[SVII-1]

Comparative Genome Analysis of *Sphingomonas chungbukensis* DJ77

Dang Sy Hai, Young-Pil Kim, Bum-Sun Choi, Hyun-Ju Um, Young-Chang Kim*

School of Life Sciences, Chungbuk National University

Abstract

The assemblies of our partial genomic sequence data of *Sphingomonas chungbukensis* DJ77, with the total size of 877,928 bp, was done by TIGR Assembler. The total size of our current obtained contigs was about 0.73 Mb. A comparative genome analysis between our uncompleted genome and the other completed genomes was performed by taking advantage of the availability of multiple complete genomes in COGs database (Clusters of Orthologous Groups of proteins) to produce the genomic prediction of our *S. chungbukensis* DJ77. This analysis based on homologues search among completed genomes provides good initial step to our better assigning putative function to predicted coding sequences .

Introduction

While analysis of a single genome provides tremendous biological insights on any given organism, comparative analysis of multiple genomes provides substantially more information on the physiology and evolution of microbial species and expands our ability to better assign putative function to predicted coding sequences. The comparative genome analysis was applied to our *Sphingomonas chungbukensis* DJ77 strain. *S. chungbukensis* was classified in the genus Sphingomonas and was found to be able to degrade a remarkably broad range of aromatic hydrocarbons including biphenyl, naphthalene, phenanthrene, phenol, salicylate, toluene, benzoate, etc., and their end-products consisted of sphingolipid, mucous polysaccharide, and many other unconfirmed biopolymers. These end-products were encoded by the species-specific genes and our plan is to do the gene prediction and identify the species-specific genes feature by apply the comparative genome analysis and genome annotation on our *S. chungbukensis* DJ77 strain. The shotgun library of our *S. chungbukensis* DJ77 was constructed using partial restriction digests to fragment the genomic DNA with *Eco*R1, *Bam*H1, *Hind*III and *Sau*3AI. Our genome sequencing process is on-going and the size of our current library was about 0.73 Mb. Contig assembly was performed by using TIGR Assembler and the preliminary annotation of the uncompleted genome sequence was performed by applied our developed data mining tools with the aid of NCBI ORF Finder and NCBI Cognitor.

Materials and Methods

The genome annotation process was used to 1) build the contigs, 2) annotate features. This process was complex and would continue to be refined.

The input data for the contig assembly included our DNA sequence fragments. The sequence data were first screened for contaminating sequences. Sequences are blasted against a collection of all sequences of our *S. chungbukensis* DJ77 clones for redundancy detecting. Any clone containing redundant sequence is entirely removed from the input data set.

The annotation process identified sequence features on the contigs - such as known and predicted genes, and gene models. This stage provides contig, and protein records with added feature annotation. All possible open reading frames (ORFs) in the contigs were identified by the online version of NCBI ORF Finder using the alternative genetic code 11 for bacterial.

All putative ORFs from our *S. chungbukensis* DJ77 genome were searched against multiple complete genomes from COGs database using the COGnitor program.

All the data mining steps above were carried out automatically by our rational database driven tools written in PERL scripts using BioPerl packages. Socket protocol was used to connect to the remote internet data sources and online programs.

Results and discussion

Genomic comparative analysis greatly enhances our abilities to predict and detect molecular function using sequence information. Below is some of our primitive results in annotating our *S. chungbukensis* DJ77 genome.

Table 1. General features of *S. chungbukensis* DJ77 genome and *Novosphingobium Aromaticivorum* F199

General info	DJ77	F199
Number of DNA fragments	1422	
Number of all contigs	1123	178
Number of overlapped contigs	216	
Total contig size (bp)	734495	4,191,461
G+C content (%)	59.86	65.1
Open reading frames (ORFs)		
Total number of putative ORFs	7562	3895
Maximum length of the ORF (bp)	2124	
Minimum length of the ORF (bp)	100	
Number of ORFs with the length > 1000 bp	49	
Number of ORFs with the length range from 500 bp to 1000 bp	645	

From 1422 DJ77 DNA fragments (877928 bp) as the sequence input for contig assembly, 1123 contigs were obtained which have the total size of 734495 bp, but there were only 216 contigs which were built from the overlap of more than 1 DNA fragments. The percentage of guanine plus cytosine (GC content) in the total contigs was 59.86%. There were 7562 possible open reading frames (ORFs) in 1123 contigs with the size range between 100 to 2124 bp. There were not too much ORFs which have the length longer than 1000 bp (only 49). Most of the ORFs' lengths were under 500 bp.

