J. Microbiol. Biotechnol. (2006), **16**(5), 748–756

JOURNAL OF MICROBIOLOGY AND BIOTECHNOLOGY

© The Korean Society for Microbiology and Biotechnology

Genomic Tree of Gene Contents Based on Functional Groups of KEGG Orthology

KIM, JIN SIK[1] AND **SANG YUP LEE**[1,2]*

[1]Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea
[2]Department of BioSystems, BioProcess Engineering Research Center and Bioinformatics Research Center, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea

Received: November 4, 2005
Accepted: January 19, 2006

Abstract We propose a genome-scale clustering approach to identify whole genome relationships using the functional groups given by the Kyoto Encyclopedia of Genes and Genomes Orthology (KO) database. The metabolic capabilities of each organism were defined by the number of genes in each functional category. The archaeal, bacterial, and eukaryotic genomes were compared by simultaneously applying a two-step clustering method, comprised of a self-organizing tree algorithm followed by unsupervised hierarchical clustering. The clustering results were consistent with various phenotypic characteristics of the organisms analyzed and, additionally, showed a different aspect of the relationship between genomes that have previously been established through rRNA-based comparisons. The proposed approach to collect and cluster the metabolic functional capabilities of organisms should make it a useful tool in predicting relationships among organisms.

Key words: KEGG, orthology, self-organizing tree algorithm, hierarchical clustering, functional category, gene contents

Increasing number of publicly available complete genomes has cleared the way for large-scale comparative analyses of the information encoded in the various genomes. At the time of this writing, more than 1,575 genome sequencing projects had either been completed or were in progress. Of these sequencing projects, 297 complete genomes had been published (Genomes Online Database; http://www.genomesonline.org). Various methods have been developed for analyzing genome data, with the practical requirements of computation time and memory constraints in mind. To date, homology analysis has primarily been used to compare genome information [4, 14]. Phylogenetic analysis using genes or conserved nucleotide sequences such as 16S rRNA has been the major method for classifying organisms [18, 25]. For example, Henson *et al.* [10] used the *nifD* gene sequences to analyze the evolutionary history of nitrogen fixation among 58 organisms, including cyanobacteria, proteobacteria, green-sulfur bacteria, and archaea. However, homologous sequences in the genomes of different organisms may not be properly aligned because of transposition, translocation, and inversion of the sequences [3, 33]. Therefore, novel and generally applicable methods that overcome the shortcomings of previous approaches are required. Since the mid-1990s, various new approaches to genome-scale analysis have been proposed, including the distinct function composition without clustering [30], gene content of unicellular species [29], whole proteome comparison [31], gene families [8], distance-based approach [2], the super-tree model [3], conservation profile model [23], phylogenetic extent model [26], metabolic pathway model [13, 22], and sequence similarity and gene content [21]. The reason for the various ways of genome tree construction was explained as the difficulties in the extension of a sequence-based approach to the genome-scale method as well as the complexities of genomes compared with genes [28].

In the present study, we analyzed the gene contents of organisms based on the functional groups defined in the Kyoto Encyclopedia of Genes and Genomes Orthology (KO). The KO concept was developed to consider information on both regulatory and metabolic pathways and to overcome the limitations of enzyme notation such as the Enzyme Commission (EC) number [15]. In the analysis, a self-organizing tree algorithm (SOTA) is first applied to the data, and then each group of organisms generated by the SOTA is clustered separately using an unsupervised hierarchical clustering method. This two-step approach is

*Corresponding author
Phone: 82-42-869-3930; Fax: 82-42-869-3910;
E-mail: leesy@kaist.ac.kr

expected to provide an efficient method for analyzing noisy genomic data. It means that the large-scale data such as metabolic contents can be classified into the clustering groups that are composed of the members with similar characteristics. Further analyses of the clusters can reveal veiled relationships among the members.

METHODS

Acquisition of Source Data

We obtained a list of genomes and information on gene functional classification from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Web site (http://www.genome.ad.jp/KEGG). All genes within the genomes were categorized by their functions, based on the KEGG Orthology (KO) information. Among the 174 genomes listed in the KEGG database, we obtained orthology information for 164 genomes of archaea, bacteria, and eukaryotes. The remaining 10 genomes were either not yet classified according to KO or not yet annotated completely to categorize the genes based on KO. Recently, the KO reorganized the hierarchy of functional classification by representing the name like "Kxxxxx," where the "xxxxx" is a number. However, in this study, we used the previous data of functional classification expressed by two-step numbers such as "1.1," because it is more useful for clustering and classifying the gene contents. The data of functional classification is given in the Supplementary Information, which is available at http://mbel.kaist.ac.kr/koanalysis/.

Preparation of Input Data for the Clustering Analysis

The number of genes in each functional category described in the KO was calculated by counting the number of genes in each category in each genome. Unassigned genes (group number 26) was initially eliminated to reduce the noise of the hierarchical clustering due to unclassified information.

KEGG Orthology (KO). KO is a recently introduced concept for identifying orthologs developed by the KEGG project [15]. KO was initially developed to solve the limitations of Enzyme Commission (EC) nomenclature and to expand the procedures to include regulatory pathways as well as metabolic pathways.

