Published online 11 March 2007

Nucleic Acids Research, 2007, Vol. 35, No. 7 ***2125–2140***
doi:10.1093/nar/gkl1114

Hierarchical classification of functionally equivalent genes in prokaryotes

Hongwei Wu, Fenglou Mao, Victor Olman and Ying Xu*

Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

Received September 22, 2006; Revised November 15, 2006; Accepted December 6, 2006

ABSTRACT

Functional classification of genes represents a fundamental problem to many biological studies. Most of the existing classification schemes are based on the concepts of homology and orthology, which were originally introduced to study gene evolution but might not be the most appropriate for gene function prediction, particularly at high resolution level. We have recently developed a scheme for hierarchical classification of genes (HCGs) in prokaryotes. In the HCG scheme, the functional equivalence relationships among genes are first assessed through a careful application of both sequence similarity and genomic neighborhood information; and genes are then classified into a hierarchical structure of clusters, where genes in each cluster are functionally equivalent at some resolution level, and the level of resolution goes higher as the clusters become increasingly smaller traveling down the hierarchy. The HCG scheme is validated through comparisons with the taxonomy of the prokaryotic genomes, Clusters of Orthologous Groups (COGs) of genes and the Pfam system. We have applied the HCG scheme to 224 complete prokaryotic genomes, and constructed a HCG database consisting of a forest of 5339 multi-level and 15 770 single-level trees of gene clusters covering ~93% of the genes of these 224 genomes. The validation results indicate that the HCG scheme not only captures the key features of the existing classification schemes but also provides a much richer organization of genes which can be used for functional prediction of genes at higher resolution and to help reveal evolutionary trace of the genes.

INTRODUCTION

Genes that have evolved from the same ancestral gene, generally called *homologs*, can be classified into two categories, *orthologs* and *paralogs*, where orthologs refer to genes that have evolved from the same ancestral gene via speciation and have maintained the same biological function, while paralogs refer to genes that have evolved from the same ancestral gene via speciation and duplication, and perform similar but distinct biological functions. Numerous research efforts have been devoted to the identification of orthologs, among which the most widely used are (a) the sequence similarity based approach such as the bi-directional best hit, the Clusters of Orthologous Groups (COGs) of proteins (1,2), and Pfam (3) and (b) gene phylogenetic tree based approach such as (4–6).

While homology and orthology are widely used concepts for studying gene evolution, our recent studies suggest that they might not necessarily be the most adequate concepts for functional prediction of genes. For example, Figure 1 shows a graphical representation of 291 genes and their functional equivalence relationships (as measured by their BLASTP e-values (7)) among each other, where each node represents a gene and each edge represents that the reciprocal BLASTP e-value between the two corresponding genes are ≤ 1.0. According to their NCBI annotations (release of March 2005), most of these 291 genes encode the two-component system's regulatory proteins of either the *sporulation* or *chemotaxis* family, and belong to four different orthologous gene groups, cheB, cheY, spo0A or spo0F. Interestingly, these genes form a few *natural* dense clusters in this graph, and the four orthologous gene groups (marked in different colors) each roughly correspond to one of these natural clusters. However, the layered structure within these clusters seems to convey much richer information than the four orthologous groups, and seems to suggest that the concepts of homology and orthology might not be adequate to capture the overall richness of the functional

*To whom correspondence should be addressed. Email: xyn@csbl.bmb.uga.edu
The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

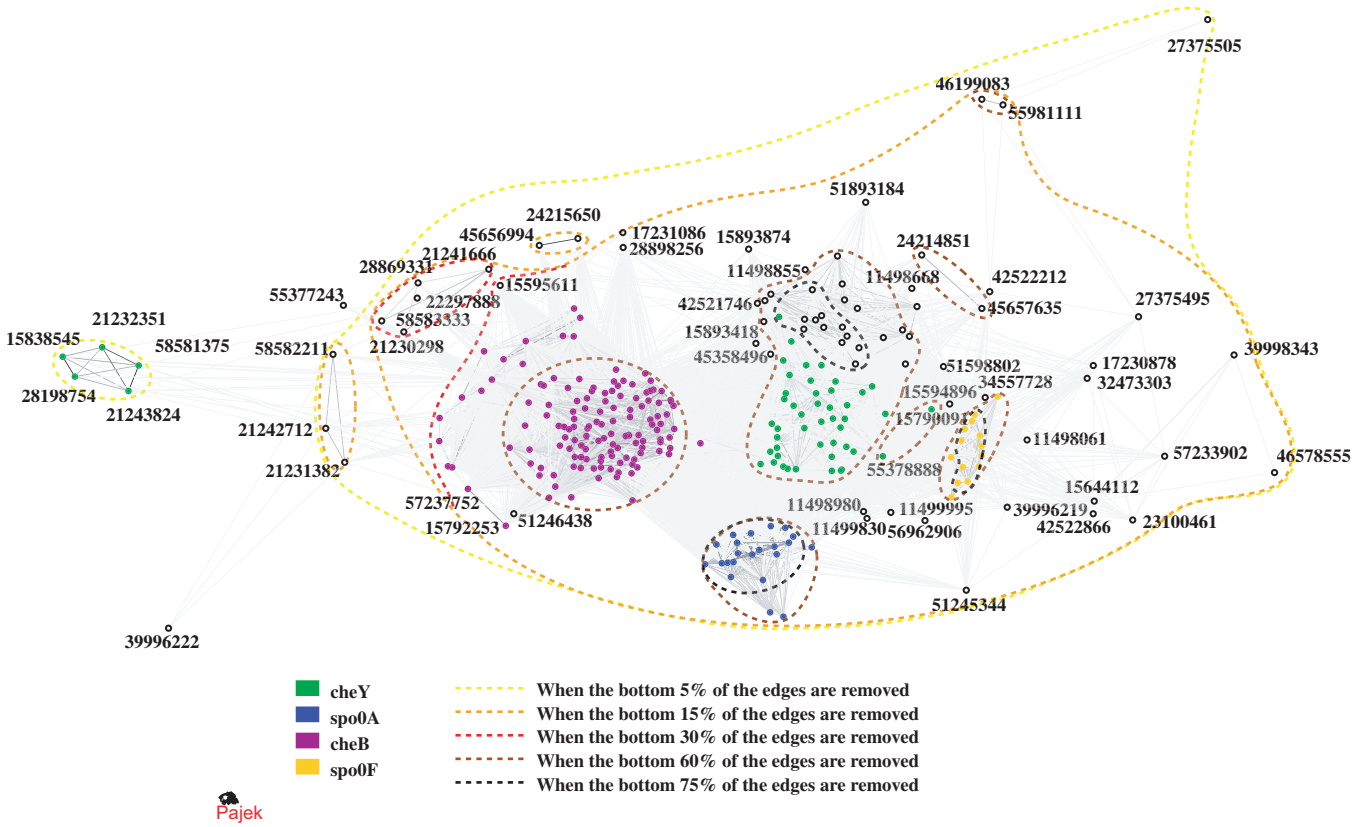

Figure 1. A graphical representation of 291 genes and their functional equivalence relationships (as measured by their BLASTP e-values). Each node represents a gene, and each edge indicates that the reciprocal BLASTP e-values between the two genes ≤ 1.0. The layout of the nodes and edges is generated by using the Pajek Software (http://vlado.fmf.uni-lj.si/pub/networks/pajek/), where the Euclidean distance between two genes and the darkness of their connecting edge are both roughly proportional to their BLASTP e-value. That is, the smaller their BLASTP e-value is, the closer their two nodes are located, and the darker their connecting edge is. Most of these genes encode the two-component system regulatory proteins of either the *sporulation* or the *chemotaxis* family. See Tables S-1.1–S-1.5 in the Supplementary Data for descriptions of those genes that do not have accompanying IDs. Based on their COG, GO, Pfam and NCBI annotations, these genes fall into five different groups, *cheB* (■), *cheY* (■), *spo0A* (■) *spo0F* (■), and genes without further specifications (■). Each dotted ellipse contains genes that form a cluster via the *guilty-by-association* rule when a certain percentage of insignificant (bottom) edges are removed, where an edge is less significant if it is associated with a higher BLASTP e-value. See Figure S-1 in the Supplementary Data for additional information of these genes and their functional equivalence relationships.

equivalence relationships among these genes. We believe that a richer framework is needed to represent the functional equivalence relationships among genes so that it can then be used for functional inference of new genes at a more detailed or higher resolution level.

We present, in this article, a framework for hierarchical classification of genes (HCGs) for prokaryotes to represent the genes' functional equivalence relationships at multi-resolution levels. We have first defined the level of functional *equivalence* between a pair of genes by using both sequence similarity and genomic neighborhood information. We have then applied this equivalence measure to all gene pairs in genomes under study, and have represented the genes and their equivalence relationships using a graph similar to the one in Figure 1. Under such a graphical representation, we consider each *densely* intra-connected sub-graph with *sparser* connections with the rest of the graph as a cluster of *functionally equivalent* genes at a certain resolution level. We have noted that the sub-graphs (clusters) naturally form a hierarchical (or tree) structure, and have therefore used a graph-theoretic technique, called a minimum spanning tree (MST)-based clustering algorithm (8,9), to identify gene clusters from this graphical representation. By applying this computational scheme to our target genomes, we have obtained a collection of functionally equivalent gene clusters, which are naturally organized as a forest. The root-level cluster of each tree represents a cluster of genes that are functionally equivalent at a certain level defined by sequence similarity alone, and a child-level cluster always contains a sub-set of genes of its parent cluster that are functionally equivalent at higher resolution than the genes in the parent cluster.

