2822–2837 *Nucleic Acids Research, 2005, Vol. 33, No. 9*
doi:10.1093/nar/gki573

Prediction of functional modules based on comparative genome analysis and Gene Ontology application

Hongwei Wu[1,2], Zhengchang Su[1,2], Fenglou Mao[1], Victor Olman[1] and Ying Xu[1,2,*]

[1]Department of Biochemistry and Molecular Biology, University of Georgia, 120 Green Street, Athens, GA 30602-7229, USA and [2]Computational Biology Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Received January 28, 2005; Revised April 5, 2005; Accepted April 25, 2005

ABSTRACT

We present a computational method for the prediction of functional modules encoded in microbial genomes. In this work, we have also developed a formal measure to quantify the degree of consistency between the predicted and the known modules, and have carried out statistical significance analysis of consistency measures. We first evaluate the functional relationship between two genes from three different perspectives—phylogenetic profile analysis, gene neighborhood analysis and Gene Ontology assignments. We then combine the three different sources of information in the framework of Bayesian inference, and we use the combined information to measure the strength of gene functional relationship. Finally, we apply a threshold-based method to predict functional modules. By applying this method to *Escherichia coli* K12, we have predicted 185 functional modules. Our predictions are highly consistent with the previously known functional modules in *E.coli*. The application results have demonstrated that our approach is highly promising for the prediction of functional modules encoded in a microbial genome.

INTRODUCTION

The worldwide sequencing efforts of microbial genomes (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html, http://www.sanger.ac.uk/Projects/Microbes, http://www.tigr.org/tdb/mdb/mdbcomplete.html and http://microbialgenome.org) have led to the completion of over 200 microbial genomes, and this number will continue to increase very rapidly. We expect to see over a thousand sequenced genomes within the next few years. This wealth of genomic data provides unprecedented opportunities for computational biologists to unveil the enormous amount of information encoded in the genomes about the biological machinery of these micro-organisms.

The complex biological processes in a living microbial cell, including metabolism, regulations and signal transduction, are carried out by a large set of functional modules at various levels. These functional modules are made up of interacting biomolecules and serve as the basic building blocks of the complex biological machinery in a microbial cell. Some of the functional modules are organized in a hierarchical manner while others could serve in multiple levels forming a complex organizational network. At the very basic level in the functional hierarchy is the set of operons (1,2) (we also consider single-gene operons here), each of which are arranged in tandem in the genome and share a common promoter and a common terminator. A regulon is a group of operons that are regulated by a common transcriptional regulator (1,2), and a modulon is a group of regulons that are controlled by more global regulators and respond to more general physiological states (1,2). At the top of this functional hierarchy is a set of stimulons, each of which consists of a collection of operons, regulons and/or modulons that respond to a common environmental stimulus (1,2). In general, functional modules at different levels are made up of combinations of operons, regulons, modulons and possibly stimulons, many of which might have 'conserved' components and structures across multiple (related) microbial organisms. We expect that the interactions of these functional modules play the essential roles in the entire functionality of a microbial organism (3,4).

In this paper, we present a new computational framework for the prediction of functional modules in a microbial organism through comparative genome analysis and application of Gene Ontology (GO) (5) information. Our focus will be on the identification of genes involved in a functional module rather than the detailed interaction relationships among these genes. This study provides a basis for further prediction of detailed

*To whom correspondence should be addressed. Tel: +1 706 542 9779; Fax: +1 706 542 9751; Email: xyn@bmb.uga.edu

gene functions and prediction of biological (metabolic, signaling and regulatory) networks.

Comparative genome analysis is one of the most powerful tools to unravel the information encoded in a genome (6,7). By identifying conserved elements across multiple genomes, researchers have been able to uncover biological functions and structures at different levels of the biological machinery in a microbial cell. Successful applications of using such a strategy include predictions of gene functions (8–10), *cis*-regulatory elements (11,12), operons (13,14) and regulons (11,15).

Our approach to the identification of functional modules is based on three classes of information: (i) co-evolutionary information of genes that are encoded in phylogenetic profiles (8); (ii) conserved gene neighborhood information, which reflects if genes have 'conserved' adjacency relationship across multiple microbial genomes and, hence, possibly suggest their functional relatedness; and (iii) functional relatedness information of genes that are encoded in the GO classification (5,16). GO is a dynamically controlled vocabulary that can be applied to all organisms and has been used to measure protein or gene relationships (17–20). In this paper, we define a similarity measure among GO terms to evaluate the functional relationship of genes.

