nodes in the two networks have a high connection degree; that is, a small number of nodes with a high connection degree, which are linked to a myriad of nodes, dominate the entire network and maintain its structure. The same feature was presented by the topological coefficients of the networks (Figure 7).

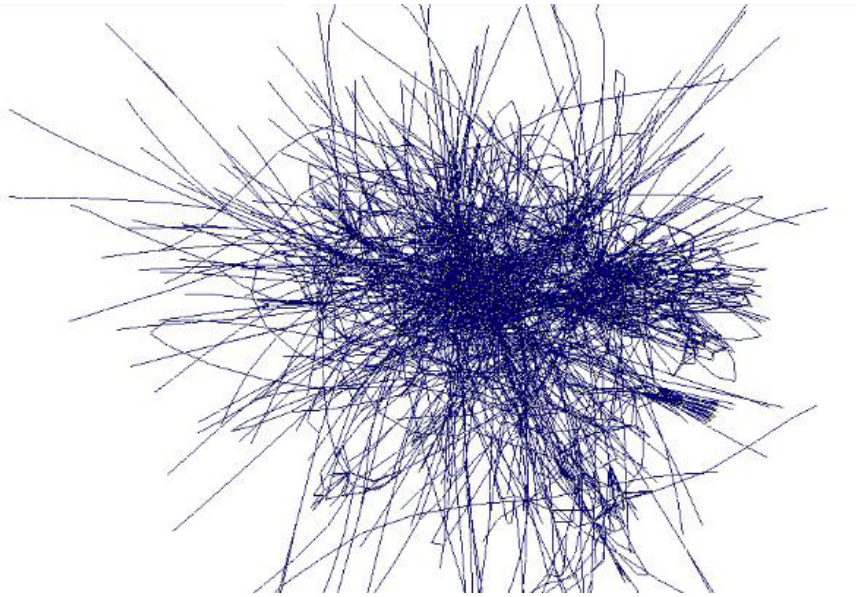

Figure 5. Visualization and topological analysis of the metabolic network.

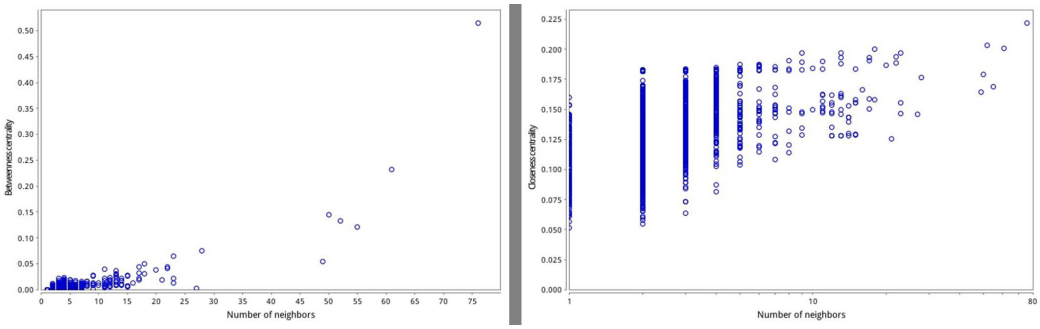

Figure 6. Connection degree of the nodes and topological coefficients for nodes in the metabolic network of HS-ML5.

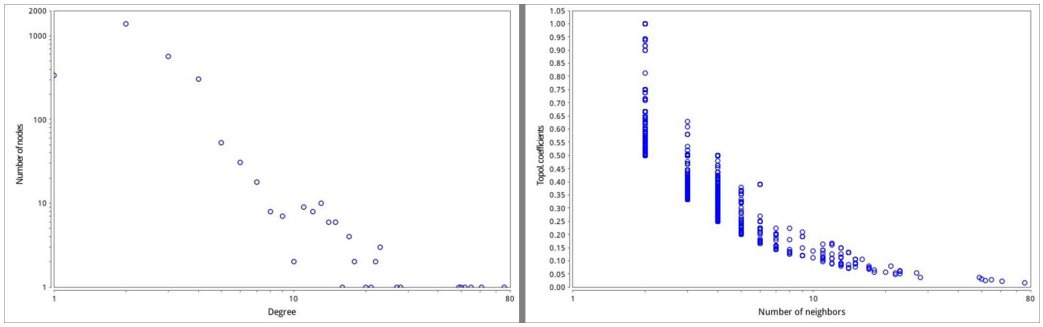

Figure 7. Betweenness centrality and closeness centrality of nodes in the metabolic network of HS-ML5.