Compared to the existing methods used for functional classification of genes, e.g. the COGs of proteins (1,2) and Pfam (3), our HCG approach is unique in the following aspects. First, both the COG and Pfam are only based on the sequence similarity information of genes. Whereas, since the HCG is specifically designed for prokaryotic genomes, it combines the genomic neighborhood information, which is complementary to the sequence similarity information and reveals the functional relatedness among

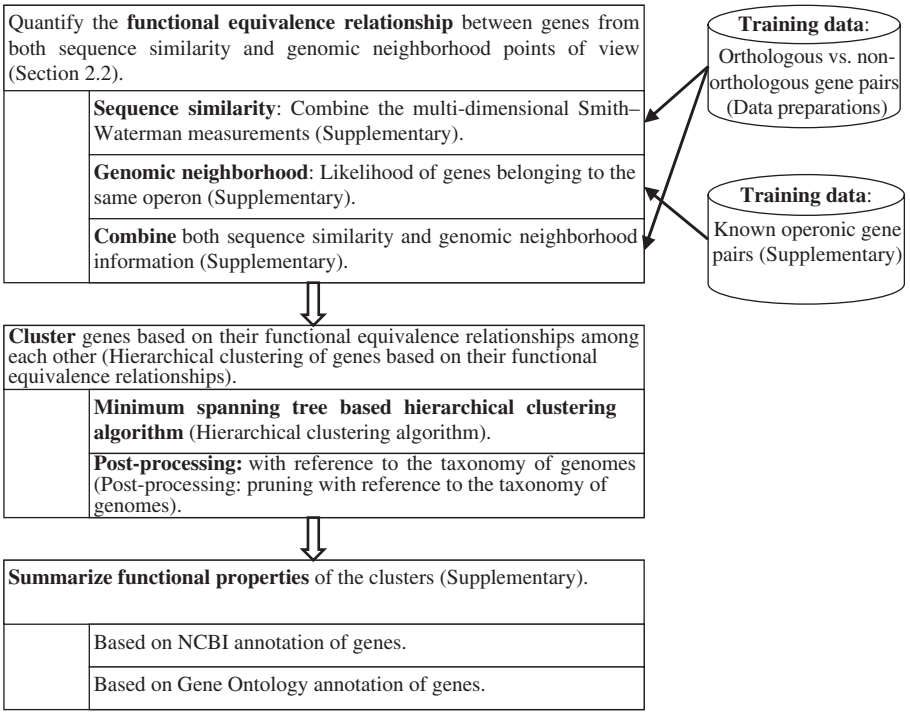

Figure 2. A flowchart of the procedure for establishing the HCG.

genes in prokaryotic genomes, with the sequence similarity information to capture the functional equivalence relationships among genes in a more comprehensive fashion. Second, unlike the Pfam, which is organized around functional or conserved domains of genes, the HCG is organized around full-length genes and captures the equivalence relationships among genes. Third and more importantly, unlike the COG and Pfam where clusters are parallel to each other though some may have overlaps but no cluster is contained inside another, the HCG adopts a hierarchical structure of clusters so that clusters are either *parallel-to* or *part-of* each other. Similar ideas of organizing clusters hierarchically have been used for structure-based protein classification such as in the SCOP database (10), whose three-level classification scheme (fold, superfamily and family) has provided much richer and more useful information than a one-level, homologous/non-homologous, structure-based classification scheme for protein structure prediction.

The validation of our gene clustering is carried out computationally through comparisons with the taxonomy of genomes, and COG and Pfam classifications. Through the comparisons, we have demonstrated that (1) the HCG classification essentially contains the information provided by the COG and Pfam classification schemes, including the functional clusterability of genes, and (2) the functional diversity and taxonomic diversity of each HCG cluster become increasingly lower as we move down from the root of a clustering tree along the HCG hierarchy, indicating that genes in each cluster are functionally equivalent at increasingly higher resolution.

MATERIALS AND METHODS

The establishment of the HCG classification is based on three key steps, where the first one is to quantify the functional equivalence relationships among genes, the second one is to cluster genes in a hierarchical manner based on their functional equivalence relationships among each other, and the last one is to functionally annotate each identified cluster. The flowchart of this procedure along with additional details is given in Figure 2. Due to the space limitations, we only provide brief descriptions for the first two steps—*scoring the functional equivalence between genes* and *hierarchical clustering of genes* and leave the details of these two steps as well as the third step, *automated annotations of clusters*, in the Supplementary Data.

Data preparations

Among the 658 174 genes in the 224 complete prokaryotic genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, release of March 2005), 610 811 genes each have at least one hit when searching against the 224 genomes using the reciprocal BLASTP with both e-values $\leq$1.0. Throughout the rest of this article, we consider only those gene pairs with both their reciprocal BLASTP e-values $\leq$1.0.

As depicted in Figure 2, our scoring scheme for functional equivalence between genes has a few parameters to be optimized. We have therefore created a positive training set consisting of known orthologous gene pairs and a negative set consisting of non-orthologous

gene pairs so that the scoring function can optimally distinguish between the two data sets. For our training sets, a gene pair (g_1, g_2) from two genomes is considered to be orthologous if and only if they are both enzymes with identical EC numbers and have exactly the same enzymatic property description in the Enzyme Database (http://ca.expasy.org/enzyme/). A gene pair $(\hat{g}_1, \hat{g}_2)$ from two genomes is considered to be non-orthologous if and only if an orthologous gene pair $(\hat{g}_1, g_2)$ can be found in the positive set such that (i) g_2 and $\hat{g}_2$ are different genes in the same genome, and (ii) the probability for g_2 and $\hat{g}_2$ to be in the same operon is negligible (Section Scoring functional equivalence), suggesting little chance for them to be duplicated genes. We have randomly selected one genome from each of the 115 genera covered by the 224 genomes, and created the positive and negative training sets from these 115 genomes. The resulting positive set consists of 218 539 orthologous gene pairs, and the negative set consists of 117 959 non-orthologous gene pairs.

Scoring functional equivalence

It should be noted that most of the existing methods for the prediction of orthologous genes, e.g. COG (2), only uses sequence similarity information. Little genomic context information has been used, though such information is complementary to the sequence similarity information and provides useful information for detection of orthologous genes, as shown in (11–16). We have observed that in our training sets, an orthologous gene pair (g_1, g_2) is much more likely than a non-orthologous gene pair to have a companion gene pair $(\hat{g}_1, \hat{g}_2)$ of high sequence similarity, with g_i and $\hat{g}_i$ $(i = 1, 2)$ being in the same genomic neighborhood (A gene g_1 is considered to be in the genomic neighborhood of another gene $\hat{g}_1$ if and only if the probability for g_1 and $\hat{g}_1$ to belong to the same operon is non-negligible See the Supplementary Data.), 20% versus 3%. This is a key piece of information that we have used to score how functionally equivalent (Note that we cannot quantify *orthology* as it is more a concept of *yes* or *no* and hence have introduced the concept of *functional equivalence*.) We have shown that genomic context information (specifically operons) can substantially help to improve the prediction accuracy of orthologous gene groups for prokaryotes (17).

Given a pair of genes (g_1, g_2), we use the following to score their functional equivalence.

$$\begin{aligned} f(g_1, g_2) &\equiv f(h(g_1, g_2), \text{genomic neighborhood of } g_1, \\ &\qquad \text{genomic neighborhood of } g_2) \\ &\equiv f\big(h(g_1, g_2), \{g_i : P((g_1, g_i) \in \text{Operon}) > 0\}, \\ &\qquad \times \{g_j : P((g_2, g_j) \in \text{Operon}) > 0\}\big) \\ &= h(g_1, g_2)\Bigg[1 + \lambda \sum_{i,j} P((g_1, g_i) \in \text{Operon}) P((g_2, g_j) \\ &\qquad \times \in \text{Operon}) I(h(g_i, g_j) \geq t_{\text{high quality}})\Bigg] \end{aligned} \quad (1)$$

In Equation (1), $h(\cdot, \cdot)$ represents a gene pair's sequence similarity—higher $h(\cdot, \cdot)$ values mean higher sequence similarities; the genomic neighborhood of a gene g consists of those (consecutively) adjacent genes g_i whose probabilities to be in the same operon of g are *non-negligible*; $\Sigma_{i,j}$ is over all gene pairs (g_i, g_j) with g_i and g_j belonging to the genomic neighborhood of g_1 and g_2, respectively; $P((\cdot, \cdot) \in \text{Operon})$ is the probability for two genes to belong to the same operon; $I(\cdot)$ is an indicator function; $t_{\text{high quality}}$ is a threshold so that a gene pair with their sequence similarity above $t_{\text{high quality}}$ is more likely to be orthologous than non-orthologous based on the sequence similarity information alone; and λ determines the relative strength of the genomic neighborhood information relative to the sequence similarity information.

The rationale for using Equation (1) to assess functional equivalence among genes lies in the following. First, we believe that a gene pair's sequence similarity information $[h(\cdot, \cdot)]$ should be a dominating factor in determining their level of function equivalence, while their genomic neighborhood information should play only a secondary role. Second, the more likely two genes belong to the same operon, the more functionally related the two genes are, and the more functional information one gene reveals about the other gene. Therefore, we have set the impact of (g_i, g_j) on the equivalence relationship between g_1 and g_2 proportional to the probability that g_1 and g_i (as well as g_2 and g_j) belong to the same operon. And finally, for g_1 and g_2, the gene pairs in their genomic neighborhoods should be reliable enough to be considered as supporting evidence for (g_1, g_2)'s equivalence relationship. Therefore, by using the indicator function and $t_{\text{high quality}}$, only those pairs that are more likely to be orthologous than non-orthologous are incorporated to enhance the functional equivalence relationship between g_1 and g_2.

A discussion on how $h(\cdot, \cdot)$ and $P((\cdot, \cdot) \in \text{Operon})$ are computed and how the relevant parameters are optimized are given in the Supplementary Data.

With the parameters of function $f(\cdot, \cdot)$ being optimized on the training data, we have achieved a classification error rate 13.00% when applying a Bayesian classifier to distinguish between the orthologous and non-orthologous training data. As a comparison, when the Bayesian classifier is performed directly on BLASTP e-values for the same purpose, the best classification error rate we can achieve is 18.47%. This indicates that $f(\cdot, \cdot)$ is better than BLASTP e-values in distinguishing between the positive and negative data.

Hierarchical clustering of genes based on their functional equivalence relationships

We use a graph $G(V^{\text{all}}, E^{\text{all}})$ to represent all the 609 887 genes of 224 genomes and their functional equivalence, where $V^{\text{all}}$and E^{all} denote the node and edge set, respectively. In this graph representation, genes are represented as nodes; and for each gene pair, the functional equivalence relationship between the two genes is represented as an edge with weight $f(\cdot, \cdot)$. Given a gene pair (g_1, g_2), the level of their functional

equivalence is reflected by both the weight on their connecting edge, $f(g_1, g_2)$, and the weights of the edges that connect g_1 (and/or g_2) with other genes g_k ($k \neq 1$ or 2), $f(g_1, g_k)$ [and/or $f(g_2, g_k)$].

Ideally $G(V^{\text{all}}, E^{\text{all}})$ should consist of a number of unconnected sub-graphs with each sub-graph representing a group of genes that are functionally equivalent at a certain level. However, we have found that some of the genes are included in the same sub-graph due to coincidental edges. Hence we have applied a graph partitioning algorithm, the Markov Cluster Algorithm (MCL) (18), to partition $G(V^{\text{all}}, E^{\text{all}})$ by removing edges sparsely linking sub-graphs, and have obtained 21 109 sub-graphs, $G(V^i, E^i)$ ($i = 1, \ldots, 21\,109$). A discussion is given in the Supplementary Data on the distribution of the number of genes being included in and the number of genomes being covered by each of these sub-graphs.

