Gene List of *Sphingomonas chungbukensis* DJ77

Kwon Hae-Ryong[1] mrjesus@just.chungbuk.ac.kr
Hyun-Ju Um[1] statu7@hotmail.com
Jeong-Su Oh[2] misojs@hanmail.net
Wan-Sup Cho[2] wscho@chungbuk.ac.kr
Young-Chang Kim[1] youngkim@chungbuk.ac.kr

[1] School of Life Sciences, Chungbuk National University, Cheongju 361-763, Korea
[2] Department of Management Information Systems, Chungbuk National University, Cheongju 361-763, Korea

Keywords: *Sphingomonas chungbukensis* DJ77, gene prediction, COGs

1 Introduction

We are performing the genome project for Spingomonas chungbukensis DJ77 screened from sediment by our laboratory, MBB(molecular biology and bioengineering). To efficiently analyze the data of the genome project, we developed the algorithm to make out the gene list with raw data derived from the genome project. For genome project to be accomplished well, it is definitely necessary to have not only high cost but also much time and endeavor for analyzing its result. The genome project is needed high cost, much time and endeavor for analysis of its product. To overcome this problem, we tried for industrially useful genome to be used on the initial time of the genome project, our study is focused on useful gene in the industries pointed.

2 Methods and Systems

We developed algorithm that can make a gene list automatically without whole genome sequencing. We use PHRED for each fragment's base calling, and make fasta format by using PHD2FASTA. Moreover, we trim vector sequence to obtain pure S. chungbukensis DJ77 sequence by using CROSS_MATCH. This is first step for obtaining pure fragment sequence from automatic sequencer(Fig. 1).

After tying pure fragment sequences completed in first step to fasta format, we perform BLASTX to compare them with protein database of all over the world. And then, we store all extracted hit information for results of BLASTAX from the primary database. We filter function of gene based on price of the expect, identities, word, and range among the value of a primary database. It is to be extract exactly from information which considered as the function of each fragment sequence(Fig. 1).

A secondary database is puting the information about the function of each fragment sequence and compare with the COGs database which is produced the gene which is completed the microbial genome project of 66 species. There is essential gene at the life phenomenon, because the function of the fragment sequence which matched with COGs database do not exist at the genus to be found out. And it is possible to be the species-specific gene and the genus-specific gene to the function of the fragment sequence which is not matched with COGs database.

3 Results

2908 numbers fragment same function of COGs database. Number of spcies-specific or genus-specific gene is.

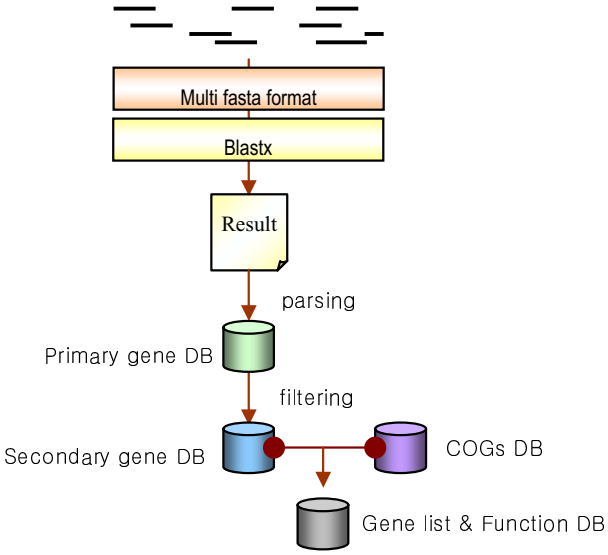

Figure 1: New system for gene prediction. Black line : DNA fragment sequences. First box is a File which is multi fasta format sequences of DJ77. Second box is blastx, annotation program. The Blastx is compare translated DNA fragment sequence with protein database. Primary gene DB is result of blastx result parsing. This DB has all data of blastx results. Secondary gene DB is result of first gene DB filtering. Filtering condition is Three. Finally gene list DB is compare second gene DB with COGs DB.

4 Discussion

1. Most of fragment is conserved function.

2. Species-specific gene, genes which biosynthesize exopolysaccharide, genes which biosynthesize sphingolipid, and genes which degrade aromatic hydrocarbon are analysize through this methods.

This work was supported by grant No.R01-2001-000-00097-0 from Korea Science & Engineering Foundation and No.DS0031 from Korea Research Foundation.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215:403–410, 1990.

[2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25:3389–3402, 1997.

[3] Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. Improved microbial gene identification with GLIMMER, *Nucleic Acids Res.*, 27(23)4636–4641, 1999.

[4] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V., The COG database: now developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res.*, 29(1):22–28, 2001.

[5] `http://www.ncbi.nlm.nih.gov/COG/`