Each of the three measures provides a different perspective about functional relationships among genes. Information derived through each of them is then combined using a Bayesian inference framework. Using this combined score, we predict whether two genes belong to the same functional module. We use a graph representation to describe such a functional relatedness relationship. That is, if two genes are predicted to belong to the same functional module, they will have an edge linking their representative nodes in this graph representation. We believe that a functional module, in general, should be represented by a group of genes that are highly connected in this graph representation. In addition, each 'highly connected' subgraph may be part of a larger and also highly connected subgraph representing for functional modules at different levels in the biological machinery of a microbial cell. We have applied this computational procedure to the genome of *Escherichia coli* K12. On the 2579 genes of *E.coli* that have been assigned biological process GO terms, we have predicted a large interaction network involving these genes. Then using a particular segmentation strategy on this network, we have obtained 185 highly connected modules covering 654 genes. By comparing these predicted modules to the known pathways, regulons and operons in the Eco Cyc (21) and KEGG (22) databases, we have observed that the highest matching degrees (see definition later) achieved by these predicted modules are significantly higher than those achieved by randomly generated modules. We believe that this large interaction network contains a great amount of information about functional modules and relationships among them in *E.coli*. The predicted modules presented in this paper represent probably only a small subset of the functional modules in *E.coli*. In our future work, we intend to fully investigate the functional modules and their organizational structures by further refining the large network prediction and exploring more sophisticated strategies to identify many more such functional modules.

There have been numerous efforts in the past few years, devoted to the discovery of molecular modules through computational methods, as exemplified by the previous works (19,20,23–25). Lee *et al*. (19) compared different classes of data (including the physical and genetic interaction datasets, mRNA co-expression data, functional links extracted through literature search, prediction of gene fusion events and phylogenetic profile analysis) and integrate them by using a Bayesian framework, and Lee *et al*. have applied this capability for the prediction of functional relatedness of genes in *Saccharomyces cerevisiae*. von Mering *et al*. (20) first developed quantitative ways to measure functional relationship among genes from three different sources of information (including gene fusion, chromosomal proximity and phylogenetic profiles) and predicted functional modules by using a clustering algorithm. Spirin and Mirny (23) developed algorithms to analyze the structural properties of a predicted interaction network to identify the subsets of genes that are densely connected among themselves but sparsely connected with others. Yamada *et al*. (24) extracted gene modules from metabolic pathways by identifying genes that share similar phylogenetic profiles. Yanai and Delisi (25) predict gene links by using the same sources of information as described previously (20) and use the union operation to combine the three types of links to predict gene modules. The gene modules predicted by all these studies have shown some level of consistency with the well-established biological concepts as described in COGG (26), Eco Cyc (21) and KEGG (22).

Our approach differs significantly from the previous methods, as summarized below: (i) we utilize both the phylogenetic and the neighborhood profiles obtained from the comparative genome analysis; (ii) we explicitly incorporate the GO information into our evaluation of functional relationships of genes; (iii) we combine different sources of information in the framework of the Bayesian inference; and (iv) we develop a formal measure to quantify the degree of consistency between the predicted and the known modules, and we provide analysis of statistical significance for such comparisons.

MATERIALS AND METHODS

We first evaluate the functional relationships among genes from three different perspectives: one based on GO assignments and the other two based on comparative genome analysis. We, then, combine these different measures by using a Bayesian inference to predict functional modules.

Gene Ontology

The GO Consortium (5) has developed three separate ontologies—molecular function, biological process and cellular component—to describe the attributes of gene products, where molecular function defines what a gene product does at the biochemical level without specifying where or when the event actually occurs or its broader context; biological process describes the contribution of a gene product to a biological objective; and cellular component refers to where in the cell a gene product functions. Each GO is structured as a directed acyclic graph, wherein each term is a child of one or multiple parents, and child terms are instances or components of parent terms. For example, in Figure 1, the term carbohydrate biosynthesis (GO: 0016051) is an instance of the term carbohydrate metabolism (GO: 0005975) as well as an instance of the term macromolecule biosynthesis (GO: 0009059).