Each of these 21 109 sub-graphs can be viewed as to represent a group of genes that are functionally equivalent at a certain level. Recall from Figure 1 that within each graph $G(V^i, E^i)$ ($i = 1, \ldots, 21\,109$), some sub-sets of genes are functionally more equivalent with each other than they are with the other genes, and these sub-sets form a natural hierarchy of densely intra-connected sub-graphs.

In the rest of this article, we use the term, (sub-)graphs and clusters inter-changeably, i.e. $G(V, E)$ denotes a (sub)graph as well as a cluster.

Hierarchical clustering algorithm. Our clustering algorithm is based on a MST representation of the graph. This idea has been successfully used in the identification of regulatory binding motifs (9), clustering gene expression data (8), etc. The details are as follows:

Disassociation measure

Given a gene pair (g_1, g_2), any other gene g_k ($k = 1, 2, \ldots$) that is functionally equivalent to both g_1 and g_2 provides a piece of evidence supporting the functional equivalence relationship between g_1 and g_2, and the product $f(g_1, g_k)f(g_2, g_k)$ reflects the strength of support by g_k. By combining their equivalence scores with the supporting evidence from other genes linked to them, we define the *disassociation* measure for (g_1, g_2), $d(g_1, g_2)$, as

$$d(g_1, g_2) = \left(f^2(g_1, g_2) + \frac{\rho}{r} \sum_{k=1}^{r} f(g_1, g_{(k)}) f(g_{(k)}, g_2) \right)^{-1} \quad (2)$$

where $g_{(k)}$ is the kth ranked gene in terms of the value of $f(g_1, g_k)f(g_2, g_k)$, r is the maximum number of the supporting genes allowed to be included in, and ρ defines the level of support by g_k. The two parameters r and ρ provide the flexibility for a user to tailor the clustering method to their specific application. In this study, we have set $\rho = 0.6$, and r to be a function of the number of genes being included in $G(V^i, E^i)$ ($i = 1, \ldots, 21\,109$)

$$r = \left\lfloor \frac{\mu_0 |V^i|}{|V^i| + \mu_1} \right\rfloor \quad (3)$$

with $|\cdot|$ standing for the cardinality, $\lfloor \cdot \rfloor$ for the *floor* function, $\mu_0 = 40$ and $\mu_1 = 100$.

Cluster identification

A *spanning tree* (19) of a weighted graph $G(V, E)$ is a connected sub-graph that includes all nodes of V but does not contain any cycle, while a MST is a spanning tree with the minimum total edge-dissociation measure. Based on the definition of a cluster (9), for which the disassociation measures among neighbors within a cluster should be smaller than inter-cluster disassociation measures, our problem of identifying densely inter-connected sub-graphs can be solved by the following two-step procedure (9). As shown in Figure 3, given a graph $G(V^i, E^i)$ ($i = 1, \ldots, 21109$), the first step is to determine the sequential representation of the graph by constructing a MST using Prim's algorithm (20); and the second step is to identify valleys in this sequential representation of $G(V^i, E^i)$. It has been proved that relatively dense clusters in such a graph have a one-to-one correspondence to the valleys in its sequential representation, and the hierarchical structure among the clusters is well preserved in the hierarchical structure among the valleys. Hence all the (relatively) dense clusters can be found through finding all the valleys that stand out with high *statistical significance*. Details of this two-step procedure and related mathematical proofs can be found in (8,9).

Statistical significance of clusters

Given the sequential representation of a graph $G(V, E)$, $\{g_{(1)}, \ldots, g_{(|V|)}\}$, if there are no dense clusters (and hence valleys), then the disassociation measures between adjacent genes, $d(g_{(l)}, g_{(l+1)})$ ($l = 1, \ldots, |V|-1$), comply to the Dirichlet distribution (9,21). So the statistical significance for a sub-set of genes, $\{g_{(m)}, \ldots, g_{(n)}\}$, to form a cluster can be measured by the p-value computed for the hypothesis that the disassociation measures between adjacent genes, $d(g_{(l)}, g_{(l+1)})$ ($l = m, \ldots, n-1$), comply to the Dirichlet distribution. The smaller the p-value is, the less likely these values comply to the Dirichlet distribution, and therefore the more statistically significant that $\{g_{(m)}, \ldots, g_{(n)}\}$ form a cluster. The details for the statistical significance analysis of clusters can be found in (9).

Post-processing: pruning with reference to the taxonomy of genomes

The 224 genomes used in our study can be classified hierarchically, based on their ribosomal RNA genes and their morphological and physiological characteristics (22,23). The resulting taxonomic lineages are organized as a seven-level tree, with the tree root being *superkingdom* (SK), followed by *phylum*, *class*, *order*, *family*, *genus* and *species*. We have used this taxonomy tree as a reference to perform a post-processing step on the clustering results obtained through our MST-based clustering algorithm to make the hierarchical system of genes biologically solid. The basic idea of this post-processing is that the relationship between a parent-level cluster and its child cluster should be kept if it reflects

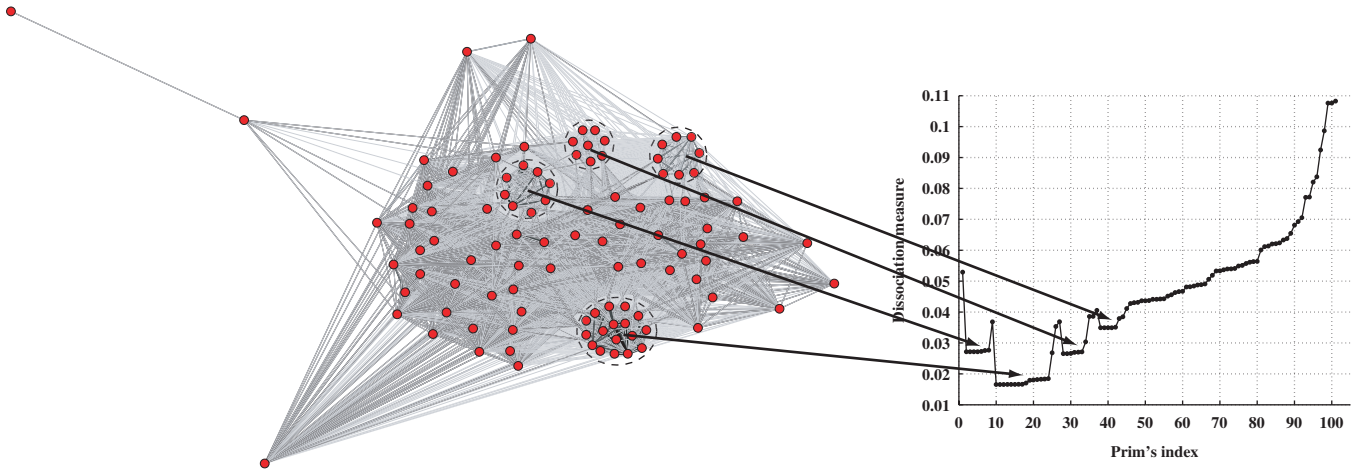

Figure 3. The MST-based hierarchical clustering algorithm: the first step is to determine the sequential representation of a graph through constructing a MST using Prim's algorithm, and the second step is to search for the valleys in the sequential representation.

a parent–child relationship from the taxonomic point of view. Specifically, a cluster $G(V, E)$ is kept if and only if $G(V, E)$ corresponds to a root-level cluster, or there are no two genes of the same genome belonging to $G(V, E)$, or the genomes covered by $G(V, E)$ are more common in terms of their taxonomic lineages than the genomes covered by $G(V, E)$'s parent cluster.

RESULTS

Using the scheme outlined in the Section Materials and methods, we have generated 51 205 clusters, covering 609 887 (~93%) of the total 658 174 genes of the 224 genomes. About 35 435 of these clusters are organized into 5339 multi-level trees covering 534 818 (87.7% = 534 818/609 887) genes, and the other 15 770 are organized into 15 770 single-cluster trees covering 75 069 gene. The root-level cluster of each tree can be considered as functionally equivalent genes at a coarse level, and as moving from the root level down to the leaf level, each cluster contains genes having functional equivalence at increasingly higher resolution. For each cluster, given its genes' NCBI annotation and the Gene Ontology (GO) annotation (http://www.ebi.ac.uk/GOA/), we have applied our annotation algorithm (which is detailed in the Supplementary Data) to select representative annotations to describe the functional features of this cluster.

Figure 4 shows the distributions of the number of clusters and the depth (i.e. the maximum number of levels) of a tree. We can observe that both distributions can be well approximated by power-law functions. When the 15 770 single-cluster trees are excluded, the number of clusters per tree ranges from 2 to 203 with the average being 6.64; and the depth of a tree ranges from 2 to 10 with the average being 3.41.

We provide a few examples here to illustrate how the HCG classification can be used to predict a gene's biological function at different resolution levels and to provide hints on the evolution trace of genes/genomes.

HCG-10: DNA-binding response regulator genes

One of the root-level clusters, HCG-10 (http://csbl.bmb.uga.edu/HCG/displaynodeclass.php?class_string=10.), contains 1730 genes from 178 bacterial genomes. Among these 1730 genes, over 99% are annotated as COG0745 (*CheY-like response regulators*) by the COG on-line system (COGNITOR) (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html); about 90% of them are annotated as *DNA-binding* or *regulator genes* by NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, release of March 2005); about 83% have GO annotations (http://www.ebi.ac.uk/GOA/) and are all annotated as GO:0003677 (*DNA binding*) and GO:0000156 (*two-component response regulator activity*); and about 83% have Pfam annotations (http://www.sanger.ac.uk/Software/Pfam/) and are all annotated to contain domains of PF00072 (*response regulator receiver domain*) and PF00486 (*transcriptional regulatory protein, C terminal*). If a gene is predicted to belong to HCG-10, we can predict that this gene is likely to be a DNA-binding response regulator gene.

It should be noted that genes of HCG-10 are further grouped into 81 sub-clusters organized as a seven-level hierarchical structure, as shown in Figure 5. As detailed below, we have observed that genes of each descendent cluster generally share a higher level of commonality than genes of its parent cluster from both the functional and evolutionary points of view.

