Clustering and Comparative Analyses of Complete Genomes for the Elucidation of Evolutionary Characteristics

Jin Sik Kim[1] Sang Yup Lee[1, 2]
[1]*Department of Chemical and Biomolecular Engineering, KAIST, Daejeon, Korea*
[2]*Department of BioSystems, BioProcess Engineering Research Center and Bioinformatics Research Center, KAIST, Daejeon, Korea*
Email : jinsikkim@kaist.ac.kr, leesy@kaist.ac.kr

ABSTRACT: Three of the genus *Pseudomonas* (*P. aeruginosa, P. putida, P. syringae*) show highly different phenotypic characteristics among them. Two of the three members are pathogenic and the other is non-pathogenic. Comparative analyses of the complete genomes can elucidate the genomic similarities and differences among them. We analyzed the three genomes and the genes of them to reveal the degree of conservation of chromosomes and similarity of the genes. The 2-dimensional dot plot between the pathogenic *P. aeruginosa* and non-pathogenic *P. putida* shared higher portion of the nucleotide sequences than other two combinations. Comparison of the nucleotide compositions by calculating the genome-scale plot of G+C contents and GC skew showed the variation of nucleotide composition according to the genomic location. Comparison of the metabolic capabilities using the functional classification of KEGG orthology revealed that the differences in the number of genes for the specific functional categories resulted in the phenotypic differences. Finally combination of the analyses using the protein homologs supported the evolutionary distance of the *P. putida* obtained from other genome-scale comparisons.

1 INTRODUCTION

As the number of genome sequencing project increases, the explosion of information involved in the genome is inevitable. Currently, more than 1,200 genome projects are registered to the genome websites and 239 genomes are already published for open access (http://www.genomesonline.org). Identification of the relationship among the genomes can be important to characterize and understand unknown organisms in a point of evolutionary aspects. Evolutionary evidences of the species can be examined by analytical methods of biological data such as genomic, proteomic and metabolomic information [1]. Recently, various comparative methods have been applied to the genomes. Kitami *et al.* examined the biochemical network and duplication of genes to elucidate the effects to the genetic buffering [2]. Cooper *et al.* analyzed the phylogenetic characteristics of mammalian genomes based on the functional elements [3]. They estimated the divergences among the mammalians by counting the substitution rates in unconstrained sites. The substitution rates were calculated from the synonymous substitution, extrapolation and multiple sequence alignment of specific regions of the sequences. Kunin *et al.* have quantified the important events occurring during the genomic evolution including gene genesis, loss, and horizontal transfer [4]. The frequencies of the genomic events and effects to the genome evolution were calculated from their approaches. von Mering *et al.* analyzed *E. coli* genome to reveal the relationships between evolutionary characteristics of a genome and metabolic network [5]. They introduced the concept of gene ontology (GO) as a criterion of the functional classification and used EcoCyc database as a reference of metabolic pathways [6].

In this study, we compared the genomic contents of three complete genome sequences of pseudomonads [7, 8, 9] to reveal the genomic similarities and differences in a genomic level and to elucidate the new evolutionary characteristics based on the distribution of metabolic functional categories and that of homologous proteins.

2 METHODS

2.1 Acquisition of source genome data

We obtained the information of functional classification and list of genomes from the KEGG website (http://www.genome.ad.jp/KEGG). All the genes of the genomes were classified by functions based on the concept of the Kyoto Encyclopedia of Genes and Genomes Orthology (KO). Information of the sequences and other known/hypothetical proteins were obtained from GenBank (http://www.ncbi.nlm.nih.gov/Genbank).

2.2 2-dimensional dot matrix plot

The dot matrix plot between two genomes was generated by the Genalysis software (http://www.genetix.com/productpages/Software/Genalysis.htm). The graphical result originally displays a 2-D plot representing matched regions between two genomes above given thresholds by green colored dots. We gathered the information of each spot from the result and generated a manual plot with different colors. The minimum threshold of matched sequences was initially set to 20 base pairs.

2.3 Exploring the GC characteristics of the genomes

Genome scale map of the G+C composition and GC skew were generated by using the Artemis software [10]. The window size for the map of GC composition and GC deviation was set to 10,000 bp for the efficient representation of the features in a graph with a one-line height.

2.4 Comparison of the metabolic capabilities using orthology clustering

All the genes of the three genomes were classified by functions based on the information of the Kyoto Encyclopedia of Genes and Genomes Orthology (KO) (http://www.genome.ad.jp/kegg/kegg.html). The functional categories were divided into several steps according to the levels of metabolism. Functional categories were classified by more than 140 groups based on the KO concepts. Each functional category was curated to remove unnecessary components such as unknown genes. The phylogenetic relationships between the functional categories were compared using the hierarchical clustering analysis [11].