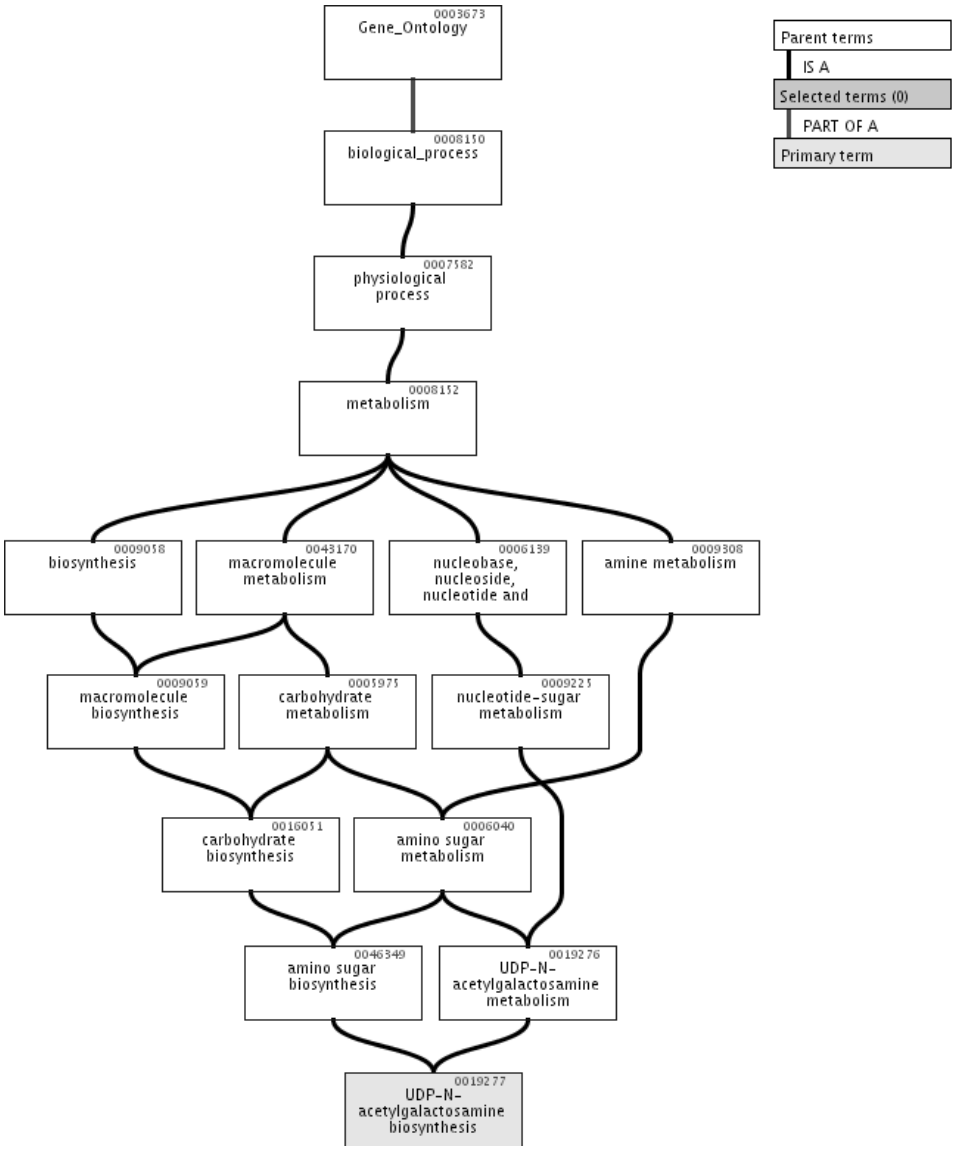


Figure 1. The directed acyclic graph induced from the GO term UDP-N-acetylgalactosamine biosynthesis (GO: 0019277), wherein at the bottommost level is the GO term of interest itself, and at the upper levels are all its ancestors, adapted from QuickGO Go Browser (http://www.ebi.ac.uk/ego/).

From another point of view, from each GO term V, we can induce a directed acyclic graph that has the following properties:

(i) The bottommost level of the graph is V itself, and at upper levels are its ancestor GO terms. Particularly, at the topmost level is the term Gene_Ontology (GO: 0003673).
(ii) Given two GO terms V_1 and V_2, if V_1 is one of the ancestors of V_2; then, the graph induced from V_1 is completely included as a subgraph in the graph induced from V_2.

Figure 1 shows the graph induced from the term UDP-N-acetylgalactosamine biosynthesis (GO: 0019277). Note that the graph induced from a GO term can also be represented by a collection of paths with each path corresponding to a complete trace from the bottommost level (i.e. the GO term of interest itself) to the topmost level (i.e. the term Gene_Ontology). For example, one possible path in the graph of Figure 1 is as follows:

UDP-N-acetylgalactosamine biosynthesis (GO: 0019277) → amino sugar biosynthesis (GO: 0046349) → carbohydrate biosynthesis (GO: 0016051) → macromolecule biosynthesis (GO: 0009059) → biosynthesis (GO: 0009058) → metabolism (GO: 0008152) → physiological process (GO: 0007582) → biological_process (GO: 0008150) → Gene_Ontology (GO: 0003673).

The number of terms along the longest path in a graph is called the depth of the graph. For example, the depth of the graph for Figure 1 is 9. The depth of a graph reflects how specific the GO term is in describing the attributes of gene products. Hence, a GO term is always more specific than its ancestor GO terms.