We can take the proteobacterial genomes as an example. Their regulator genes of the two-component systems in response to different environmental conditions are further grouped into the following first-level sub-clusters of HCG-10, namely, (a) phoB genes (of phoR–phoB, in response to phosphate) (24) in HCG-10.2, (b) kdpE genes (of kdpD–kdpE, in response to potassium) (25) in HCG-10.3, (c) irlR/yedW/cusR/czcR/copR genes (of irlS–irlR, yedV–yedW, copS–copR, cusS–cusR, and czcS–czcR, respectively, in response to heavy metal) (26–29) in HCG-10.4, (d) feuP/phoP genes (of feuP–feuQ and phoQ–phoP, in response to iron and

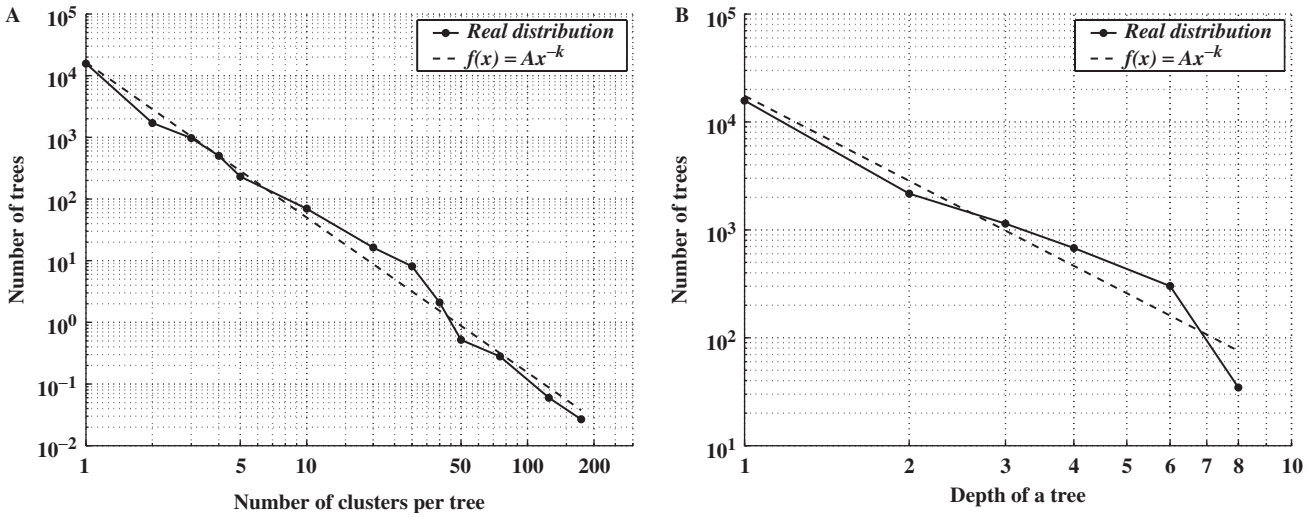

Figure 4. (**A**) Distribution of the number of clusters per tree, where the parameters for the power-law function are $A = 16\,379$ and $k = 2.51$, and the correlation coefficient between the power-law function and the real distribution curve is greater than 0.995; and (**B**) distribution of the depth of a cluster tree, where the parameters for the power-law function are $A = 17\,467$ and $k = 2.62$, and the correlation coefficient between the power-law function and the real distribution curve is greater than 0.969.

magnesium, respectively) (30,31) in HCG-10.5, (e) ompR genes (of envZ–ompR, in response to osmotic stimuli) (32) in HCG-10.6, (f) basR/qseB genes (of basE–basR and qseC–qseB, respectively) in HCG-10.7, (g) torR/arcA genes (of torS–torR and arcB–arcA, respectively, both related to respiration) (33,34) in HCG-10.8, (h) cpxR genes (of cpxA–cpxR, in response to envelope stress) (35) in HCG-10.9, (i) ctrA genes (related to the cell cycle) (36) in HCG-10.12.0, (j) baeR genes (of baeR–baeS) in HCG-10.13, (k) rstA genes (of rstA–rstB) in HCG-10.15, (l) creB genes (of creB–creC) (37) in HCG-10.18, and (k) colR genes (of colR–colS) (38) in HCG-10.24. This indicates that if an uncharacterized gene from a proteobacterial genome is predicted to belong to one of these descendent clusters, we will be able to infer to which particular two-component system this gene belongs, and hence provide higher resolution functional information than what the COG, GO and Pfam classifications can provide, since based on this we may only be able to infer that this gene is a regulator of a two-component system.

We can further go down the classification hierarchy to get more specific information about a group of functionally equivalent genes. For example, one of the child clusters, HCG-10.7, contains 43 genes that are annotated as either pmrA (of pmrB–pmrA) or ygiX (of ygiY–ygiX), where pmrA (a.k.a. baeR of baeS–baeR) genes have been found to be involved in multi-drug resistance (39), and ygiX (a.k.a. qseB of qseC–qseB) genes have been found to be involved in the regulation of flagella and motility genes (40). The functional equivalence relationships among these 43 HCG-10.7 genes are shown in Figure 6. We have observed that the $h(\cdot,\cdot)$ value for every pair of HCG-10.7 genes is above the threshold $t_{\text{high quality}}$, suggesting that all the HCG-10.7 genes are more likely to be orthologous than non-orthologous. In fact, the pmrA–pmrB and the ygiX–ygiY two-component systems exhibit high sequence identity in several genomes (41). However, the pmrA–pmrB and the ygiX–ygiY are not exactly identical, since it has been found that the PmrD protein may bind to phospho-PmrA inhibiting phospho-PmrA's dephosphorylation but does not have any effect on phospho-ygiX (41). The subtle difference between pmrA and ygiX genes, which are not reflected by any of the existing classification schemes such as COG or GO, is well captured by the two child clusters of HCG-10.7, HCG-10.7.0 and HCG-10.7.1, which contain pmrA- and ygiX-annotated genes, respectively. So, given a gene that is predicted to belong to the cluster HCG-10.7.1, we can infer that this gene is likely to be related to the regulation of flagella and motility genes, which is clearly more detailed than this gene's COG or GO annotation.

In addition, the level of functional equivalence between genes, as organized in a hierarchical fashion, may also reveal some evolutionary trace of genes, and can therefore be used for the prediction of orthology/paralogy relationship among genes. For example, one of the child clusters of HCG-10, HCG-10.0, contains 76 genes from 40 different genomes. Among these 40 genomes, 23 genomes each have multiple genes included in HCG-10.0. In particular, eight genomes from the S*taphylococcus* genus (including *Staphylococcus aureus* subsp. *aureus* COL, *S. aureus* subsp. *aureus* MW2, *S. aureus* subsp. *aureus* Mu50, *S. aureus* subsp. *aureus* N315, *S. aureus* subsp. *aureus* MRSA252, *S. aureus* subsp. *aureus* MSSA476, *S. epidermidis* ATCC 12228, and *S. epidermidis* RP62A, see Table S-5.1 in the Supplementary Data for the genes of these genomes and their HCG clustering results) each have two genes included in HCG-10.0; and eight genomes of the *Bacillus* genus (including *Bacillus anthracis* str. A2012, *B. anthracis* str. Ames, *B. anthracis* str. 'Ames Ancestor', *B. anthracis* str. Sterne, *B. cereus* ATCC 14579, *B. cereus* ATCC 10987,

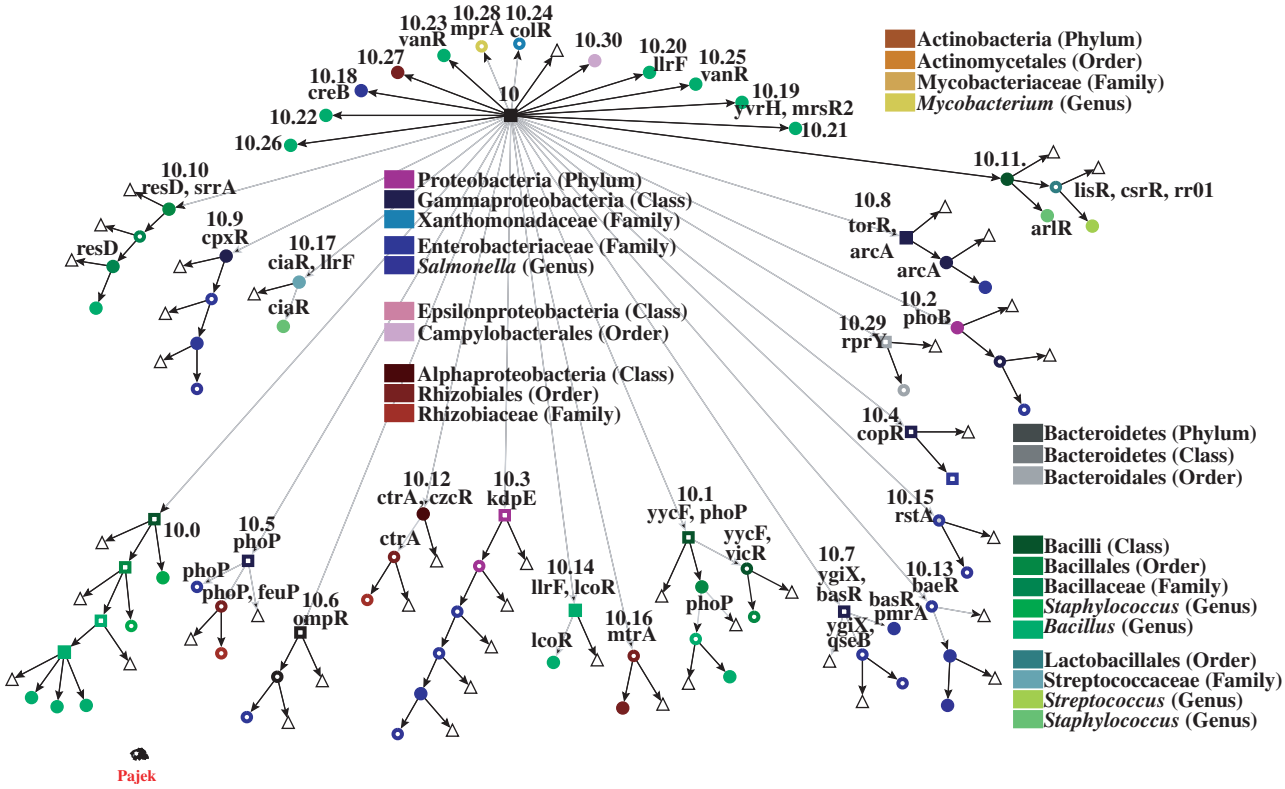

Figure 5. A hierarchical structure formed by HCG-10 and its descendant clusters, where most of the genes belonging to HCG-10 are annotated as *DNA-binding regulatory genes*. Each rectangular or circular node corresponds to a cluster, whereas a triangular node represents a group of genes that cannot be further clustered. The shape of a node reflects whether the cluster contains multiple genes from the same genome, with the rectangular standing for *yes* and the circular standing for *no*. The color of a node reflects the taxonomic lineages of the genomes being covered by the cluster, where a solid color represents that all the genomes being covered belong to the same taxonomic lineage for which the color stands, and a color with white interior represents that most (but not all) of the genomes being covered belong to the taxonomic lineage for which the color stands. The annotations accompanying the clusters are summarized from the NCBI annotations of the genes being included.