We also calculated the betweenness centrality and closeness centrality in the two networks (Figure 6). Betweenness centrality is a measure of whether a node is the center of the network, and describes the flux going through one node in the network: the larger the betweenness centrality of the node, the larger the flux going through it (Opsahl et al., 2010). It was apparent that a few nodes in the networks shared a predominant portion of the flux; these were designated as hubs. Closeness centrality represents the length of the shortest path, and the nodes closer to the center have a higher closeness centrality. In the figure, an increase in the number of adjacent nodes led to a corresponding increase in the closeness centrality.

The two metabolic networks had scale-free properties and certain hierarchies. The subnetworks with high degree of clustering constituted of modules with biological functions.

Determination of lipid degradation subnetwork and construction of functional modules

The lipid degradation subnetwork in model HS-ML5 was identified using MCODE (Saito et al., 2012). The lipid degradation pathways were closely related to the carbon metabolism pathways, with respect to the metabolic network. The biological module of lipid degradation and related metabolic reactions, in combination with annotations, showed the presence of four major pathways and 73 reactions in both models; these were the glycolysis/gluconeogenesis (30 reactions), TCA cycle (18 reactions), alanine and aspartic acid metabolism (15 reactions), and lipid degradation (10 reactions) pathways (Figure 8).

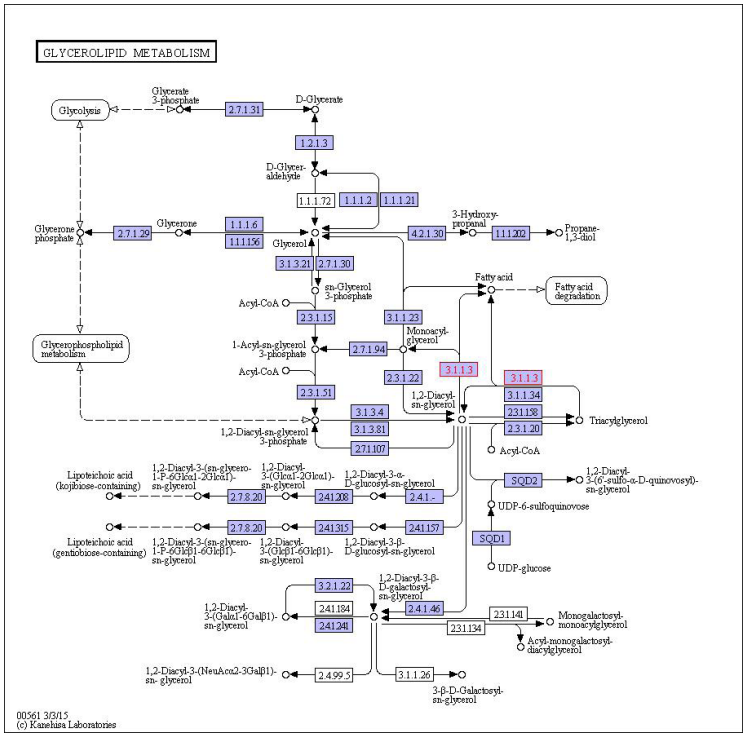

Figure 8. Lipid metabolism pathway in the metabolic network of HS-ML5.

Using the metabolic network of *Escherichia coli*, FBA was performed for this module. Minimum essential medium was used in all simulations to ensure the absorption of phosphate, sulfate, H_2O, protons, and ammonia, as well as secretions. The lower boundary of flux through exchange reactions in the network was set as -1000 mmol·gDW^{-1}·h^{-1}, at which the microbial absorption of these elements would not be restricted. Tributyrin was used as the only carbon source, and the maximum rate of carbon absorption was restricted to 10 mmol·gDW^{-1}·h^{-1}. The lower and upper boundaries of flux through other exchange reactions were set as 0 and 1000, respectively. Therefore, some small molecules produced within the cell were secreted outside the cell. The conditions of exchange reactions were configured, and the simulation experiments were initiated.