Table 2. The result of COGs database similarity search with 7562 putative ORFs from the *S. chungbukensis* DJ77

Number of No hits	4420
Number of No related COG (3 BeTs)	2626
Number of hits found	516
Number of COG Names (Group types)	250
Number of descriptive COG Names (Types of predicted function)	196

The COGs can be employed for annotation of newly-sequenced genomes using the COGnitor program. This program assigns new proteins to COGs by comparing them to protein sequences from all genomes included in the COG database and detecting genome-specific best hits (BeTs). When three or more BeTs fall into the same COG, the query protein is considered a likely new COG member. All putative protein sequences encoded by 7562 ORFs from our *S. chungbukensis* DJ77 genome were searched against multiple complete genomes (49) from COGs database using the COGnitor program. The similarity search result showed that there were 4420 protein sequences returned no hits and it meant that these protein were not predicted to belong to any of the currently-defined COGs. Only 516 protein sequences were found to belong to the defined COGs. These 516 proteins falled into 250 COGs and were annotated with 196 different predicted functions.

Table 3. Functional categories of *S. chungbukensis* DJ77 genome and *Novosphingobium Aromaticivorum* F199 genome using COGs data

Information storage and processing		DJ77	F199
Translation, ribosomal structure and biogenesis	J	22	156
Transcription	K	44	175
DNA replication, recombination and repair	L	32	122
Cellular processes			
Cell division and chromosome partitioning	D	2	24
Posttranslational modification, protein turnover, chaperones	O	13	105
Cell envelope biogenesis, outer membrane	M	24	100
Cell motility and secretion	N	20	55
Inorganic ion transport and metabolism	P	37	172
Signal transduction mechanisms	T	25	111
Metabolism			
Energy production and conversion	C	58	213
Carbohydrate transport and metabolism	G	27	119
Amino acid transport and metabolism	E	47	167
Nucleotide transport and metabolism	F	13	53
Coenzyme metabolism	H	18	110
Lipid metabolism	I	41	122
Secondary metabolites biosynthesis, transport and catabolism	Q	43	0
Poorly characterized			
General function prediction only	R	36	97
Function unknown	S	14	385

Each COG consists of proteins that likely share a common function or domain, which in turn has a role in a given cellular process (or processes).

Efficiency and accuracy of our genomic annotation will be improved in our next phase of the project.