Clustering Analysis

The data were analyzed in two steps: (1) a self-organizing tree algorithm (SOTA) was applied to the data, and (2) each group of organisms generated by the SOTA was clustered separately using an unsupervised hierarchical clustering method. The data sheet was extracted in a tab-spaced text as an input of the clustering methods.

SOTA Method. The SOTA is a neural network algorithm that generates a number of clusters from an input data set. Although it was developed based on self-organizing maps (SOMs) [19], the SOTA method classifies data sets on the basis of binary topology (Fig. 1) [23]. In other words, it is a divisive algorithm. Since divisive methods proceed from the top of the hierarchical structure to the bottom, they are not useful for generating a complete partition of a data set. Rather, the SOTA is more useful for generating a highly important part of the hierarchy [5]. Generation of a binary tree from a given root point separates the data into two clusters. The tree is continuously expanded if there are more terminal nodes than those calculated from the given variability threshold conditions such as 0.90 or 0.99. The variability of a node is the maximum distance among all the profiles associated with the node and the variability threshold can be used to determine convergence of the network. We used 0.99 as the value of threshold condition in this study. An important feature of neural network algorithms that favors their use is their superior ability to handle noisy data, compared with algebraic methods such as classical hierarchical clustering. However, the SOTA has the drawback that it does not produce either a hierarchical classification or a proportional clustering at the final step of the process (Fig. 1).

Hierarchical Clustering. Each group of organisms generated by the SOTA method was clustered separately using an unsupervised hierarchical clustering method [11].

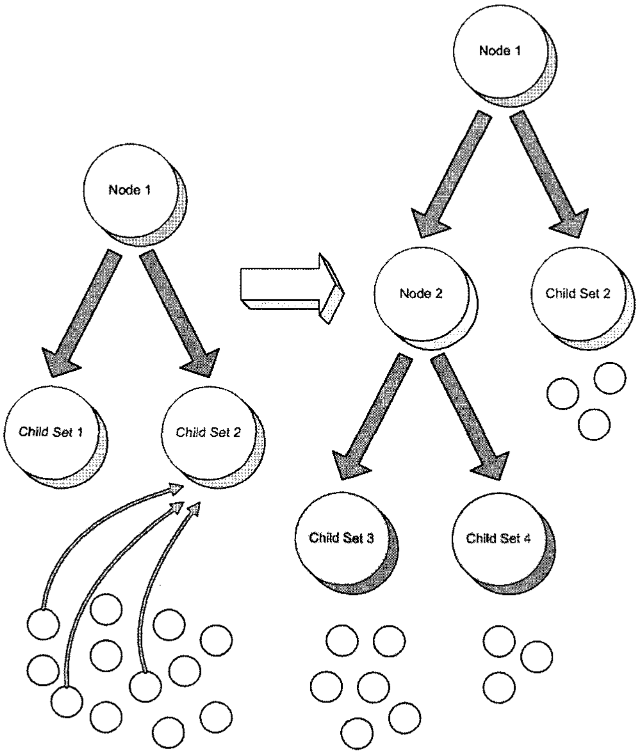

Fig. 1. Schematic diagram of the SOTA clustering algorithm. The system is initialized as a binary tree with three nodes, where the elements at each node are selected by computing the distances from each data point.

This clustering was performed using the Web-based program of the Gene Expression Pattern Analysis Suite (GEPAS) [12]. In the results of this clustering, each row represents the organism and each column represents a functional category based on the KO. The distribution levels of genes are represented on a color scale, ranging from green to red, where green corresponds to low level and red corresponds to high level. We selected the color scale representation from −20 (green) to 20 (red).

Additional Analysis by Hierarchical Clustering as a Reference Result. To establish whether the two-step clustering method is superior to the one-step hierarchical clustering method, we applied both methods to the same data set and compared the results. In this work, we used the Cluster program for the clustering analysis [7]. The results were represented as a tree structure using the TreeView program [7].

Comparison with rRNA-Based Clustering Information

We compared the results of our clustering analyses, which used the number of genes in each functional category, with previous results based on the homology of rRNA sequences [24]. We performed this comparison to determine whether the results obtained by the clustering algorithms were biologically relevant.

Table 1. List of complete genomes and taxonomic classification of the species.

Taxonomic classification			No. of organisms[a]	No. and name of KO-unavailable organisms[b]	
Eukaryotes (10)	Animals	Mammals	3		
		Fish	1	1	dre
		Insect	1		
		Nematode	1		
	Plants	Dicotyledon	1		
		Monocotyledon	1	1	osa
	Protists	Protozoa	4	3	tbr, lma, cpv
		Cellular slime mold	1	1	ddi
	Fungi	Budding yeast	1		
		Fission yeast	1		
		Microsporidia	1		
Bacteria (136)	Proteobacteria	Gamma	35	1	ypm
		Beta	8		
		Epsilon/delta	7		
		Alpha	12		
	Firmicutes	Bacillales	12		
		Lactobacillales	13		
		Clostridia	4		
		Mollicutes	8	1	poy
	Actinobacteria		13	1	cgl
	Fusobacteria		1		
	Planctomyces		1		
	Chlamydia		7		
	Spirochete		4		
	Bacteroid		2		
	Cyanobacteria		9	1	syc
	Green sulfur bacteria		1		
	Radioresistant bacteria		1		
	Hyperthermophilic bacteria		2		
Archaea (18)	Euryarchaeota		13		
	Crenarchaeota		4		
	Nanoarchaeota		1		
Total			174	10	

[a] The list of organisms for the taxonomic classification is summarized in Supplementary Information (http://mbel.kaist.ac.kr/koanalysis/).