The medium containing triglyceride was used with the objective of achieving maximum fatty acid synthesis. The main factors influencing fatty acid synthesis were concentrations of triglyceride, glycerol, and acetyl CoA. Glycerol accumulation would impede fatty acid synthesis through triglyceride degradation. The synthesis of acetyl CoA is the major reason for the consumption of fatty acids (74%). Excess glycerol can lead to an increase in the synthesis of phosphomonoglyceride and diglyceride, which in turn leads to a decrease in fatty acid synthesis. Some studies have indicated that glycerol inhibits the activity of glycerol lipases.

DISCUSSION

The simulations indicated that the efficiency of fatty acid synthesis from lipids could be improved by three methods. Lipase EC3.1.1.3 catalyzes the conversion of long-chain acyl triglyceride and diglyceride into acyl monoglyceride and fatty acids, respectively, and not the conversion of acyl monoglyceride into glycerol. Therefore, lipid degradation by glycerol can be inhibited by increasing the purity of lipase EC3.1.1.3, which in turn could reduce the contamination of monoglyceride lipase EC3.1.1.23, which catalyzes the degradation of monoglyceride. Another method would be to separate the monoglyceride produced during the reaction using a new industrial design, which inhibits excess accumulation and increases the rate of degradation. Lastly, through the lipase reaction, diglyceride catabolism can only give fatty acids and glycerol as the final products, and the degradation of fatty acids can be inhibited by the addition of acetyl CoA. Thus, the removal of acetyl CoA can activate the lipase in the bacterial strain HS-ML5, as described in the present study. Moreover, the efficiency of *in vitro* industrial production using natural enzymes for catalytic reactions could be improved by process optimization and simulation of reaction environments and conditions within the organisms.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundations of China (#31371893, #31071653).

REFERENCES

Aziz RK, Bartels D, Best AA, DeJongh M, et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9: 75. http://dx.doi.org/10.1186/1471-2164-9-75

Barrett CL, Herring CD, Reed JL and Palsson BO (2005). The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states. *Proc. Natl. Acad. Sci. USA* 102: 19103-19108. http://dx.doi.org/10.1073/pnas.0505231102

Ha SH, Mai NL, Sang HL, Hwang SM, et al. (2007). Lipase-catalyzed biodiesel production from soybean oil in ionic liquids. *Enzyme Microb. Technol.* 41: 480-483. http://dx.doi.org/10.1016/j.enzmictec.2007.03.017

Jaeger KE and Reetz MT (1998). Microbial lipases form versatile tools for biotechnology. *Trends Biotechnol.* 16: 396-403. http://dx.doi.org/10.1016/S0167-7799(98)01195-0

Kamini N and Iefuji H (2001). Lipase catalyzed methanolysis of vegetable oils in aqueous medium by *Cryptococcus* spp. S-2. *Process Biochem.* 37: 405-410. http://dx.doi.org/10.1016/S0032-9592(01)00220-5

Li RQ, Zhu HM, Ruan J, Qian WB, et al. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20:265-272.

Ma H and Zeng AP (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 19: 270-277. http://dx.doi.org/10.1093/bioinformatics/19.2.270

Mahadevan R and Schilling CH (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* 5: 264-276. http://dx.doi.org/10.1016/j.ymben.2003.09.002

Opsahl T, Agneessens F and Skvoretz J (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Soc. Networks* 32: 245-251. http://dx.doi.org/10.1016/j.socnet.2010.03.006

Orth JD, Thiele I and Palsson BØ (2010). What is flux balance analysis? *Nat. Biotechnol.* 28: 245-248. http://dx.doi.org/10.1038/nbt.1614

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, et al. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297: 1551-1555. http://dx.doi.org/10.1126/science.1073374

Saito R, Smoot ME, Ono K, Ruscheinski J, et al. (2012). A travel guide to Cytoscape plugins. *Nat. Methods* 9: 1069-1076. http://dx.doi.org/10.1038/nmeth.2212

Shannon P, Markiel A, Ozier O, Baliga NS, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13: 2498-2504. http://dx.doi.org/10.1101/gr.1239303

Watts DJ and Strogatz SH (1998). Collective dynamics of 'small-world' networks. *Nature* 393: 440-442. http://dx.doi.org/10.1038/30918