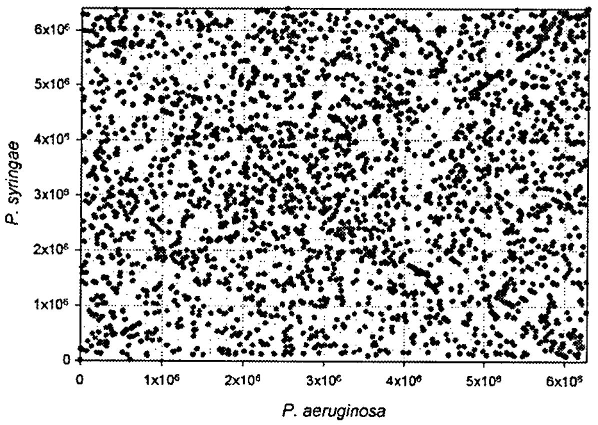
2.5 Analysis of the protein homolog

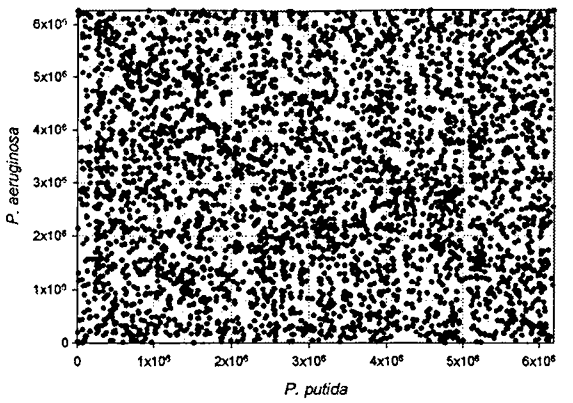
The taxonomic characteristics of the genes of the three genomes were compared using the taxonomic comparison tool of the NCBI website (http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi).

3 RESULTS AND DISCUSSION

3.1 Comparison of the 2-dimensional dot matrix plot

Two of the three genomes were compared sequentially to reveal the conserved regions between the genomes (Fig. 1). Pairwise alignment using the two dimensional dot plot method was performed.





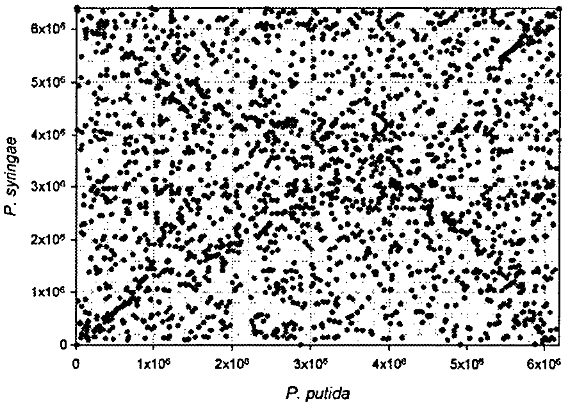


Figure 1 : Comparison of the 2-dimensional dot matrix between (a) *P. syringae* and *P. aeruginosa* (b) *P. aeruginosa* vs *P. putida* and (c) *P. syringae* vs *P. putida*. The region of synteny between two genomes appears as continuous series of matched sequences.

3.1.1 *P. aeruginosa* vs *P. syringae*.

Two pathogenic genomes were compared first (Fig. 1a). As a result, we found that 3.87% of the sequence of *P. syringae* was conserved to that of *P. aeruginosa*. 9,944 fragments with at least 20 base pairs were conserved in both sequences. Among the fragments, the longest length was found to be 473 bps.

3.1.2 *P. aeruginosa* vs *P. putida*.

Secondly, the animal pathogenic *P. aeruginosa* and non-pathogenic *P. putida* were analyzed. Compared with the above result, more than 2.5% of the two sequences share conserved regions. Along with the increase of the matched percentages, the number of matched fragments was also increased to 16,370 (Fig. 1b). However in this case, the longest region of conservation was shorter than the previous one (377 bps).

3.1.3 *P. syringae* vs *P. putida*.

Finally, the plant pathogenic *P. syringae* and non-pathogenic *P. putida* were analyzed. The degree of conservation was located between the previous two plots (4.82% conserved regions and 11,874 fragments). In this case, the longest length for the match was 453 bps.