[b] The abbreviations are as follows: dre, *Danio rerio*; osa, *Oryza sativa* japonica (Chr 1); tbr, *Trypanosoma brucei* (Chr 1, 2); lma, *Leishmania major* (Chr 1, 3); cpv, *Cryptosporidium parvum* (Chr 6); ddi, *Dictyostelium discoideum*; ypm, *Yersinia pestis* bv. Mediaevails; poy, *Phytoplasma* sp. Onion yellows; cgl, *Corynebacterium glutamicum*; syc, *Synechococcus* sp. PCC6301.

Results

Data Acquisition by Applying the SOTA Method and Hierarchical Clustering Method

We obtained data on the classified genes of 164 genomes based on the KO from the KEGG database (http://www.genome.ad.jp/kegg). The classification of genomes based on taxonomy is listed in Table 1.

The collected data were manually curated to reduce the effects of genes that do not have proper functions assigned by the annotation. By this process, the functional categories of unknown genes were removed from the input data. Figure 2 shows a phylogenetic tree constructed using the SOTA method, with upper tree options enabled, based on the functional groups defined in the KO. The degree to which each organism is involved in each functional category is expressed on a color scale, ranging from green to red.

The SOTA grouped the 164 genomes into 13 major clusters, where the smallest clusters contained just 2 elements and the largest cluster contained 51 elements. The results obtained by applying the SOTA are summarized in Table 2. We then applied unsupervised hierarchical clustering to each group identified by the SOTA method. Figure 2 shows two of the initial clusters obtained by applying the SOTA (clusters #9 and #10) along with the results obtained after applying hierarchical clustering to each of these clusters.

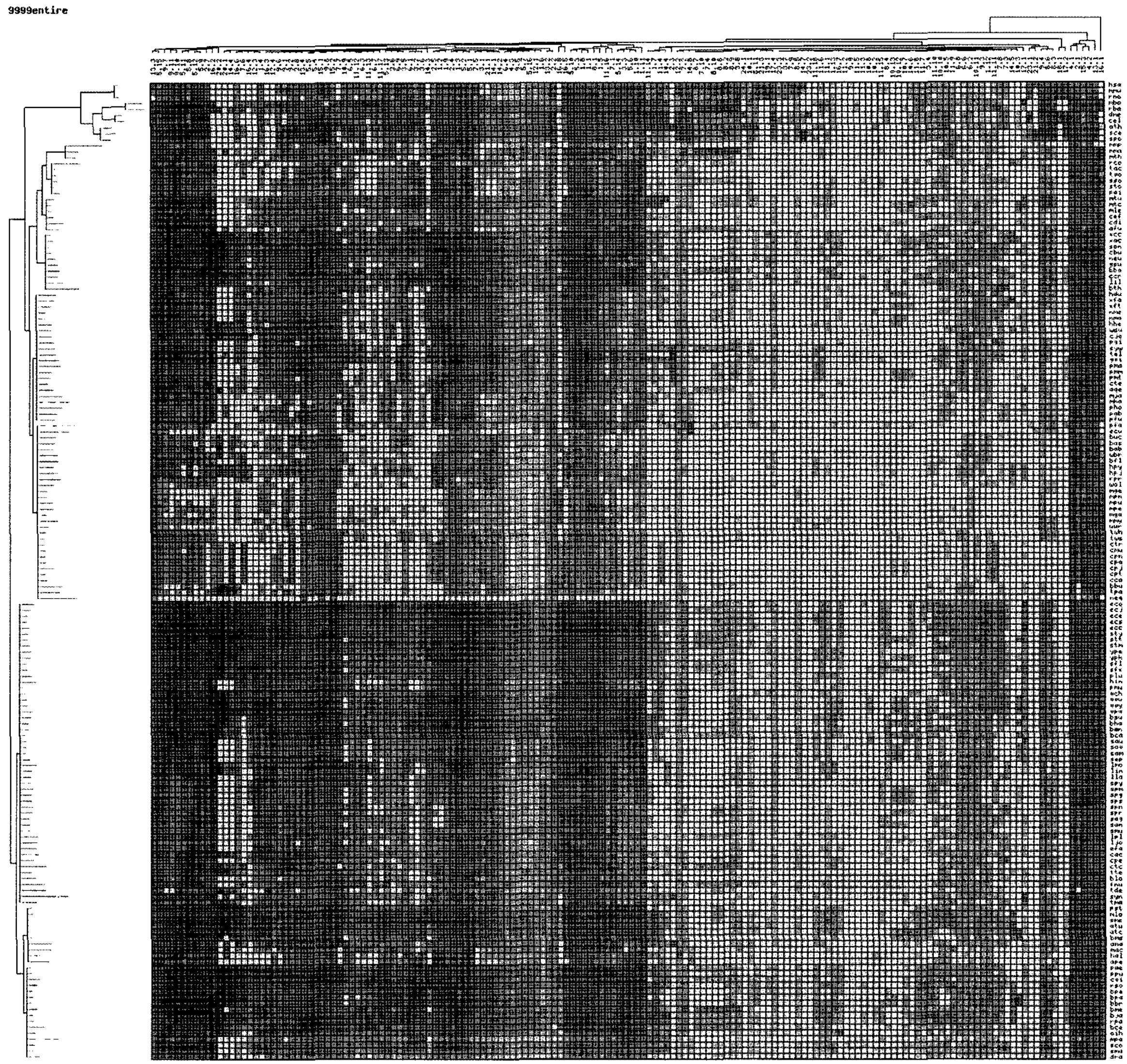

Fig. 2. Result of applying the self-organizing tree algorithm (SOTA) to the functional information of 164 genomes, using a variability threshold of 99.99%.