B. cereus ZK, and *B. thuringiensis* serovar *konkukian* str. 97-27, see Table S-5.2 in the Supplementary Data for the genes of these genomes and their clustering results) each have three genes included in HCG-10.0. This indicates that HCG-10.0 contains genes that may have resulted from gene duplication. At the child level of HCG-10.0, the 16 genes from the eight *staphylococcus* genomes are grouped into different clusters. Specifically, eight genes, each of which belongs to a distinct *staphylococcus* genome, are grouped into the cluster HCG-10.0.1; and the other eight genes, each of which also belongs to a distinct *staphylococcus* genome, are grouped together with 38 other HCG-10.0 genes into the cluster HCG-10.0.0. It suggests that the eight HCG-10.0.1 genes may be more equivalent to each other than they are to the eight HCG-10.0.0 genes. From an evolutionary point of view, it suggests that the gene duplication event that has led to the two DNA-binding response regulator genes in each of these eight *staphylococcus* genomes may have occurred before the speciation of the eight *staphylococcus* genomes from each other. We have made similar observations about HCG-10.0.0 and its descendent clusters. Particularly, at the child level of HCG-10.0.0, the eight genes of the eight *staphylococcus* genomes are grouped together with a gene of *Geobacillus kaustophilus* HTA426 into the cluster HCG-10.0.0.1; and at the leaf level of HCG-10.0.0, the 24 genes from the eight *bacillus* genomes are grouped into three different clusters, HCG-10.0.0.0.0, HCG-10.0.0.0.0.1 and HCG-10.0.0.0.0.2, respectively, with each cluster having only one gene from each and every of these eight *bacillus* genomes. It indicates that the gene duplication event that has led to three DNA-binding response regulator genes in each of these eight *bacillus* genomes may have occurred before the speciation of these eight *bacillus* genomes from each other but after the speciation of these eight bacillus genomes from the eight *staphylococcus* genomes.

HCG-424: Ribonucleotide reductases

It has been documented that there are three major classes of ribonucleotide reductases (42). Class I, encoded by the nrdA–nrdB or nrdE–nrdF genes, is dependent on oxygen and requires the substrate to be a ribonucleotide diphosphate. Additionally, the ribonucleotide reductases of class I can be further divided into two sub-classes, where one sub-class (Ia) is encoded by the nrdA–nrdB genes, and the other sub-class (Ib) is encoded by the nrdE–nrdF genes. These two sub-classes are different in their

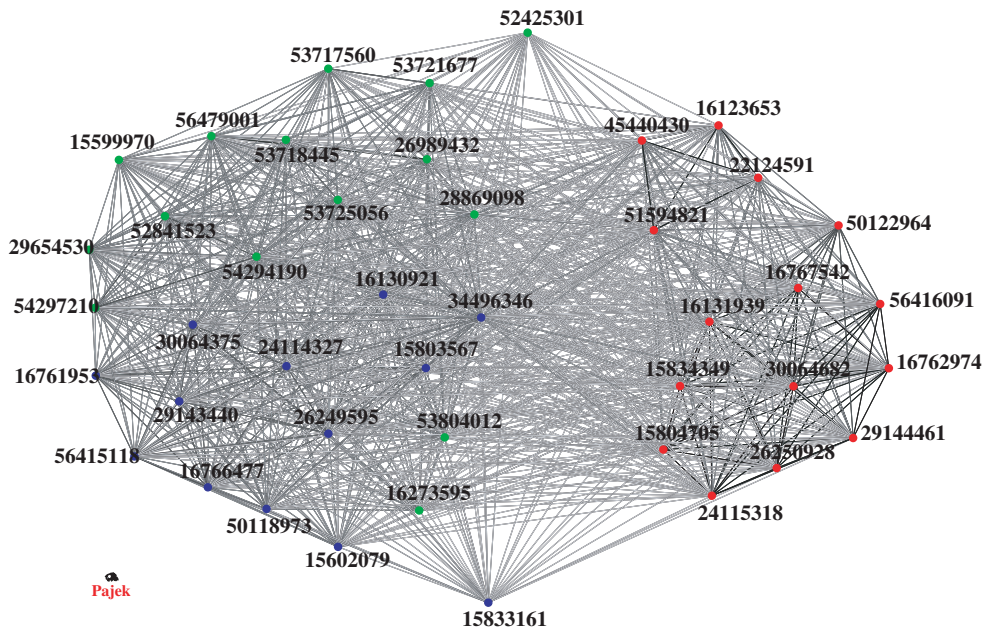

Figure 6. Genes of HCG-10.7 and their functional equivalence relationships (as measured by their $f(\cdot,\cdot)$ values), as represented by nodes and edges, respectively. The layout of the nodes and edges is generated using the Pajek Software, where both the Euclidean distance between two genes and the darkness of their connecting edge are roughly proportional to their $f(\cdot,\cdot)$ value. That is, the larger their $f(\cdot,\cdot)$ value is, the closer their two nodes are located, and the darker their connecting edge is. The red-colored genes, most of which are annotated as *basR*/*pmrA*, and the blue-colored genes, most of which are annotated as *ygiX*/*qseB*, are grouped into two different child clusters of HCG-10.7; whereas, the green-colored genes are those that cannot be further grouped.

allosteric regulations, where the class Ia is inhibited by the binding of dATP to an allosteric activity site, but the class Ib is insensitive to the inhibitor. Class III, encoded by the nrdD gene, is oxygen sensitive and requires the substrate to be a ribonucleotide triphosphate. Class II, encoded by the nrdJ gene, is oxygen tolerant and its preference for diphosphates or triphosphates differs for different organisms.

One root-level cluster, HCG-424 (http://csbl.bmb.uga.edu/HCG/displaynodeclass.php?class_string=424.), contains 292 genes from 216 genomes. Among these genes, ~96% are annotated by the COG online system as COG0209 (*ribonucleotide reductase alpha subunit*); ~91% are annotated by NCBI as *ribonucleotide reductase*; ~82% have GO annotations and are mostly annotated as GO:0004748 (*ribonucleoside-diphosphate reductase activity*); and ~81% have Pfam annotations and are largely annotated as PF00317 (*ribonucleotide reductase, all-alpha domain*) and PF02867 (*ribonucleotide reductase, barrel domain*). All these indicate that HCG-424 contains ribonucleotide reductase genes of classes I and II; and that a gene can be inferred as a ribonucleoside-diphosphate reductase gene if it is predicted to belong to HCG-424.

It should be noted that genes of HCG-424 are further grouped into 23 clusters organized as a tree structure with seven levels, as shown in Figure 7. While the ribonucleotide reductase genes of classes I and II are indistinguishable by COG, GO or Pfam classification, HCG-424 contains such information in its child clusters. For example, the class II ribonucleotide reductase (nrdJ) genes are separated from class I genes, and are grouped into cluster HCG-424.3; and at the grandchild level of HCG-424, the class Ib (nrdE) genes are separated from the class Ia (nrdA) genes, and are grouped into cluster HCG-424.0.0. This indicates that the HCG system can differentiate these ribonucleotide reductase genes that are functionally similar but yet distinctive.

From the evolutionary point of view, that the nrdA genes are grouped into different clusters at different levels reveals more about the evolutionary trace of the nrdA genes than about functional differences among them. For example, one of the grandchild clusters of HCG-424, HCG-424.0.1, contains 38 nrdA genes and covers 36 genomes, most of which belong to the gammaproteobacteria class. The descendent clusters of HCG-424.0.1 each still only contain nrdA genes but cover a different range of genomes. In particular, at the grandchild level of HCG-424.0.1, the cluster HCG-424.0.1.0.0 covers 17 grammaproteobacterial genomes, most of which belong to the enterobacteriaceae family; and at the great-grandchild level of HCG-424.0.1, the cluster HCG-424.0.1.0.0.0 covers six genomes, most of which belong to the *Salmonella* genus. As we have observed, from the hierarchy consisting of the cluster HCG-424.0.1 and its descendent clusters as well as other hierarchies, it is generally true that a lower level cluster tends to cover a narrower range of genomes.

VALIDATION OF THE HCG CLASSIFICATION

While it is desirable to go through each of the 5339 multi-leveled and 15 770 single-leveled cluster trees and

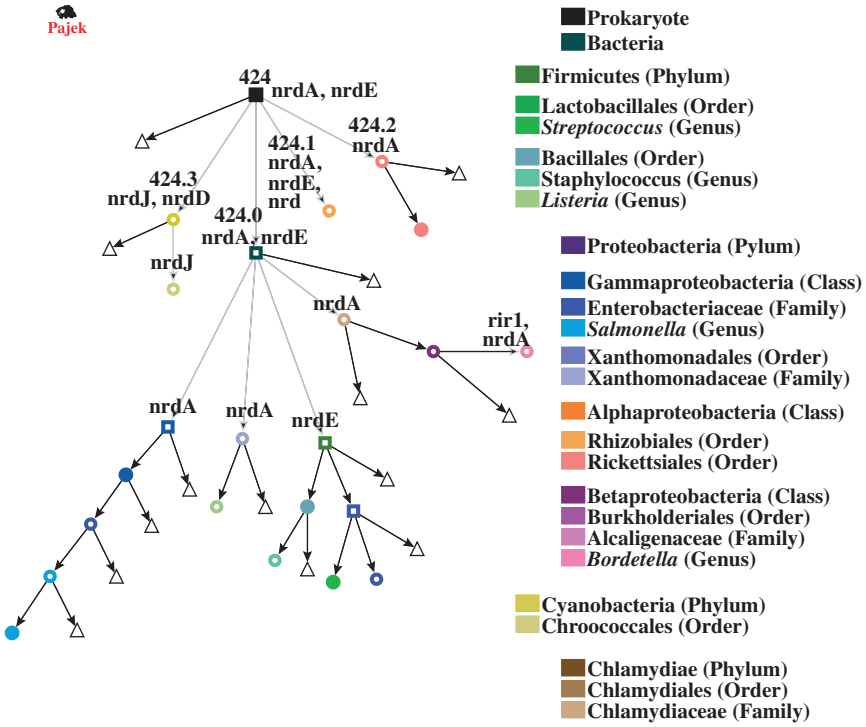

Figure 7. Tree structure formed by cluster HCG-424 and its descendant clusters, where all the 292 genes belonging to HCG-424 are annotated as *ribonucleotide reductase genes*. Each rectangular or circular node corresponds to a cluster, whereas a triangular node represents a group of genes that cannot be further clustered. The shape of a node reflects whether the cluster contains multiple genes from the same genome, with the rectangular standing for *yes* and the circular standing for *no*. The color of a node reflects the taxonomic lineages of the genomes being covered by the cluster, where a solid color represents that all the genomes being covered belong to the same taxonomic lineage for which the color stands, and a color with white interior represents that most (but not all) of the genomes being covered belong to the taxonomic lineage for which the color stands. The annotations accompanying the clusters are summarized from the NCBI annotations of the genes being included.

manually check their soundness as we have done on the two cluster trees in the Results section, it may not be practical for the time being. So here we present a computational validation on the whole clustering result, by comparing it with the taxonomic lineages of the 224 genomes, with the COG classification and with the Pfam classification, respectively.