Comparison of the linear conservation between two genomes showed that strong correlations between the arrangements were not found from all the cases. These results can be explained by the modification of genomes by important events as explained earlier such as insertion, deletion or substation to the genomic sequences. We can estimate several meaningful results involved in these events from the first and second comparisons. Although the conservation of the sequences between *P. aeruginosa* and *P. putida* were higher than that between *P. aeruginosa* and *P. syringae*, possible mutational events may cause the higher value of evolutionary distance of the former than the latter.

3.2 Comparison of the genomic map – estimation of the frequency of evolutionary events

Basic characteristics of the genomes were analyzed by calculating genome-scale G+C composition and GC skew analyses (Fig. 2). Genomic islands that can be found from the bacterial species can be explained by the result of lateral gene transfer or internal recombination [12]. The result showed that the plot of *P. aeruginosa* and *P. putida* had higher rates of deviations in the G+C contents. However, the plant pathogen *P. syringae* did not have many deviations compared with other two genomes. This result supported the explanation of the previous section. Higher frequency of the evolutionary events of *P. putida* resulted in the higher rates of the deviation of the genomic contents.

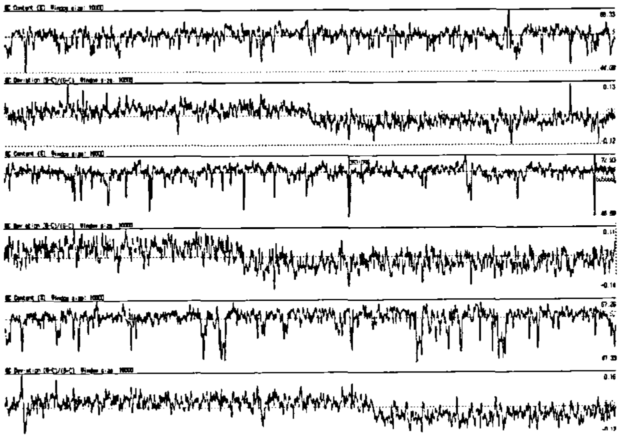


Figure 2 : Comparison of the G+C contents and GC skew of (a) *P. syringae* (b) *P. aeruginosa*, and (c) *P. putida.*

3.3 Comparison of the functional groups based on the KEGG Orthology (KO) – evolutionary evidence from the classification of genes

As a result of genome-scale comparison of the three species obtained from the previous sections, we found that the two species, *P. aeruginosa* and *P. putida* shared the most homologous genome compared with other combinations including *P. syringae*. To provide additional evidences of the genomic evolution, we introduce the concept of KEGG orthology for the classification of the genomes to the specified functional categories. This approach is based on the assumption that evolutionary events also affect the metabolic capabilities of microorganisms. The number of genes in each group was calculated and summarized in Table 1. Each functional group can be divided into more detailed metabolic categories defined in the KEGG orthology. As the three genomes were classified into functional categories, we compared the metabolic capabilities by calculating the density of genes on the specific pathways. In most functional groups, *P. aeruginosa* possesses the largest number of genes involved in the metabolism. For an example, in case of the functional group 2 (energy metabolism), *P. aeruginosa* had 190 genes, *P. putida* had 164 genes and *P. syringae* had only 143 genes. Buell *et al.* described that the important metabolism including glycolysis, TCA cycle and PP pathway of *P. syringae* were shared by other two pseudomonads [7]. The deviations of G+C contents and GC skews for the three pseudomonads were already shown in the previous section (Fig. 2). Higher deviations of G+C contents were found in both *P. aeruginosa* and *P. putida*. This phenomenon can be explained by the acquisition or loss of genes for the organisms to adapt to the environments [13].

Number of functional group and group names	PA, number of genes (PA/PS, %)	PP, number of genes (PP/PS, %)	PS, number of genes, (PS/PS, %)
1. Carbohydrate metabolism	354 (131.6%)	316 (117.5%)	269 (100.0%)
2. Energy metabolism	190 (132.9%)	164 (114.7%)	143 (100.0%)
3. Lipid metabolism	163 (191.8%)	116 (136.5%)	85 (100.0%)
4. Nucleotide metabolism	122 (111.9%)	105 (96.3%)	109 (100.0%)
5. Amino acid metabolism	549 (143.7%)	482 (126.2%)	382 (100.0%)
6. Metabolism of other amino acids	113 (163.8%)	89 (129.0%)	69 (100.0%)
7. Metabolism of complex carbohydrates	95 (114.5%)	86 (103.6%)	83 (100.0%)
8. Metabolism of complex lipids	85 (163.5%)	54 (103.8%)	52 (100.0%)
9. Metabolism of cofactors and vitamins	199 (112.4%)	183 (103.4%)	177 (100.0%)
10. Biosynthesis of secondary metabolites	15 (93.8%)	20 (125.0%)	16 (100.0%)
11. Biodegradation of xenobiotics	121 (165.8%)	97 (132.9%)	73 (100.0%)
12. Transcription	62 (140.9%)	53 (120.5%)	44 (100.0%)
13. Translation	124 (115.9%)	105 (98.1%)	107 (100.0%)
14. Sorting and degradation	101 (129.5%)	65 (83.3%)	78 (100.0%)
15. Replication and repair	75 (101.4%)	77 (104.1%)	74 (100.0%)
16. Membrane transport	374 (99.7%)	341 (90.9%)	375 (100.0%)
17. Signal transduction	9 (69.2%)	13 (100.0%)	13 (100.0%)
20. Cell motility	67 (91.8%)	66 (90.4%)	73 (100.0%)
26. Unassigned	257 (111.3%)	176 (76.2%)	231 (100.0%)