The numbers in the upper tree represent the functional categories defined in the KEGG orthology (KO). The functional categories were also grouped into patterns similar to that as shown in different colors. The bar at the bottom of the figure represents the color scale, ranging from green to red. The characteristics of the SOTA method are shown in the left tree of the figure. Application of the SOTA generates clustered groups rather than clustered results.

The corresponding results for the other clusters are given in Supplementary Fig. 1 (http://mbel.kaist.ac.kr/koanalysis/).

Detailed Analyses of the Clusters

In the clustering results, the major taxonomic classification did not exactly match the result of RNA-based clustering. In the unrooted phylogenetic trees reported by Nelson *et al.* [24], the taxonomic groups were clearly grouped into separate branches on the tree. In our analysis, however, the genomes of two eukaryotes assigned to cluster #10 by the SOTA, *Plasmodium falciparum* (pfa) and *Encephalitozoon cuniculi* (ecu), were not grouped into the same or derived groups from the binary trees. *P. falciparum* is a human malaria parasite whose genome was completely sequenced recently, and *E. cuniculi* is also a eukaryotic parasite [9, 17]. A decisive similarity of the genomes of these two organisms is seen, however, in their functional capacities (Figs. 2 and 3; see Supplementary Table 2; http://mbel.kaist.ac.kr/koanalysis/). Specifically, they have only a small number of genes involved in important metabolic categories such as amino acid metabolism, a characteristic they share with various organisms that are also grouped into cluster #10, such as mollicutes.

Cluster #10 contains organisms belonging to all three major taxonomic groups. The 30 organisms in this cluster include 27 bacteria, the 2 eukaryotes pfa and ecu, and the archaea *Nanoarchaeum equitans* (neq). *N. equitans* is another organism that has been completely sequenced only recently, and has been used as a model of archaeal evolution [32]. The neq genome, which is the smallest microbial genome sequenced to date (total size 491 kb), lacks various metabolic capabilities involved in the biosynthesis of amino acids, nucleotides, and cofactors. Therefore, the SOTA cluster #10 contains various parasitic organisms and organisms with

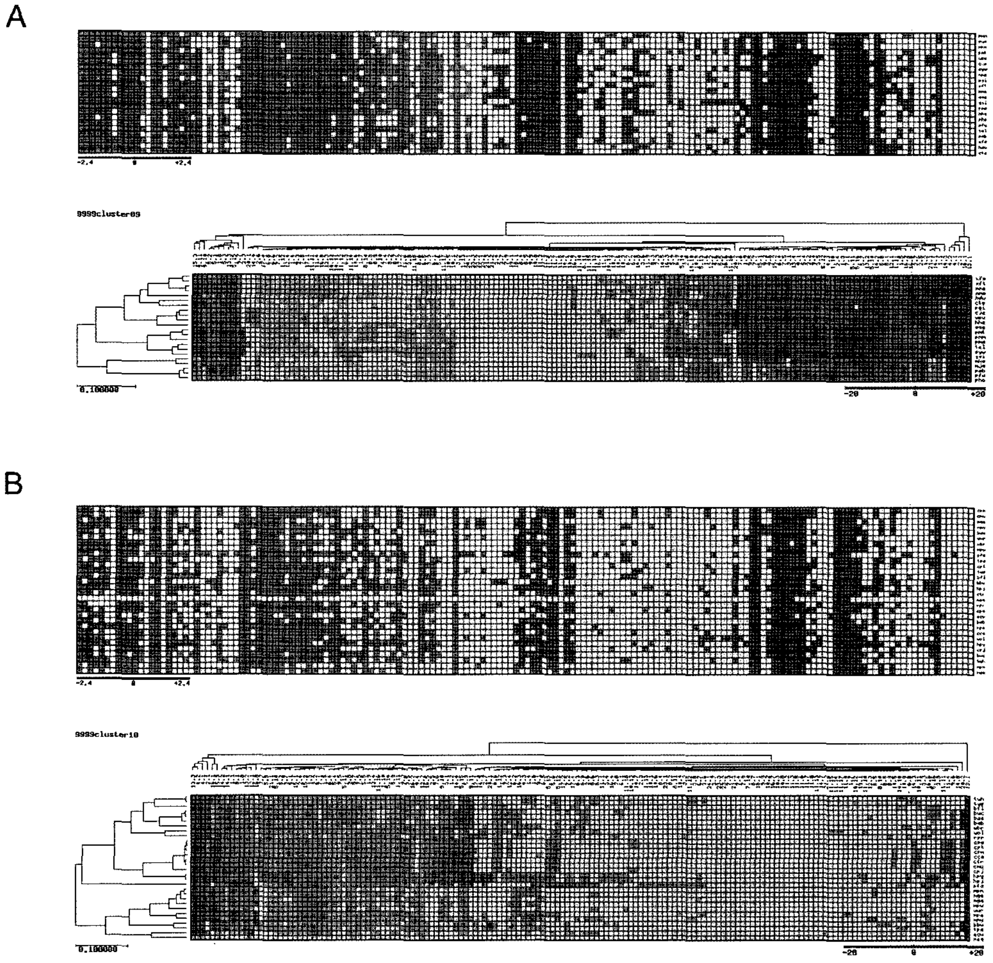

Fig. 3. Results of applying unsupervised hierarchical clustering to two clusters generated by the SOTA: cluster #9 (**A**), and cluster #10 (**B**).