In the rest of this paper, we do not differentiate between a GO term and the graph induced from it. When quantifying the similarity between two GO terms, it is desired that both their commonality and individual specificities (in describing the attributes of gene products) can be captured simultaneously. Let V_s and V_t be the graphs induced from two GO terms, respectively. We define their similarity $s(V_s, V_t)$ as follows:

$$S \equiv \max_{L_s \in V_s, L_t \in V_t} \left\{ \begin{array}{l} \text{the number of common} \\ \text{terms between } L_s \text{ and } L_t \end{array} \right\}, \quad \mathbf{1}$$

where L_s and L_t are the paths of V_s and V_t, respectively. If $s(V_s, V_t)$ is large, it means that the two GO terms are both highly specific and share much commonality in describing the attributes of gene products; if $S_{GO}(V_s, V_t)$ is small, it means that either the two GO terms are not highly specific or they do not share much commonality; and, if $s(V_s, V_t) > s(V_u, V_v)$, it means that the most recent common ancestor of V_s and V_t is more specific than the most recent common ancestor of V_u and V_v.

The above defined similarity measure is then used to assess the functional relationship among genes (through their products) based on their biological process GO assignments. Because a gene product may be involved in more than one biological process, it may be assigned with multiple GO terms, and, therefore, multiple GO graphs may be induced. Let $\mathbf{V}(g)$ denote all the GO terms assigned to a gene g. We define the GO similarity S_{GO} for a pair of genes g_i and g_j as the maximum similarity of all possible combinations of $\mathbf{V}(g_i)$ and $\mathbf{V}(g_j)$, i.e.

$$S_{GO}(g_i, g_j) \equiv \max_{V_i \in \mathbf{V}(g_i), \mathbf{V}_j \in \mathbf{V}(g_j)} s(V_i, V_j), \quad \mathbf{2}$$

where V_i and V_j are the GO terms assigned to g_i and g_j, respectively. If $S_{GO}(g_i, g_j)$ is large, then at a very specific level the two genes are involved in at least one common biological process (e.g. both of them are involved in the UDP-N-acetylgalactosamine biosynthesis); and if $S_{GO}(g_i, g_t)$ is small, then only at a very general level can the two genes be considered to be involved in the same biological process (e.g. both of them are involved in the physiological process).

Our similarity measure for GO annotations is very similar in concept to the information-content based semantic similarity defined by Lord *et al*. (27), although the two definitions treat in different ways how specific the most recent common ancestor of two GO terms is. In our definition, the specificity of a GO term is reflected by the number of GO terms along the longest path (distance) to the topmost level GO term, whereas in (27) the specificity is reflected by the number of genes that are assigned with it or its descendant GO terms. When assessing the functional relationship among genes using the similarity measure of GO terms, we take a different approach from that of Lord *et al*. (27). Given all GO terms assigned to two genes, we use the maximum similarity of all term pairs, whereas Lord *et al*. use the average similarity of all term pairs. We believe that the difference between the implementations of ours and Lord *et al*.'s is minor compared with their commonality at the conceptual level.

As we focus on the prediction of functional modules, which are basic building blocks of the biological machinery of a microbial cell for carrying out complex biological processes, we are most interested in whether a specific gene is involved in related biological processes and have consequently only used the biological process GO annotations of genes. Out of the 4311 genes in *E.coli* K12 (release of December 2003), 2579 genes have been assigned biological process GO terms by the GO Annotation project (16) (release of September 2004). In this paper we focus on these 2579 genes. These genes form 3 324 331 gene pairs, among which 46 009 pairs whose two genes belong to the same functional module (i.e. an operon, regulon or pathway) according to Eco Cyc (21) are considered to form the positive set, and the remaining 3 278 322 pairs are considered to form the background (or called random) set. Note that a random set is not necessarily an equivalent to a negative set, where the former consists of the pairs whose two genes have not been confirmed by experiments to be involved in the same functional module, and the latter consists of the pairs whose two genes have been confirmed by experiments not to be involved in the same functional module. Since our knowledge about the role of a protein/gene in biological processes is still accumulating and evolving, it might be difficult to identify a true negative set. Hence, we use a random set rather than a negative set.