Consistency between HCG classification and taxonomy of genomes

The taxonomy of prokaryotic genomes is established based on the classification of the ribosomal RNA genes and morphological/physiological characteristics of these genomes (22,23) (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/). Given two genomes, their relative positions on the taxonomic tree roughly reflect their evolutionary/morphological/physiological distances. We have defined a taxonomic distance between two genomes G_1 and G_2, $d_{\text{taxonomy}}(G_1, G_2)$, as the level of the most specific taxonomic lineage that is common to both genomes, and the taxonomic distance between two genes g_1 and g_2, $d_{\text{taxonomy}}(g_1, g_2)$, as the taxonomic distance between the two pertinent genomes.

The distribution of $d_{\text{taxonomy}}(g_1, g_2)$ for the whole gene population in HCG, denoted as the background distribution $d_{\text{taxonomy}}(g_1, g_2)$, is shown in Figure 8. We can see from the figure that the background distribution of $d_{\text{taxonomy}}(g_1, g_2)$ is peaked at the level of *super-kingdom*, indicating that two genes randomly picked from the whole gene population are most likely to simultaneously belong

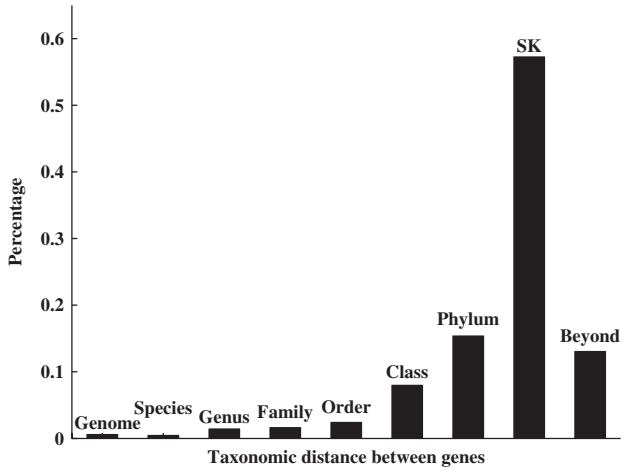

Figure 8. The distribution of $d_{\text{taxonomy}}(g_1, g_2)$ for all the genes being covered by the HCG prediction, where *SK* stands for super-kingdom, and *beyond* means that two genes do not even belong to the same super-kingdom.

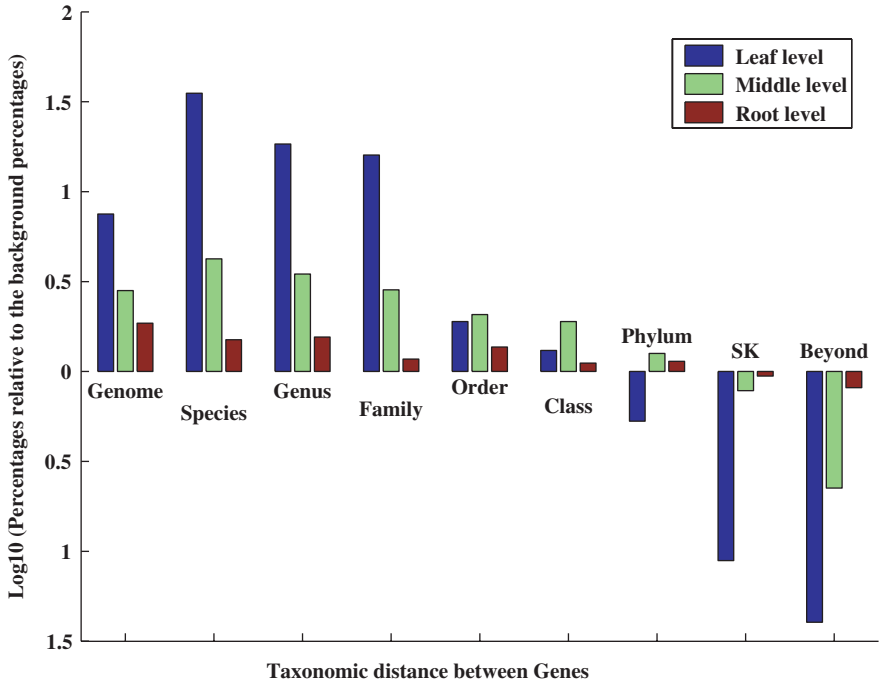

Figure 9. The distribution of $d_{\text{taxonomy}}(g_1, g_2)$ at the root, middle and leaf levels of the HCG, relative to the background distribution of $d_{\text{taxonomy}}(g_1, g_2)$, where SK stands for super-kingdom, and *beyond* means that two genes do not even belong to the same super-kingdom. Each bin represents the ratio between the percentage of the gene pairs at a particular HCG level and the percentage of the background gene pairs that have the same taxonomic distance level.

to the same *super-kingdom* but not to any more specific taxonomic lineage.

To better understand how the distribution of $d_{\text{taxonomy}}(g_1, g_2)$ for genes of the same cluster varies along different levels of the HCG classification, we have selected from the whole HCG hierarchy those paths, each of which starts from a leaf-level cluster and ends at a root-level cluster with at least one level in between, and have then taken from these paths the HCG clusters at the root, leaf and middle (which is equally distant to the root and the leaf) levels, respectively. The distributions of $d_{\text{taxonomy}}(g_1, g_2)$ at these different levels are compared with the background distribution of $d_{\text{taxonomy}}(g_1, g_2)$ and are shown in Figure 9, from which we have observed the following:

- The distribution of $d_{\text{taxonomy}}(g_1, g_2)$ at the leaf level is substantially different from the background distribution $d_{\text{taxonomy}}(g_1, g_2)$. This is reflected by the fact that the former distribution is much more dominant than the latter at the taxonomic lineage levels of *genome*, *family*, *genus* and *species*, but is much less dominant at the levels of *super-kingdom* and *beyond*. Compared to a pair of genes randomly chosen from the whole gene population, a pair of genes randomly chosen from a leaf-level cluster is much more likely to belong to the same of *genome*, *family*, *genus* or *species*, and is much less likely to belong to different *phyla* or *super-kingdoms*. This indicates that the taxonomic lineages of genomes covered by a leaf-level cluster tend to be much less diverse than the taxonomic lineages of all the genomes considered.
- The distribution of $d_{\text{taxonomy}}(g_1, g_2)$ at the root level is only slightly different than the background distribution $d_{\text{taxonomy}}(g_1, g_2)$, indicating that the taxonomic diversities of the root-level clusters tend to be as great as the taxonomic diversity of all the genomes covered by HCG.

These observations suggest that from the root level down to the leaf level, a cluster tends to include genes that are increasingly less taxonomically diverse, and the HCG classification is to some extent consistent with the taxonomy of prokaryotic genomes. It should be noted, however, that due to the prevalence of horizontal gene transfers in prokaryotes (43), the hierarchical classification established for some genes may not necessarily agree with the taxonomy; and from this point of view the HCG classification provides much more information than the taxonomy about the evolutionary trace of genes.

HCG versus COG classification

By using the COGNITOR system (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html) (2), we have obtained 3179 different COGs covering 459 955 (~70%) of the 658 174 genes from the 224 genomes. We have observed that (1) about 10% of the COGs each have fewer than 16 genes (therefore called *small* COGs); (2) about 12% of the genes being covered by COG each are assigned with more than one COGs (therefore called *complex* genes); and (3) about 20% of the COGs each have more than 20% of its covered genes being complex (therefore called *complex* COGs). Overall, there are 2272 COGs that

are neither small nor complex. We compare the HCG classification only against these neither-small-nor-complex COGs.

It should be noted that the COG and HCG classifications have different structures. In the COG system, all COGs are parallel to each other, some of which may have overlaps but none of which is contained inside another. In HCG, clusters are organized in a hierarchical way, in which clusters may be either *parallel-to* or *part-of* each other. Therefore, comparing the consistency between these two classification schemes requires a new comparison method, which we give below.

Consistency between COG and HCG classifications. MD measures. To quantify the consistency between these two different classification schemes, we define the *matching degree* (*MD*) between a COG, *COG*, and a HCG cluster, *HCG*, as

$$MD(COG, HCG) \equiv \frac{|COG \cap HCG|}{|COG \cup HCG|} = \frac{|COG \cap HCG|}{|COG| + |HCG| - |COG \cap HCG|} \quad (4)$$

Note that $|\mathrm{COG} \cap \mathrm{HCG}|/|\mathrm{COG}$ and $|\mathrm{COG} \cap \mathrm{HCG}|/|\mathrm{HCG}$ can be interpreted as the *sensitivity* and *specificity* measures of *HCG* with respect to *COG*, respectively. Hence the *MD* measure is basically a non-linear combination of the sensitivity and specificity measures (44,45). The *MD* measure increases monotonically as either the sensitivity or the specificity or both increases. Particularly, when $MD(COG, HCG) \geq 2/3$, both the sensitivity and specificity are guaranteed to be no smaller than 2/3. Therefore, we have used $MD_0 \equiv 2/3$ as a threshold to indicate strong consistency between *COG* and *HCG*.

HMD measures. We also define for each *COG* the *highest matching degree* (*HMD*) that can be achieved by considering all clusters in HCG as

$$HMD(COG) = \max_{HCG} MD(COG, HCG) \quad (5)$$

This measure reflects the consistency between the COG and HCG classifications from each COG's perspective. Without loss of generality, we consider that the HCG classification achieves strong consistency with a given *COG* if $HMD(COG) \geq MD_0 \equiv 2/3$.