Cluster #9 contains two major domains of life, bacteria and archaea, and cluster #10 contains all three major taxonomic classifiers, bacteria, archaea, and eukaryotes. The first figures of **A** and **B** show the cluster groups of organisms generated by SOTA, before application of unsupervised hierarchical clustering to these clusters. Application of the hierarchical clustering analysis generated the second figures of **A** and **B**, which show the species relationship based on the functional categories and number of genes.

Table 2. Results of the SOTA classification followed by unsupervised hierarchical clustering.

Cluster (SOTA node)	No. of elements	Organisms[a]	3 major taxonomic category[b]
#1 (node 6)	3	hsa, mmu, rno.	E
#2 (node 9)	2	mbo, rba.	B
#3 (node 10)	2	dme, cel.	E
#4 (node 7)	3	ath, sce, spo.	E
#5 (node 18)	3	mmp, mma, mth.	A
#6 (node 17)	6	rco, tac, tvo, sso, sto, pai.	B, A
#7 (node 15)	6	mtu, mtc, mle, cef, cdi, afu.	B, A
#8 (node 13)	10	xcc, xac, son, cbu, neu, gsu, bba, ccr, lil, bth.	B
#9 (node 20)	22	hdu, xfa, xft, nme, nma, hhe, wsu, cje, pgi, syw, tel, gvi, pma, pmm, pmt, cte, aae, mja, mka, pho, pab, pfu.	B, A
#10 (node 19)	30	pfa, ecu, buc, bas, bab, wbr, bfl, hpy, hpj, rpr, wol, mge, mpn, mpu, mpe, mga, mmy, uur, twh, tws, ctr, cmu, cpn, cpa, cpj, cpt, cca, bbu, tpa, neq.	E, B, A
#11 (node 22)	51	eco, ecj, ece, ecs, ecc, sty, stt, stm, ype, ypk, sfl, sfx, plu, hin, pmu, vch, vvu, vvy, vpa, bsu, bha, ban, bca, sau, sav, sam, sep, lmo, lin, lla, spy, spm, spg, sps, spn, spr, sag, san, smu, lpl, ljo, efa, cac, cpe, ctc, tte, blo, fnu, tde, syn, tma.	B
#12 (node 23)	10	pst, mlo, sme, atu, atc, bms, ana, mac, hal, ape.	B, A
#13 (node 24)	16	pae, ppu, cvi, rso, bpe, bpa, bbr, bme, bja, rpa, bce, oih, mpa, sco, sma, dra.	B
Total	164		

[a] Abbreviations are summarized in Supplementary Information (http://mbel.kaist.ac.kr/koanalysis/).
[b] A, archaea; B, bacteria; E, eukaryotes.

small genome size. These SOTA results cannot be obtained from the traditional unsupervised hierarchical clustering or multiple sequence alignment of conserved sequences.

Cluster #11 is characterized by organisms in the gamma proteobacteria and in the firmicutes, except for the mollicutes (see Supplementary Fig. 1; http://mbel.kaist.ac.kr/koanalysis/). These two divisions contain organisms with large genome sizes. Analyses of the genome of *Fusobacterium nucleatum* (fnu) revealed that the important metabolic characteristics of this organism are similar to those of two other members of Cluster #11, *Enterococcus* spp. and *Clostridium* spp. [16].

An interesting feature of the clustering results is that the 18 archaeal organisms included in the analysis are distributed across 6 SOTA clusters. This phenomenon can be explained in terms of the evolutionary relationships between archaeal and bacterial genomes. For example, Aravind *et al.* [1] established that gene exchanges occurred between archaea and hyperthermophile bacteria. This is further supported by considering cluster #9, in which the hyperthermophilic bacterium *Aquifex aeolicus* (aae) has functional similarity with the archaeal organism *Methanococcus jannashii* (mja). Among the eukaryotic organisms, the two parasites pfa and ecu were assigned to cluster #10, as mentioned above, and the remaining 8 eukaryotic organisms were assigned to three clusters (#1, #3, and #4). If, however, the SOTA was applied using a variable threshold of 0.90 instead of 0.99, the latter 8 eukaryotic organisms were assigned to the same cluster (data not shown). We selected the later variability of threshold to yield a better clustering result.