The means and standard deviations of $S_{GO}(g_i, g_j)$ for both the positive and the random sets are summarized in Table 1. We have performed a χ^2-test (28) to check if the distribution of $S_{GO}(g_i, g_j)$ is different for the positive and the random sets. The χ^2-statistics (four bins have been used for the χ^2-test, so that there are ~32, 41, 15 and 27% of positive pairs, and 37, 37, 16 and 10% random pairs falling into the four bins, respectively) is corresponding to a P-value less than 10^{-4}, which reveals that the distribution of $S_{GO}(g_i, g_j)$ for the positive set is significantly different from the random set. Figure 2 shows the distribution of $S_{GO}(g_i, g_j)$ for the positive and the random sets. From this figure, we can see that a pair of genes of the same functional module (operon, regulon or pathway) are more likely to have a high degree of GO similarity than a random pair.

Comparative genome analysis

Among the 145 complete genome sequences of bacteria and archaea that are available (release of December 2003), we have used 135 genomes for our comparative genome analysis, including *E.coli* K12 as the target genome and the other 134 as the reference genomes. Let G_0, N_0 and g_i denote the genome, the number of genes and the i-th ($i = 1, \ldots, N_0$) gene of *E.coli* K12, and G_k and N_k denote the genome and the number of genes of the k-th ($k = 1, \ldots, K$, $K = 134$ in our study) reference genome, respectively.

The first step in our comparative genome analysis is to predict orthologous genes for each gene of *E.coli* K12 in the reference genomes. We have used the PSI-BLAST (29) with an E-value of 10^{-6} to search for the bi-directional best

Table 1. Means and standard deviations of $S_{GO}(g_i,g_j)$, $d(g_i, g_j)$, $S_N(g_i, g_j)$ and $C_{ombined}(g_i, g_j)$ for the positive and the random sets

	$S_{GO}(g_i, g_j)$		$d(g_i, g_j)$		$S_N(g_i, g_j)$		$C_{ombined}(g_i, g_j)$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Positive set	3.652	1.871	23.273	11.365	0.864	0.436	0.286	1.192
Random set	3.111	1.244	26.882	16.077	0.720	0.266	−0.262	0.813

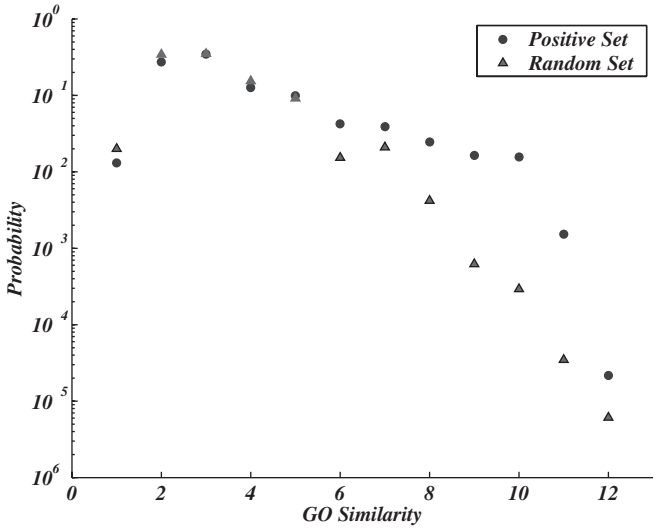


Figure 2. Distribution of $S_{GO}(g_i, g_j)$ for the positive (circles) and the random (triangles) sets.

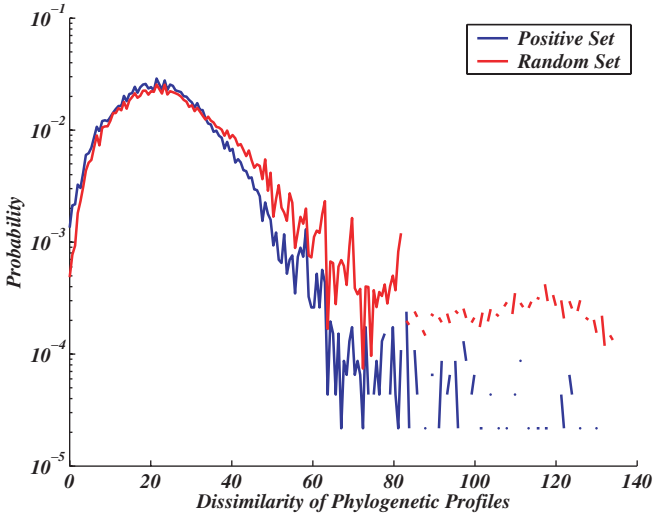


Figure 3. Distribution of $d(g_i, g_j)$ for the positive (blue) and the random (red) sets.

hits (BDBH) and have obtained the following two profiles for each gene of *E.coli* K12, g_i ($i = 1, \ldots, N_0$):

(i) The phylogenetic profile of g_i is a K-dimensional binary vector indicating the presence or absence of the orthologous genes of g_i in the reference genomes.
(ii) The neighborhood profile is a K-dimensional vector with each element indicating the absence or the order of the orthologous genes of g_i along the reference genomes.