Table 1 : Distribution of genes based on the functional categories defined in the KEGG orthology

As a result of the classification, we found that three genomes had evidences of duplication events in various functional categories. Especially in case of *P. aeruginosa*, the fatty acid biosynthesis and metabolism, valine, leucine and isoleucine degradation and tryptophan metabolism showed multiple copies of genes compared with other functional categories and genomes. On the other hand, *P. putida* lacks inositol phosphage metabolism and phospholipids degradation groups. Portion of this groups are involved in the expression of pathogenic characteristics of the organisms. Loss of the genes involved in the pathogenicity in *P. putida* can be a reason for the high deviation of the G+C contents (Fig. 2). Both metabolic groups were found in the two pathogenic *P. aeruginosa* and *P. syringae* genomes. Hierarchical clustering analysis of the metabolic contents of the three organisms revealed that they showed similar pattern of distribution through the most part of the functional categories (Fig. 3).

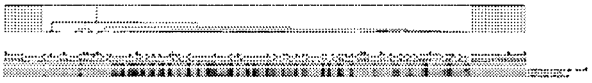


Figure 3 : Result of the hierarchical clustering of the functional categories.

3.4 Protein homologs in the genomes – the orthologous relationships of the genes among the three genomes

To find the similarities of protein sequences among the three microorganisms, we analyzed the orthologous characteristics of the genes. Three genomes share high portion of genes based on the analyses of the protein homologs. The distribution of *P. syringae* homologs showed that 5,608 query proteins produced 3,842 hits to the *P. aeruginosa* and *P. putida* proteins. Among the 3,842 hits, 2,668 hits were homologous to *P. putida* and other 1,143 were homologous to *P. aeruginosa*. Remained 31 hits were shared by two species (Fig. 4a). Similarly, we calculated the distribution of *P. aeruginosa* and *P. putida* homologs. In case of *P. aeruginosa*, 4,474 hits were distributed into *P. putida* (2,585 hits), *P. syringae* (1,832 hits) and shared hits (57 hits) (Fig. 4b). Finally, 4,227 homologs of *P. putida* were distributed into *P. aeruginosa* (1,543 hits), *P. syringae* (2,663 hits) and shared hits (21 hits) (Fig 4c). These results were summarized in Table 2. Divergence of the *P. putida* resulted in the smallest number of proteins that were homologous to other two species.

Querying species (# of queries)	*P. syringae*	*P. aeruginosa*	*P. putida*	Shared hits[a]
P. syringae (5,608)	-	2,668	1,143	31
P. aeruginosa (4,474)	1,832	-	2,585	57
P. putida (4,227)	2,663	1,543	-	21
Total	4,495	4,211	3,728	109

Table 2 : Distribution of protein homologs among the genes of the three genomes

[a]Shared hits are generated between the two species which are not querying organism.