Upper Tree Comparison

Unsupervised clustering of the upper tree, using the SOTA result, showed that the gene distribution trends can be estimated from the clustering result. Figure 2 shows the grouping trends for all of the organisms used in this study, and Fig. 3 shows the detailed grouping for the organisms clustered by the SOTA method. In Fig. 2, the functional category that is most distant from the other ones is shown at the rightmost branch of the tree. The category in this position is category 16.1, which contains genes involved in the prokaryotic ABC transporter system. Most of the organisms showing low levels of genes (green squares) in this category were eukaryotes. In the middle of the column corresponding to functional category 16.1, intermediate levels of genes (dark orange squares) are observed with five organisms (buc, bas, bab, wbr, bfl), all of which are species that lack cell-component genes and regulator genes [27]. Therefore, clusters distant from other grouped clusters can be used as specific estimators for the classification of organisms. The next group in the outermost branch consists of functional categories 4.1, 2.1, 15.3, 4.2, and 13.1. Categories 4.1 and 4.2 are involved in nucleotide metabolism, 2.1 is involved in oxidative phosphorylation for energy generation, and 15.3 and 13.1 are involved in the translation process (ribosome) and replication and repair (other factors), respectively.

The clustering result indicates that the distribution levels of the genes involved in nucleotide metabolism (categories 4.1 and 4.2) are high for all of the organisms considered, and therefore, this function is insensitive to organism class. Moreover, we find that the distribution of genes involved in categories 4.1, 2.1, 15.3, 4.2, and 13.1 is clearly distinguished from those of other categories (Fig. 2).

The result of the unsupervised hierarchical clustering is apparently different from that of SOTA (Fig. 3). In particular, the functional category of the most distant branch varies among the clusters derived from the SOTA method. For example, the outermost branches in cluster #1 contain categories 17.2 and 18.1, which are involved in signal transduction and the ligand-receptor interaction, respectively, whereas the outermost branches in cluster #4 contain category 13.1, which is involved in replication and repair. The latter functional category was also located at the outer edge of the clusters obtained without unsupervised hierarchical clustering, as explained above (Figs. 2 and 3). Analysis of the outermost branches of all the clusters showed that the functional categories located in these branches of the trees in Fig. 3 and Supplementary Fig. 1 (http://mbel.kaist.ac.kr/koanalysis/) were also found in the outside branches in Fig. 2. To determine the major species-independent factors, we collated and analyzed the outermost and second outermost functional groups (Table 3).

For most of the clusters, the outermost and second outermost functional groups were those described above. Analysis of the outermost functional group can be used to relate clusters based on the distribution of functional groups on the outermost branch. A similar analysis of the functional capabilities, using only the second outermost group, does not yield useful results. However, the information on the second outermost groups can be used to reduce the uncertainty in the determination of the relationship among the groups, using the distribution of functional groups on the outermost branch. As a basis of selection, the members of the second outermost group were chosen from branches that had less than 30 leaves in the hierarchical clustering results. We used the value of 30 leaves for the meaningful selection of the branches from the data. From the distribution of functional groups in the outermost and second outermost functional groups, we can predict that clusters #11, #12, and #13 show similar clustering results. This prediction is supported by the fact that these three clusters contain predominantly prokaryotic genomes of proteobacteria and firmicutes, and that the genomes in these clusters exhibit a similar distribution of genes involved in amino acid metabolism (functional category 5), as shown in Table 3.

Clusters #5 and #6 are mainly composed of archaeal species. Cluster #5 contains three methanogenic archaea: *Methanococcus maripaludis* (mmp), *Methanosarcina mazei* (mma), and *Methanobacterium thermocutotrophicum* (mth). Among these three organisms, the genome of *M. mazei* is about 2.5 times larger than those of the other two species. Cluster #6 contains two Euryarchaeota species and three Crenarchaeota species. Along with the functional similarities obtained from the clustering analysis, the genomes of the five

Table 3. List of distant functional categories from the result of upper tree analyses.

Cluster (SOTA node)	1st distant categories	2nd distant categories
#1 (node 6)	17.2, 18.1	17.4, 4.2, 3.3, 1.1, 13.1, 5.14, 19.1, 21.2, 4.1, 2.1, 8.1, 7.1, 11.16, 9.4, 8.6, 8.3, 23.1
#2 (node 9)	11.16, 8.1, 8.6, 1.5, 9.4, 7.1	16.1, 5.14, 5.6, 1.10, 1.11
#3 (node 10)	8.6, 8.3, 11.16, 13.2, 4.2, 7.1, 17.2, 8.1, 10.4, 1.7, 11.14, 23.1, 11.4, 13.1, 4.1, 2.1	14.8, 1.1, 1.2, 2.2, 9.4, 3.3
#4 (node 7)	13.1	4.1, 2.1, 9.4, 8.3, 11.16, 8.0, 13.2, 4.2, 7.1, 10.4
#5 (node 18)	13.1, 9.7	8.1, 5.11, 1.5, 9.11, 5.12, 16.1, 4.2, 4.1
#6 (node 17)	13.1	5.15, 1.1, 2.5, 9.10, 5.3, 5.10, 5.1, 13.3, 1.2, 9.11, 10.1, 4.2, 4.1, 2.1, 1.10, 1.8, 1.11
#7 (node 15)	11.16, 2.1, 1.10, 1.11, 3.3, 15.3, 4.2, 13.1, 4.1, 16.1	-
#8 (node 13)	9.11, 4.2, 20.2, 15.3, 2.1, 16.1, 14.3, 13.1, 4.1	-
#9 (node 20)	4.2, 4.1, 2.1, 16.1, 13.1	15.3, 9.11, 9.10, 5.2, 5.1, 1.8, 5.3, 13.3, 5.15, 9.7, 2.3
#10 (node 19)	13.1	13.3, 4.2, 4.1, 2.1, 16.1, 15.3
#11 (node 22)	16.1	5.3, 1.11, 2.1, 8.1, 1.8, 1.1, 1.5, 16.2, 20.2, 2.7, 16.5, 20.2, 2.7, 16.5, 15.3, 4.2, 4.1, 16.8, 13.1
#12 (node 23)	16.1	-
#13 (node 24)	16.1	-

[a] The members in the second outermost group were chosen from the branches that had less than 30 leaves in the trees.