These two profiles are then used to evaluate the functional relationships of genes.

Dissimilarity of phylogenetic profiles. Let $\mathbf{x}_i \equiv [x_{i1}, x_{i2}, \ldots, x_{ik}]^T$ denote the phylogenetic profile of gene g_i, with $x_{ik} = 1$ representing the presence and $x_{ik} = 0$ for the absence of g_i in G_k. Given the phylogenetic profiles $\mathbf{x}_i$ and $\mathbf{x}_j$ for a pair of genes g_i and g_j, we define their dissimilarity $d(g_i, g_j)$ as follows:

$$d(g_i, g_j) \equiv \frac{d_{\text{Hamming}}(\mathbf{x}_i, \mathbf{x}_j)}{1 + \gamma E_{\text{ntropy}}(\mathbf{x}_i, \mathbf{x}_j)}, \quad \mathbf{3}$$

where $d_{\text{Hamming}}(\mathbf{x}_i, \mathbf{x}_j)$ represents the Hamming distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, γ is a non-negative constant and has been set as 2 in our study, and $E_{\text{ntropy}}(\mathbf{x}_i, \mathbf{x}_j)$ is the entropy of the common part of $\mathbf{x}_i$ and $\mathbf{x}_j$ and is computed as follows:

$$E_{\text{ntropy}}(\mathbf{x}_i, \mathbf{x}_j) = -p \log p - (1-p) \log (1-p) \quad \mathbf{4}$$

with p being the frequency of 1's in the common part. The dissimilarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ ranges from 0 to K, with 0 and K corresponding to the identical and complementary phylogenetic profiles, respectively, and is smaller when $\mathbf{x}_i$ and $\mathbf{x}_j$ have a smaller Hamming distance and/or a more diverse common part. We omit further details about how to choose the value of the constant γ to balance between the Hamming distance and the entropy of $\mathbf{x}_i$ and $\mathbf{x}_j$.

To check if the distribution of $d(g_i, g_j)$ is different for the positive and the random sets, we have performed both Kolmogorov–Smirnov test [by treating $d(g_i, g_j)$ as an observation of a continuous random variable] and χ^2-tests (28). The Kolmogorov–Smirnov test rejects the hypothesis that the distribution of $d(g_i, g_j)$ is the same for the positive and the random sets with a significance level less than 10^{-4}, while from the χ^2-test (nine bins have been used for the χ^2-test, so that there are at least 6% of positive pairs and 5.58% of random pairs in each bin) we have obtained the χ^2-statistics value of 1341 corresponding to a P-value less than 10^{-4}. Therefore, both tests demonstrate that the distribution of $d(g_i, g_j)$ for the positive set is significantly different from the random set. Figure 3 shows the distributions of $d(g_i, g_j)$ for the positive and the random sets. Note that a pair of genes from the same functional module are less likely to have a large measure of dissimilarity than a random pair.

Likelihood of neighborhood profiles. Let $\mathbf{y}_i \equiv [y_{i1}, y_{i2}, \ldots, y_{ik}]^T$ denote the neighborhood profile of a gene g_i. When the k-th element of $\mathbf{y}_i$, y_{ik}, is 0, it stands for that the orthologous gene of g_i is absent from G_k; when y_{ik} is $n \neq 0$, it stands for that the orthologous gene of g_i is ordered as the n-th gene along G_k. We make the following assumptions about the statistical model for the neighborhood profiles:

(i) For each gene g_i and each reference genome G_k, g_i is present in G_k with probability p_{ik}; and when present, the order of g_i along G_k is uniformly distributed over $\{1, 2, \ldots, N_k\}$, i.e.

$$p(y_{ik} = n) = \begin{cases} 1 - p_{ik}, & \text{when } n = 0 \\ p_{ik}/N_k, & \text{when } n = 1, 2, \ldots, N_k. \end{cases}$$

(ii) For each gene g_i, the elements of $\mathbf{y}_i$ are independent, i.e.

$$P(y_{ik}, y_{il}) = P(y_{ik})P(y_{il}), \quad k, l \in \{1, 2, \ldots, K\} \text{ and } k \neq l.$$

This means that a gene's behavior (i.e. the presence/absence and order) is independent among different reference genomes.