References

1. Akopyants NS, Fradkov A, Diatchenko L, Hill JE, Siebert PD, Lukyanov SA, Sverdlov ED, Berg DE. 1998. PCR-based subtractive hybridization and differences in gene content among strains of Helocobater pylori. Proc Natl Acad Sci USA
2. Harayama, S., M. Rekik, A. Wasserfallen, and A. Bairoch, 1987. Evolutionary relationships between catabolic pathways for aromatics: conservation of gene order and nucleotide sequence of catechol oxidation genes of pWW0 and NAH7 plasmids. Mol. Gen. Genet. 210, 241-247.
3. JGI (Doe Joint Genome Institute) - Microbial Genomics - Novosphingobium Aromaticivorum Genome Project. http://www.jgi.doe.gov/JGI_microbial/html/sphingomonas/sphingo_homepage.html
4. Kim, S., H.-J. Shin, Y. S. Kim, S. J. Kim, and Y. C. Kim 1997. Nucleotide sequence of the Pseudomonas sp. DJ77 phnG gene encoding 2-hydroxymuconic semialdehyde dehydrogenase. Biochem. Biophys. Res. Commun. 240, 41-45.
5. Kim, S., Kweon, O.K., Kim, Y., Kim, C.-K., Lee, K.-S., and Kim, Y.C. 1997. Localization and sequence analysis of the phnH gene encoding 2-hydroxypent-2,4-dienoate hydratase in Pseudomonas sp. strain DJ77. Biochem. Biophys. Res. Commun. 238, 56-60.
6. Kim, S., Y. C. Park, C. K. Kim, J. Y. Kim, K. S. Lee, K. H. Min, and Y. C. Kim 1997. Nucleotide sequence of the phnR gene encoding Rieske-type Ferredoxin from Pseudomonas sp. strain DJ77. Kor. J. Appl Microbiol. Biotechnol. 25:367-373.
7. Kim, Y.C., K.S. Youn, M.S. Shin, H.S. Kim, M.S. Park, and H.J. Park, 1992. Molecular cloning of a gene cluster for phenanthrene degradation from Pseudomonas sp. DJ77 and its expression in Escherichia coli. Kor. J. Microbiol. 30, 1-7.
8. Kim, Y.C., M.S. Shin, K.S. Youn, Y.S. Park, and U.H. Kim, 1992. Nucleotide sequence of the phnE gene encoding extradiol dioxygenase from Pseudomonas sp. strain DJ77. Kor. J. Microbiol. 30, 8-14.
9. Michael Y Galperin and Eugene V Koonin. 1999. Searching for drug targets in microbial genomes. Current Opinion in Biotechnology. 10, 571-578
10. Frederick R. Blattner, Guy Plunkett III, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode. 1997. The complete Genome Sequence of Escherichia coli K-12. Science. 277, 1453-1462.
11. Walter M. Fitch. 2000. A personal view on some of the problems. Homology. 16(5):227-231
12. Carol J. Bult, and J. Craig Venter. 1996. Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus jannaschii. Science. (23)273, 1058-1073.
13. Roman L. Tatusov, Micheal Y. Galperin, Darren A. Natale and Eugene V. Koonin. 2000. The COG

database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acid Research. 28(1), 33-36.

14. J. Craig Venter, Mark D. Adams, and Xiaohong Zhu et al. 2001. The Sequence of human genome. Science. 291, 1304-1351
15. Micheal Y Galperin. 2001. Conserved 'hypothetical' proteins: new hints and new puzzles. Comparative and Functional Genomics. 2, 14-18.
16. Clemens Suter-Crazzolara, gunther Kurapkat. 2000. An infrastructure for comparative Genomics to functionally characterize Genes and Proteins. Genome informatics. 11, 24-32.
17. C. K. Stover, X. Q. Pham. 2000. Complete genome sequence of Psedomonas aeruginosa PAO1, an opportunistic pathogen. Nature. 406, 959-964
18. Seong-Jae Kim, Young-chang Kim. 2000. Polyphasic assignment of an aromaticdegrading Psedomonas sp., strain DJ77,. in the genus Sphingomonas as Sphingomonas chungbukensis sp. nov. IJSEM. 50, 1641-1647.
19. Mark W. Perlin, 1997, Method and system for sequencing genomes, Unites States Patent. US5604100.
20. Micheal N. Gould, 1997, Identification of differentially expressed genes, Unites States Patent. US5700644.
21. Periannan Senapathy, 1999, Method for contiguous genome sequencing, Unites States Patent. US5994058.
22. Ian Noel Hampson, 2000, Global amplification of nucleic acid, Unites States Patent. US6066457.
23. Mark W. Perlin, 2000, Method and system for sequencing genomes, Unites States Patent. US6068977.
24. Alison Abbott, 1999, A post-genomic challenge: learning to read patterns of protein systhesis, Nature, 402:715-720
25. In Kuan Cheang, Young Bae Choi, and Adrian Tang, 1994, Overview of the Structures of Heterogeneous Genome Databases, IEEE,
26. Jacqueline Courteau, 1991, Genome Databases, Science, October, vol. 254, 201-207.
27. Lincoln D. Stein, Jean Thierry-Mieg, "AceDB : A Genome Database Management System", Computer in Science & Engineering
28. Minoru Kanehisa, Post-Genome Informatics, Oxford University Press (1999).
29. Andreas D. Baxevanis and B. F. Francis Ouellette, Bioinformatics,
30. Wiley Interscience (1998).
31. Thomas Dietterich, Bioinformatics, MIT Press(1998).
32. The Genome International Sequencing Consortium, Nature 409, 860 (2001).
33. Claire M. Fraser, et al, Science 270, 397 (1995).
34. Eric S. Lander and M. S. Waterman, Genomics 2, 231 (1988).