HMD_r measures. We have observed cases where the union of multiple HCG clusters is more consistent with a *COG* than any individual HCG cluster. Hence we revise Equation (5) as follows to allow for the union of multiple HCG clusters being considered when assessing the consistency that the entire HCG system can achieve for a given *COG*.

$$HMD_r(COG) = \max_{HCG_j} MD\left(COG, \bigcup HCG_j\right) \quad (6)$$

Note that HMD_r is achieved by maximizing over all possible combinations of HCG clusters.

For the 2272 neither-small-nor-complex COGs, 1938 (85.3%) have their *HMD* values higher than the threshold MD_0; and 2053 (90.4%) have their HMD_r values higher than MD_0. This comparison indicates that the vast majority of the neither-small-nor-complex *COG*s are essentially included in the HCG classification either as a single HCG cluster or as the union of a few HCG clusters. Hence the information conveyed by the COG system (e.g. the clusterability of genes, functional difference between different COGs) is essentially conveyable by HCG.

Functional equivalence of genes at different HCG levels. Since the COG classification is considered one of the best functional classifications for genes, and genes of the same COG are generally considered to have similar functions, we have used the COG assignments of genes to analyze how the functional diversity within a HCG cluster varies from the root down to the leaf levels of the hierarchy. More specifically, we have used the following *a*verage *e*ntropy (AE_{COG}) of the COGs covered by each HCG cluster as a measure of functional diversity of the cluster, i.e. the larger the AE_{COG} of a HCG cluster is, the more functionally diverse it is.

$$AE_{COG}(HCG) = \frac{1}{K_{HCG}} \sum_{k=1}^{K_{HCG}} -[p_k \log p_k + (1-p_k)\log(1-p_k)] \quad (7)$$

HCG denotes a HCG cluster, K_{HCG} denotes the number of different *COG*s being covered by *HCG* and p_k ($k = 1, \ldots, K_{HCG}$) is the percentage of the *HCG* genes annotated as the kth *COG*. For the kth ($k = 1, \ldots, K_{HCG}$) *COG* being covered by *HCG*, the term $-[p_k \log p_k + (1-p_k)\log(1-p_k)]$ measures its entropy, which is a monotonically increasing function of p_k when p_k is in the range (0, ½) and a monotonically decreasing function of p_k when p_k is in the range (½, 1). In particular, when only half of the *HCG* genes are annotated as the kth *COG* but the other half are not (i.e. $p_k = ½$), the entropy of this *COG* reaches its maximum value of 0.693; and if all the *HCG* genes are annotated as the kth COG (i.e. $p_k = 1$), then the entropy of this kth *COG* reaches its minimum value of 0. Note that AE_{COG} captures more information about the functional diversity of a HCG cluster, in terms of the COG annotation of its included genes, than simply counting the number of COGs covered by this HCG cluster.

We have selected from the HCG hierarchy those paths each of which starts from a leaf cluster and ends at a root cluster with at least one level in between, have taken from these paths the clusters at the root, leaf and middle levels, and calculated AE_{COG} for each selected HCG cluster. Table 1 summarizes some statistics of AE_{COG} at different levels of the HCG hierarchy. We can see from the table that the mean values of AE_{COG} are very different at different levels, which is monotonically decreasing from the root level down to the leaf level. In particular, only for 40.78% of the root-level HCG clusters, genes of the same cluster have identical COG annotations (so that AE_{COG} of the corresponding HCG cluster is zero); and this percentage increases

Table 1. Statistics of the AE_{COG} measures at different levels of the HCG hierarchy

	Minimum	Maximum	Mean	Standard deviation	Percentage of the HCG clusters whose $AE_{COG}=0$ (%)
Root level	0	0.6931	0.1179	0.1690	40.78
Middle level	0	0.6931	0.0776	0.1463	68.61
Leaf level	0	0.6931	0.0340	0.1064	88.97
Pfam	0	0.3466	0.0151	0.0563	49.10

Table 2. Statistics of the AE_{Pfam} measures at different levels of the HCG hierarchy and for COG

	Minimum	Maximum	Mean	Standard deviation	Percentage of the clusters with $AE_{Pfam}=0$ (%)
Root level	0	0.6931	0.0911	0.1338	47.51
Middle level	0	0.6931	0.0486	0.1201	79.27
Leaf level	0	0.6931	0.0162	0.0801	95.06
COG	0	0.3466	0.0040	0.0190	42.99

to 68.61% for the middle-level HCG clusters, and to 88.97% for the leaf-level clusters. This indicates that from the root level down to the leaf level, the functional diversity of a HCG cluster tends to be increasingly lower, so the functional commonality shared by genes of the same cluster is increasingly more specific.

Consistency between HCG and Pfam classifications

Pfam is a database of multiple sequence alignments of protein domains or conserved protein regions, where each domain or conserved region is considered to be associated with some biological function. Pfam consists of two sets of families, in which *Pfam-A* families are derived based on curated sequence alignments, whereas *Pfam-B* families are obtained through an automated clustering procedure to supplement Pfam-A families, whose classification quality may not be as good as that of *Pfam-A*. The most recent version of Pfam (http://www.sanger.ac.uk/Software/Pfam/, 05/2006) provides annotations for 448 871 (~68%) of the 658 174 genes of the 224 genomes, covering 56 031 different Pfam families. In particular, 402 358 of these annotated genes have Pfam-A annotations, spanning 4510 Pfam-A families. We compare the HCG classification only to these Pfam-A families.

The Pfam classification is structured in a similar fashion to that of COG. So we compare the HCG and Pfam classifications using the same scheme described in the section Consistency between COG and HCG classifications, to measure their consistency from the Pfam's point of view, i.e. to replace *COG* in Equations (4)–(6) with a Pfam family. Among the 4510 Pfam-A families covered, 3268 (72.5%) have their *HMD* values higher than the threshold MD_0; and 3716 (82.4%) have their HMD_r higher than MD_0. This indicates that the majority of these Pfam-A families are essentially included in the HCG classification either as single or unions of a few HCG clusters. It should be noted that the discrepancy between the *HMD* and HMD_r values is mainly resulted from that a Pfam family represents a domain used in different proteins and therefore may be better matched by the union of multiple HCG clusters than by individual HCG clusters. For example, the GTP-binding domain is used in the protein synthesis initialization factor IF-2, elongation factors EF-Tu and EF-G, release factor RF-3 (46), the lepA protein (47) and the selenocysteine-specific elongation factor selB (48); and we have observed that the PF00009 family (*elongation factor Tu GTP-binding domain*) is better matched by the union of four HCG clusters, namely, HCG-210 (including tufB/COG0050, cysN/COG2895 and selB/COG3276 genes), HCG-226 (including fusA/COG0480 and prfC/COG0480 genes), HCG-294 (including lepA/COG0481 and typA/COG1217 genes), and HCG-611 (including infB/COG0532 genes), than by any individual HCG clusters.

Note that proteins belonging to the same Pfam family are considered to share a certain level of functional commonality. We have used the Pfam annotation of genes to check how the functional diversity of each HCG cluster varies at different levels in the HCG hierarchy. We have defined the average entropy (AE_{Pfam}) of the Pfam families covered by each HCG cluster, which is same as Equation (7) except that the annotations are for Pfam families rather than for COGs, and have used AE_{Pfam} to assess the functional diversity of a HCG cluster. Similar to AE_{COG}, AE_{Pfam} captures more about the diversity of the Pfam annotation of genes in a HCG cluster than the number of Pfam annotations covered by the HCG cluster. Table 2 summarizes some statistics of AE_{Pfam} at different levels in the HCG hierarchy. We can see from the table that the mean values of AE_{Pfam} are very different at different levels in HCG, which is monotonically decreasing from the root level down to the leaf level. In particular, for only 47.51% of the root-level HCG clusters, all the genes in the same cluster have identical Pfam annotations (so that AE_{Pfam} of the corresponding HCG cluster is zero); and this number increases to 79.27% for the middle-level HCG clusters, and to 95.06% for the leaf-level clusters. This indicates that from the root level down to the leaf levels, the functional diversity of each HCG cluster becomes increasingly smaller.

Comparison between COG and Pfam classifications

To better understand the consistency measures between the HCG and the two existing classifications, we have also compared the COG and Pfam classifications using the same framework as in Equations (4)–(6). For the same 2272 neither-small-nor-complex COGs, when being compared to the Pfam classification, 1270 (55.90%) have their *HMD* values higher than the threshold MD_0; and 1339 (58.93%) have their HMD_r values higher than the threshold MD_0. For the same 4510 Pfam-A families, when compared to the COG classification, only 3391 Pfam-A families have overlaps with COG clusters, among which 1517 (44.74% = 1517/3391) have their *HMD* values higher

then the threshold MD_0; and 1748 (51.55%) have their HMD_r values higher than the threshold MD_0. By comparing these consistency measures between the COG and Pfam with the consistency measures between the HCG and the two existing classifications (see Sections HCG versus COG classification and Consistency between HCG and Pfam classifications, respectively), it suggests that the HCG is more consistent with these two existing classifications than they are with each other.

To better understand the functional diversities of HCG clusters in the context of COG and Pfam, we have also computed the AE_{COG} values for each Pfam family and AE_{Pfam} for each COG cluster. Among the 2272 neither-small-nor-complex COGs, only for 1061 (46.70%) of them genes belonging to the same COG cluster have identical Pfam annotations. Among the 3391 Pfam-A families that have overlaps with COG clusters, only for 1665 (49.10%) of them genes belonging to the same Pfam-A family have identical COG annotations. By comparing these functional diversity measures of the COG and Pfam with the functional diversity measures of the HCG (as summarized in Tables 1 and 2), it suggests that from the middle level down along the HCG hierarchy, HCG clusters in general have lower the functional diversity than both COG and Pfam.

In summary, all the above comparisons indicate, on the one hand, that our HCG classification results are generally consistent with these two well-known classification systems, as reflected by the fact most of the COG and Pfam clusters are essentially included in the HCG classification. On the other hand, the comparison results also reveal that the functional diversity of a cluster varies along the different levels of the HCG hierarchy, indicating that the HCG classification can be used to predict biological functions of uncharacterized genes at different levels.

DISCUSSION

Most of the existing functional classification schemes of genes, such as COG and Pfam, employ a one-level classification strategy to group genes into parallel groups without clearly defining cross-group relationships. As we have established in this article, such classification schemes, as popular as they are, may not provide the most useful way for gene classification for the purpose of functional assignment of genes, particularly so when used for high-resolution functional assignment of genes. To address this issue, we have developed a novel way for classifying genes into a hierarchical structure of clusters, each consisting of functionally equivalent genes at some resolution level while finer clusters always consisting of functionally equivalent genes at higher resolution than the more coarse ones. This new classification scheme allows us to go beyond the concept of orthologous versus paralogous genes when making functional predictions, providing a richer and more useful framework for examining genes in terms of their functions.