Given a pair of genes g_i and g_j, under the hypothesis that g_i and g_j do not functionally relate to each other, their

neighborhood profiles $\mathbf{y}_i$ and $\mathbf{y}_j$ can be treated as independent random vectors; hence, the log-likelihood of $\mathbf{y}_i$ and $\mathbf{y}_j$, $L(g_i, g_j)$, is computed as follows:

$$L(g_i, g_j) = \sum_{k=1}^{K} L(g_i, g_j, G_k) \quad \mathbf{5}$$

with $L(g_i, g_j, G_k)$ standing for the log-likelihood of y_{ik} and y_{jk} and are being computed as follows:

$$\begin{aligned} L(g_i, g_j, G_k) = {} & I(y_{ik} = 0, y_{jk} = 0) \log P_{00} \\ & + I(y_{ik} = 0, y_{jk} \neq 0) \log P_{01} \\ & + I(y_{ik} \neq 0, y_{jk} = 0) \log P_{10} \\ & + I(y_{ik} \neq 0, y_{jk} \neq 0) \log P_{11}, \end{aligned} \quad \mathbf{6}$$

where $I(\cdot,\cdot)$ is 1 if and only if both criteria within the parentheses are satisfied and is 0 otherwise, P_{00} stands for the probability that neither g_i nor g_j is present, P_{01} stands for the probability that only g_j is present, P_{10} stands for the probability that only g_i is present, and P_{11} stands for the probability that both g_i and g_j are present and have a distance not more than $d_k(i, j)$ with $d_k(i, j) \equiv |y_{ik} - y_{jk}|$ being the observed distance between g_i and g_j (in terms of the number of genes in between) along G_k. Since y_{ik} and y_{jk} can be treated as independent random variables under the hypothesis that g_i and g_j are functionally unrelated, these four probabilities are computed as follows:

$$\begin{aligned} P_{00} &= (1-p_{ik})(1-p_{jk}). \\ P_{01} &= (1-p_{ik})p_{jk}. \\ P_{10} &= p_{ik}(1-p_{jk}). \\ P_{11} &= p_{ik}\,p_{jk}\,\frac{d_k(i,j)(2N_k - d_k(i,j) - 1)}{N_k(N_k - 1)}. \end{aligned}$$

Note that P_{11} is very small when $d_k(i, j)$, p_{ik} and/or p_{jk} are small. This is consistent with our intuition—it is very unlikely that two functionally unrelated genes are simultaneously present at a genome with a small distance, especially when these two genes are not highly conserved at this genome.

The likelihood $L(g_i, g_j)$ is the evidence supporting the hypothesis that the two genes do not functionally relate to each other. The larger $L(g_i, g_j)$ is, the more supportive the neighborhood profiles $\mathbf{y}_i$ and $\mathbf{y}_j$ are for this hypothesis; and the smaller $L(g_i, g_j)$ is, the more $\mathbf{y}_i$ and $\mathbf{y}_j$ are against this hypothesis (i.e. the more supportive $\mathbf{y}_i$ and $\mathbf{y}_j$ are for the alternative hypothesis that the two genes functionally relate to each other in some way).

We use the score

$$S_N(g_i, g_j) \equiv -L(g_i, g_j) \quad \mathbf{7}$$

to evaluate the strength of the functional relationship between g_i and g_j in terms of their neighborhood profiles. In this way, a larger $S_N(g_i, g_j)$ implies a stronger functional relationship between g_i and g_j.

In practice, p_{ik} is unknown and must be estimated from phylogenetic profiles. We first group all the 134 reference genomes into 14 groups so that each group corresponds to a phylum (as shown in Table 2), and then assume that p_{ik} is identical within the same group of genomes for each gene g_i. The maximum-likelihood estimation of p_{ik} is computed as the frequency of g_i in the group that G_k belongs to, i.e.

$$p_{ik} = \frac{\text{the number of genomes having } g_i \text{ in the group } G_k \text{ belongs to}}{\text{the number of genomes in the group } G_k \text{ belongs to}}. \quad \mathbf{8}$$

We have performed both the Kolmogorov–Smirnov test and the χ^2-test to check if the distribution of $S_N(g_i, g_j)$ is different for the positive and the random sets. The Kolmogorov–Smirnov test rejects the hypothesis that the distribution of $S_N(g_i, g_j)$ is the same for the positive and the random sets with significance level less than 10^{-4}; and from the χ^2-test (seven bins have been used for the χ^2-test, so that there are at least 8% of positive gene pairs and 5.7% of random gene pairs in each bin) we have obtained the χ^2-statistics value of 7583 corresponding to a P-value less than 10^{-4}. Therefore, both tests demonstrate that the distribution of $S_N(g_i, g_j)$ for the positive set is significantly different from the random set. Figure 4 shows the distribution of $S_N(g_i, g_j)$ for both the positive and the random sets. As shown in the figure, a pair of genes of the same functional module are more likely to have a large neighborhood score than a random pair.