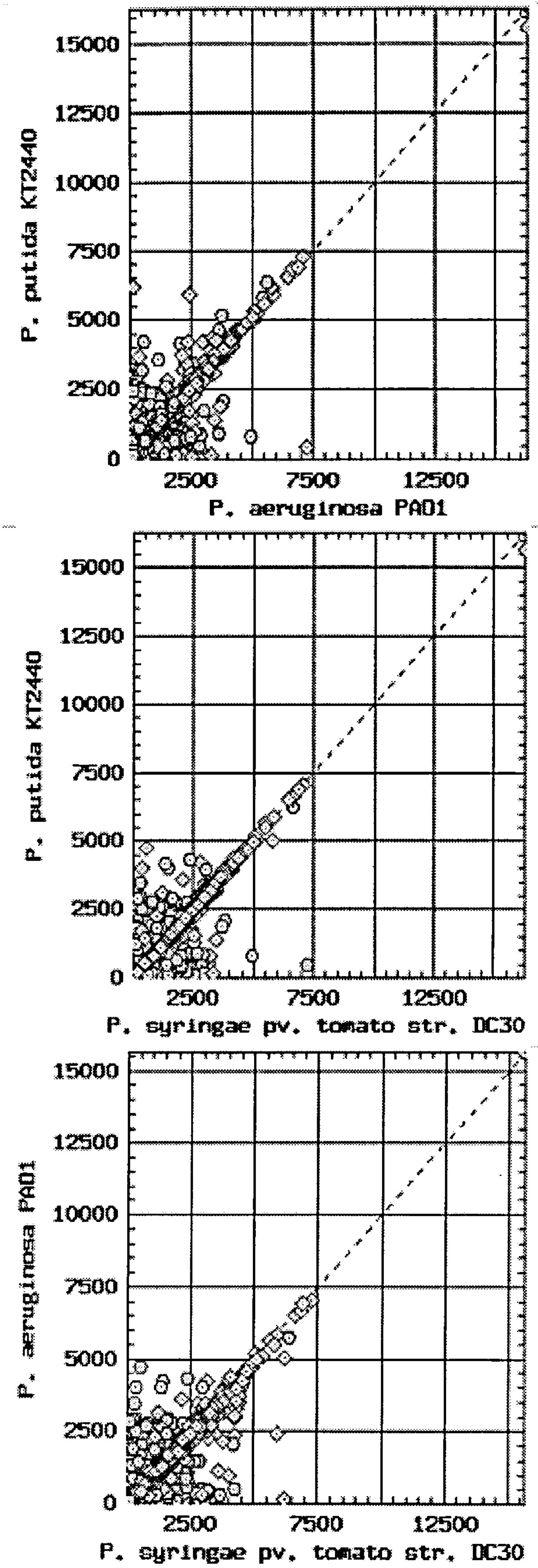


Figure 4 : Distribution of homologs among the three

genomes (a) *P. syringae* (b) *P. aeruginosa* and (c) *P. putida.*

ACKNOWLEDGEMENTS

This work was supported by a grant from the Korean Ministry of Science and Technology (Korean Systems Biology Research Grant) and by the Brain Korea 21 Project. Further supports through the LG Chemicals Chair Professorship and IBM SUR program are appreciated.

REFERENCES

[1] B.P. Kelley, R. Sharan, R.M. Karp, T. Sittler, D.E. Root, B.R. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. USA*. **100**:11394--11399, 2003.

[2] T. Kitami, and J.H. Nadeau. Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nat. Genet.* **32**:191--194, 2002.

[3] G.M. Cooper, M. Brudno, NISC Comparative Sequencing Program., E.D. Green, S. Batzoglou, and A. Sidow. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13**:81--820, 2003.

[4] V. Kunin, and C.A. Ouzounis. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* **13**:1589--1594, 2003.

[5] C. von Mering, E.M. Zdobnov, S. Tsoka, F.D. Ciccarelli, J.B. Pereira-Leal, C.A. Ouzounis, and P. Bork. Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci. USA*. **100**:15428--15433, 2003.

[6] P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, J. Collado-Vides, S.M. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro. The EcoCyc Database. *Nucleic Acids Res.* **30**:*56--58*, 2002.

[7] C.R. Buell, V. Joardar, M. Lindeberg, J. Selengut, I.T. Paulsen, M.L. Gwinn, R.J. Dodson, R.T. Deboy, A.S. Durkin, and J.F. Kolonay *et al.* The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. *Proc. Natl. Acad. Sci. USA*. **100**:10181--10186, 2003.

[8] K.E. Nelson, C. Weinel, I.T. Paulsen, R.J. Dodson, H. Hilbert, V.A. Martins dos Santos, D.E. Fouts, S.R. Gill, M. Pop, and M. Holmes *et al.* Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ. Microbiol.* **4**:799--808, 2002.

[9] C.K. Stover, X.Q. Pham, A.L. Erwin, S.D. Mizoguchi, P. Warrener, M.J. Hickey, F.S. Brinkman, W.O. Hufnagle, D.J. Kowalik, and M. Lagrou *et al.* Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**:959--964, 2000.

[10] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.A. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics* **16**:944--945, 2000.

[11] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* ***95**:14863--14868, 1998.*

[12] P. Lio, and M. Vannucci. Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics* **16**:932--940, 2000.

[13] J.G. Lawrence. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol.* **2**:519-523, 1999.

WEBSITE REFERENCES

http://www.ncbi.nlm.nih.gov; NCBI Home page
http://www.genome.ad.jp/kegg; KEGG Database Home page
http://www.genomesonling.org; GOLD^{TM} Genomes Online Database