In this classification scheme, we have assessed the functional equivalence relationships among genes based on both sequence similarity and the genomic context information, and have captured the hierarchical nature of equivalence relationships of genes. By comparing our classification results with the results of other classification schemes, we conclude that while our classification scheme in general captures the key features of the existing classification schemes such as COG and Pfam, it provides a much richer organization of genes, which facilitates functional assignment of unknown genes possibly at a higher resolution level.

A database along with a highly effective search engine of our classification results on 244 prokaryotic genomes has been set up, which is ready for application by users through the internet (http://csbl.bmb.uga.edu/HCG). The details of the database and the search engine are given in a companion article (Mao *et al.*, submitted for publication), to be published elsewhere.

The world-wide genome sequencing efforts have produced ~350 complete prokaryotic genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, as of April 2006) and this number is expected to extend beyond a few thousand within the next few years. To keep pace with this rapidly growing pool of complete genomes, we plan in the future to (1) extend the HCG database to include all sequenced prokaryotic genomes, and (2) to make our clustering algorithm and the HCG classification scheme more general. One particular extension we plan to have is to allow multi-membership of genes in parallel clusters, which we have discovered useful to deal with genes with multiple functions.

As the techniques and efforts for genome sequencing have advanced far ahead of those for experimental investigations, the gap between the pool of the completely sequenced genes and the pool of the experimentally studied genes is expected to continue to widen. Our classification scheme, as we have demonstrated here, will prove to be very useful for inferring biological functions of genes at a high-resolution level and to reveal hints on the evolutionary trace of genes/genomes, and will prove to be a powerful tool for filling the gap between the uncharacterized and the experimentally investigated pools of genes.

SUPPLEMENTARY DATA

Supplementary data are available at NAR online.

ACKNOWLEDGEMENTS

This work was supported in part by the US Department of Energy's Genomes to Life Program under project 'Carbon Sequestration in *Synechococcus* sp.: From Molecular Machines to Hierarchical Modeling' (http://www.genomes-to-life.org). The work is also supported, in part, by National Science Foundation (#NSF/DBI-0354771, #NSF/ITR-IIS-0407204, DBI-0542119), and also by Distinguished Scholar grant from the Georgia Cancer Coalition. Funding to pay for the open access publication

charges for this article was provided by the University of Georgia.

REFERENCES

1. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
2. 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. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
3. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
4. Dufayard,J.F., Duret,L., Penel,S., Gouy,M., Rechenmann,F. and Perriere,G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**, 2596–2603.
5. Storm,C.E. and Sonnhammer,E.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, **18**, 92–99.
6. Zmasek,C.M. and Eddy,S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Xu,Y., Olman,V. and Xu,D. (2002) Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning tree. *Bioinformatics*, **18**, 526–535.
9. Olman,V., Xu,D. and Xu,Y. (2003) CUBIC: identification of regulatory binding sites through data clustering. *J. Bioinform. Comput. Biol.*, **1**, 21–40.
10. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
11. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A*, **96**, 2896–2901.
12. Huynen,M., Snel,B., Lathe,W.3rd and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.
13. Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S. and Koonin,E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.*, **11**, 356–372.
14. Kolesov,G., Mewes,H.W. and Frishman,D. (2001) SNAPping up functionally related genes based on context information: a colinearity-free approach. *J. Mol. Biol.*, **311**, 639–656.
15. Notebaart,R.A., Huynen,M.A., Teusink,B., Siezen,R.J. and Snel,B. (2005) Correlation between sequence conservation and the genomic context after gene duplication. *Nucleic Acids Res.*, **33**, 6164–6171.
16. Mao,F., Su,Z., Olman,V., Dam,P., Liu,Z. and Xu,Y. (2006) Mapping of orthologous genes in the context of biological pathways: an application of integer programming. *Proc. Natl. Acad. Sci. U.S.A*, **103**, 129–134.
17. Wu,H., Mao,F., Olman,V. and Xu,Y. (2005) Accurate prediction of orthologous gene groups in microbes. *Proc. IEEE. Comput. Syst. Bioinform. Conf.*, 73–79.
18. van Dongen,S. (2000) Graph clustering by flow simulation. *PhD Thesis*. University of Utrecht.
19. Cormen,T.H. (2001) *Introduction to Algorithms*. 2nd edn. MIT Press, Cambridge, Mass.
20. Prim,R.C. (1957) Shortest connection networks and some generalizations. *Bell Sys. Technol. J.*, **36**, 1389–1401.
21. Wilks,S.S. (1962) *Mathematical Statistics*. John Wiley & Sons, New York.
22. Balows,A. (1992) *The Prokaryotes: A Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications*. Springer, New York.
23. Boone,D.R., Castenholz,R.W. and Garrity,G.M. (2001) *Bergey's Manual of Systematic Bacteriology*. 2nd edn. Springer, New York.
24. Wanner,B.L. (1993) Gene regulation by phosphate in enteric bacteria. *J. Cell Biochem.*, **51**, 47–54.
25. Epstein,W. (2003) The roles and regulation of potassium in bacteria. *Prog. Nucleic. Acid Res. Mol. Biol.*, **75**, 293–320.
26. Perron,K., Caille,O., Rossier,C., Van Delden,C., Dumas,J.L. and Kohler,T. (2004) CzcR-CzcS, a two-component system involved in heavy metal and carbapenem resistance in Pseudomonas aeruginosa. *J. Biol. Chem.*, **279**, 8761–8768.
27. Basim,H., Minsavage,G.V., Stall,R.E., Wang,J.F., Shanker,S. and Jones,J.B. (2005) Characterization of a unique chromosomal copper resistance gene cluster from Xanthomonas campestris pv. Vesicatoria. *Appl. Environ. Microbiol.*, **71**, 8284–8291.
28. Munson,G.P., Lam,D.L., Outten,F.W. and O'Halloran,T.V. (2000) Identification of a copper-responsive two-component system on the chromosome of Escherichia coli K-12. *J. Bacteriol.*, **182**, 5864–5871.
29. Yamamoto,K. and Ishihama,A. (2005) Transcriptional response of Escherichia coli to external copper. *Mol. Microbiol.*, **56**, 215–227.
30. Yeoman,K.H., Delgado,M.J., Wexler,M., Downie,J.A. and Johnston,A.W. (1997) High affinity iron acquisition in Rhizobium leguminosarum requires the cycHJKL operon and the feuPQ gene products, which belong to the family of two-component transcriptional regulators. *Microbiology*, **143**(Pt 1), 127–134.
31. Minagawa,S., Ogasawara,H., Kato,A., Yamamoto,K., Eguchi,Y., Oshima,T., Mori,H., Ishihama,A. and Utsumi,R. (2003) Identification and molecular characterization of the Mg^{2+} stimulon of Escherichia coli. *J. Bacteriol.*, **185**, 3696–3702.
32. Mizuno,T. and Mizushima,S. (1990) Signal transduction and gene regulation through the phosphorylation of two regulatory components: the molecular basis for the osmotic regulation of the porin genes. *Mol. Microbiol.*, **4**, 1077–1082.
33. Bordi,C., Ansaldi,M., Gon,S., Jourlin-Castelli,C., Iobbi-Nivol,C. and Mejean,V. (2004) Genes regulated by TorR, the trimethylamine oxide response regulator of Shewanella oneidensis. *J. Bacteriol.*, **186**, 4502–4509.
34. Iuchi,S. and Weiner,L. (1996) Cellular and molecular physiology of Escherichia coli in the adaptation to aerobic environments. *J. Biochem. (Tokyo)*, **120**, 1055–1063.
35. Ruiz,N. and Silhavy,T.J. (2005) Sensing external stress: watchdogs of the Escherichia coli cell envelope. *Curr. Opin. Microbiol.*, **8**, 122–126.
36. Quon,K.C., Marczynski,G.T. and Shapiro,L. (1996) Cell cycle control by an essential bacterial two-component signal transduction protein. *Cell*, **84**, 83–93.
37. Yamamoto,K., Hirao,K., Oshima,T., Aiba,H., Utsumi,R. and Ishihama,A. (2005) Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. *J. Biol. Chem.*, **280**, 1448–1456.
38. Horak,R., Ilves,H., Pruunsild,P., Kuljus,M. and Kivisaar,M. (2004) The ColR-ColS two-component signal transduction system is involved in regulation of Tn4652 transposition in Pseudomonas putida under starvation conditions. *Mol. Microbiol.*, **54**, 795–807.
39. Nagakubo,S., Nishino,K., Hirata,T. and Yamaguchi,A. (2002) The putative response regulator BaeR stimulates multidrug resistance of Escherichia coli via a novel multidrug exporter system, MdtABC. *J. Bacteriol*, **184**, 4161–4167.
40. Sperandio,V., Torres,A.G. and Kaper,J.B. (2002) Quorum sensing Escherichia coli regulators B and C (QseBC): a novel two-component regulatory system involved in the regulation of flagella and motility by quorum sensing in E. coli. *Mol. Microbiol.*, **43**, 809–821.
41. Kato,A. and Groisman,E.A. (2004) Connecting two-component regulatory systems by a protein that protects a response regulator from dephosphorylation by its cognate sensor. *Genes Dev.*, **18**, 2302–2313.
42. Larsson,K.M., Jordan,A., Eliasson,R., Reichard,P., Logan,D.T. and Nordlund,P. (2004) Structural mechanism of allosteric

substrate specificity regulation in a ribonucleotide reductase. *Nat. Struct. Mol. Biol.*, **11**, 1142–1149.
43. Gogarten,J.P. and Townsend,J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.*, **3**, 679–687.
44. Che,D., Li,G., Mao,F., Wu,H. and Xu,Y. (2006) Detecting uber-operons in microbial genomes. *Nucleic Acids Res*, **34**, 2418–2427.
45. Wu,H., Su,Z., Mao,F., Olman,V. and Xu,Y. (2005) Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Res.*, **33**, 2822–2837.
46. Voet,D. and Voet,J.G. (1990) *Biochemistry*. Wiley, New York.
47. March,P.E. and Inouye,M. (1985) GTP-binding membrane protein of Escherichia coli with sequence homology to initiation factor 2 and elongation factors Tu and G. *Proc. Natl. Acad. Sci. U.S.A*, **82**, 7500–7504.
48. Forchhammer,K., Leinfelder,W. and Bock,A. (1989) Identification of a novel translation factor necessary for the incorporation of selenocysteine into protein. *Nature*, **342**, 453–456.
49. Mao,F., Wu,H., Olman,V. and Xu,Y. (2006) HCG: A database for hierarchical classification of functional equivalent genes in prokaryotes. *Submitted.*