[b] In the case of clusters #7 and #8, the members of the second outermost group were not chosen because they had more than 35 leaves and were difficult to classify as a second outermost group compared with other remained branches.

species in cluster #6 contain sequences of homologous proteins with conserved archaeal transcriptional regulator domains [6].

DISCUSSION

By using the concept of KEGG orthology, we have related the number of genes in each functional category to the relationships among 164 complete genomes. These analyses yielded gene distribution similar to that based on the functional categories defined in the KO. In general, phylogenetic analyses using conserved sequences such as rRNA do not exactly reflect the phenotypic characteristics of the organisms. To better classify organisms, some investigations have tried to develop genome-scale analysis methods. For example, our group has used the concept of metabolic subpathways to classify organisms and compared our results with those obtained by multiple sequence alignment of 16S rRNA sequences [13]. Wolf *et al.* [33] performed tree-based analyses using orthologous genes to obtain the phenotypic characteristics of organisms. As mentioned in their review paper, the analysis of phylogeny on the genome scale should be performed with specific direction such as orthology or a method to convert the information contained in genomes to phylogenetic trees. As explained in the Results section, elimination of unnecessary genes before the clustering analyses generated meaningful results. Analyses of the upper tree structure revealed that the functional categories could be grouped into clusters based on the distribution of genes. Moreover, the clustering analysis identified several functional categories whose gene distribution levels were similar in all types of organisms.

Before performing the two-step clustering, we experimented with a one-step hierarchical clustering of the information derived from the KO (see Supplementary Fig. 2; http://mbel.kaist.ac.kr/koanalysis/), and found that the one-step hierarchical clustering was much more sensitive to the clustering conditions than the two-step approach. This could be explained by the characteristics of SOTA methods. Therefore, compared with simple hierarchical clustering, the two-step method proposed in the present study appears to be better able to handle noisy data.

The clustering results obtained by using the proposed method showed a mixing of three major kingdoms. Phylogenetic relationships, established based on conserved sequences such as 16S rRNA, separate the archaea from the bacteria, whereas genome-scale clustering classifies the species of archaea in a different manner [33]. In our analysis based on functional relationships, the archaeal species are distributed across several clusters, indicating that they have functional relationships with bacterial species. This grouping of the archaea with the bacteria, based on functional similarity, is supported by the relationship between archaea and hyperthermophile bacteria [1]. By contrast, the functional capabilities of most eukaryotic genomes are very different from those of bacterial and archaeal genomes. This difference was confirmed in the present results, which classified the eukaryotic organisms into independent clusters, except for two parasitic species. These two organisms were included in the clusters that were mainly composed of the parasitic or pathogenic species (cluster #10). In the evolutionary aspect, the horizontal gene transfers from the eukaryotes to the bacterial and archaeal species are also known to occur frequently [20]. Several pathogenic genes are known to have been transferred from the eukaryotes to the archaea and/or bacteria. For example, the ATP/ADP translocase was transferred from a plant to a bacterium, *X. fastidiosa*, as an unexpected phenomenon. The eukaryote-specific gene was originally transferred to *X. fastidiosa* and continuously passed to other pathogenic species. As shown in this case, the unexpected result can be explained by applying the clustering analyses based on the functional categories.

The methods introduced in the present work are expected to be increasingly proven reliable as the information in the KO database is further curated and qualified. For a given input data set, the two-step clustering analysis proposed herein generates more stable and meaningful results than that with one-step hierarchical clustering. The classification of KO is constantly being modified to make the concepts more rigid and robust; hence, future multistep clustering analyses of KO-based data should give more accurate and biologically meaningful results.

Acknowledgments

This work was supported by grants from the Korean Systems Biology Research Project of the Ministry of Science and Technology and by the Brain Korea 21 Project. Further support by the LG Chem Chair Professorship, Microsoft, and IBM SUR program are appreciated.

REFERENCES

1. Aravind, L., R. L. Tatusov, Y. I. Wolf, D. R. Walker, and E. V. Koonin. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14:** 442–444.
2. Clarke, G. D., R. G. Beiko, M. A. Ragan, and R. L. Charlebois. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J. Bacteriol.* **184:** 2072–2080.
3. Daubin, V., M. Gouy, and G. Perriere. 2002. A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Res.* **12:** 1080–1090.
4. Do, J. H., M. J. Anderson, D. W. Denning, and E. B. Bauer. 2004. Inference of *Aspergillus fumigatus* pathways by

computational genome analysis: Tricarboxylic acid cycle (TCA) and glyoxylate shunt. *J. Microbiol. Biotechnol.* **14:** 74–80.

5. Dopazo, J. and J. M. Carazo. 1997. Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.* **44:** 226–233.
6. Edmondson, S. P., M. A. Kahsai, R. Gupta, and J. W. Shriver. 2004. Characterization of Sac10a, a hyperthermophile DNA-binding protein from *Sulfolobus acidocaldarius*. *Biochemistry* **43:** 13026–13036.
7. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95:** 14863–14868.
8. Fitz-Gibbon, S. T. and C. H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27:** 4218–4222.
9. Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498–511.
10. Henson, B. J., L. E. Watson, and S. R. Barnum. 2004. The evolutionary history of nitrogen fixation, as assessed by NifD. *J. Mol. Evol.* **58:** 390–399.
11. Herrero, J., A. Valencia, and J. Dopazo. 2001. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* **17:** 126–136.
12. Herrero, J., F. Al-Shahrour, R. Diaz-Uriarte, A. Mateos, J. M. Vaquerizas, J. Santoyo, and J. Dopzo. 2003. GEPAS: A Web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.* **31:** 3461–3467.
13. Hong, S. H., T. Y. Kim, and S. Y. Lee. 2004. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Appl. Microbiol. Biotechnol.* **65:** 203–210.
14. Jin, J. H., U. S. Jung, J. W. Nam, Y. H. In, S. Y. Lee, D. H. Lee, and J. W. Lee. 2005. Construction of comprehensive metabolic network for glycolysis with regulation mechanisms and effectors. *J. Microbiol. Biotechnol.* **15:** 161–174.
15. Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32:** D277–D280.
16. Kapatral, V., I. Anderson, N. Ivanova, G. Reznik, T. Los, A. Lykidis, A. Bhattacharyya, A. Bartman, W. Gardner, G. Grechkin, L. Zhu, O. Vasieva, L. Chu, Y. Kogan, O. Chaga, E. Goltsman, A. Bernal, N. Larsen, M. D'Souza, T. Walunas, G. Pusch, R. Haselkorn, M. Fonstein, N. Kyrpides, and R. Overbeek. 2002. Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. *J Bacteriol.* **184:** 2005–2018.
17. Katinka, M. D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach, and C. P. Vivares. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414:** 450–453.
18. Kim, S. H., K. Y. Kim, C. H. Kim, W. S. Lee, M. Chang, and J. H. Lee. 2004. Phylogenetic analysis of harmful algal bloom (HAB)-causing dinoflagellates along the Korean coasts, based on SSU rRNA gene. *J. Microbiol. Biotechnol.* **14:** 959–966.
19. Kohonen, T. 1997. *Self-Organizing Maps*. Springer-Verlag, Berlin, Germany.
20. Koonin, E. V., K. S. Makarova, and L. Aravind. 2001. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* **55:** 709–742.
21. Kunin, V., D. Ahren, L. Goldovsky, P. Janssen, and C. A. Ouzounis. 2005. Measuring genome conservation across taxa: Divided strains and united kingdoms. *Nucleic Acids Res.* **33:** 616–621.
22. Ma, H. W. and A. P. Zeng. 2004. Phylogenetic comparison of metabolic capacities of organisms at genome level. *Mol. Phylogenet. Evol.* **31:** 204–213.
23. Martin, M. J., J. Herrero, A. Mateos, and J. Dopazo. 2003. Comparing bacterial genomes through conservation profiles. *Genome Res.* **13:** 991–998.
24. Nelson, K. E., I. T. Paulsen, J. F. Heidelberg, and C. M. Fraser. 2000. Status of genome projects for nonpathogenic bacteria and archaea. *Nat. Biotechnol.* **18:** 1049–1054.
25. Park, H. G., H. G. Ko, S. H. Kim, and W. M. Park. 2004. Molecular identification of asian isolates of medicinal mushroom *Hericium erinaceum* by phylogenetic analysis of nuclear ITS rDNA. *J. Microbiol. Biotechnol.* **14:** 816–821.
26. Peregrin-Alvarez, J. M., S. Tsoka, and C. A. Ouzounis. 2003. The phylogenetic extent of metabolic enzymes. *Genome Res.* **13:** 422–427.
27. Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407:** 81–86.
28. Snel, B., M. A. Huynen, and B. E. Dutilh. 2005. Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.* **59:** 191–209.
29. Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. *Nat Genet.* **21:** 108–110.
30. Tamames, J., C. Ouzounis, C. Sander, and A. Valencia. 1996. Genomes with distinct function composition. *FEBS Lett.* **389:** 96–101.
31. Tekaia, F., A. Lazcano, and B. Dujon. 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* **9:** 550–557.
32. Waters, E., M. J. Hohn, I. Ahel, D. E. Graham, M. D. Adams, M. Barnstead, K. Y. Beeson, L. Bibbs, R. Bolanos, M. Keller, K. Kretz, X. Lin, E. Mathur, J. Ni, M. Podar, T. Richardson, G. G. Sutton, M. Simon, D. Soll, K. O. Stetter, J. M. Short, and M. Noordewier. 2003. The genome of *Nanoarchaeum equitans*: Insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci. USA* **100:** 12984–12988.
33. Wolf, Y. I., I. B. Rogozin, N. V. Grishin, and E. V. Koonin. 2002. Genome trees and the tree of life. *Trends Genet.* **18:** 472–479.