Bayesian inference for information fusion

As we have shown, the GO similarity measure $S_{GO}(g_i, g_j)$, the dissimilarity measure of phylogenetic profiles $d(g_i, g_j)$, and the neighborhood score $S_N(g_i, g_j)$ reflect the possible functional relationship between two genes from different perspectives. To fully utilize the information from all the three sources, we combine them by using a Bayesian inference approach.

Given two genes g_i and g_j, we assume that their GO assignments are conditionally independent of their comparative genome analysis. Therefore, based on the three measures, $S_{GO}(g_i, g_j)$, $d(g_i, g_j)$ and $S_N(g_i, g_j)$, the odds of g_i and g_j belonging to the same functional module can be computed as follows:

$$\begin{aligned} & \frac{P(g_i \text{ and } g_j \in \text{the same module} \mid S_{GO}(g_i,g_j), d(g_i,g_j), S_N(g_i,g_j))}{P(g_i \text{ and } g_j \notin \text{the same module} \mid S_{GO}(g_i,g_j), d(g_i,g_j), S_N(g_i,g_j))} \\ & = \frac{P(S_{GO}(g_i,g_j) \mid g_i \text{ and } g_j \in \text{the same module})}{P(S_{GO}(g_i,g_j) \mid g_i \text{ and } g_j \notin \text{the same module})} \\ & \quad \times \frac{P(d(g_i,g_j), S_N(g_i,g_j) \mid g_i \text{ and } g_j \in \text{the same module})}{P(d(g_i,g_j), S_N(g_i,g_j) \mid g_i \text{ and } g_j \notin \text{the same module})} \\ & \quad \times \frac{P(g_i \text{ and } g_j \in \text{the same module})}{P(g_i \text{ and } g_j \notin \text{the same module})}. \end{aligned} \quad \mathbf{9}$$

We use the logarithm of (9), i.e.

$$\begin{aligned} & C_{\text{ombined}}(g_i, g_j) \\ & \equiv \log P(g_i \text{ and } g_j \in \text{the same module} \mid \\ & \qquad S_{GO}(g_i,g_j), d(g_i,g_j), S_N(g_i,g_j)) \\ & \quad - \log P(g_i \text{ and } g_j \notin \text{the same module} \mid \\ & \qquad S_{GO}(g_i,g_j), d(g_i,g_j), S_N(g_i, \pm g_j)) \end{aligned} \quad \mathbf{10}$$

as the combined score for g_i and g_j. The higher $C_{\text{ombined}}(g_i, g_j)$ is, the stronger the functional relationship between g_i and g_j we consider them to have.

Note that the ratios and conditional distributions in the last three rows of (9) must be known or estimated a priori in order to calculate $C_{\text{ombined}}(g_i, g_j)$ for any pair of genes. To estimate

Table 2. Group assignments of the 134 reference genomes

Phylum	Genomes
Crenarchaeota	*Aeropyrum pernix*, *Pyrobaculum aerophilum*, *Sulfolobus solfataricus*, *Sulfolobus tokodaii*
Aquificae	*Aquifex aeolicus*
Euryarchaeota	*Archaeoglobus fulgidus* DSM 4304, *Halobacterium* sp. NRC-1, *Methanococcus jannaschii*, *Methanopyrus kandleri* AV19, *Methanosarcina acetivorans* str. C2A, *Methanosarcina mazei* Goe1, *Methanothermobacter thermautotrophicus*, *Pyrococcus abyssi*, *Pyrococcus horikoshii*, *Pyrococcus furiosus* DSM 3638, *Thermoplasma acidophilum*, *Thermoplasma volcanium*
Firmicutes	*Bacillus anthracis* A2012, *Bacillus anthracis* str. *Ames*, *Bacillus cereus* ATCC 14579, *Bacillus halodurans*, *Bacillus subtilis*, *Clostridium acetobutylicum*, *Clostridium perfringens*, *Clostridium tetani* E88, *Enterococcus faecalis* V583, *Lactobacillus plantarum* WCFS1, *Lactococcus lactis* subsp. *lactis*, *Listeria innocua*, *Listeria monocytogenes* EGD-e, *Mycoplasma gallisepticum* R, *Mycoplasma genitalium*, *Mycoplasma penetrans*, *Mycoplasma pneumoniae*, *Mycoplasma pulmonis*, *Oceanobacillus iheyensis* HTE831, *Staphylococcus aureus* subsp. *aureus* MW2, *Staphylococcus aureus* subsp. *aureus* Mu50, *Staphylococcus aureus* subsp. *aureus* N315, *Staphylococcus epidermidis* ATCC 12228, *Streptococcus agalactiae* 2603V/R, *Streptococcus agalactiae* NEM316, *Streptococcus mutans* UA159, *Streptococcus pneumoniae* R6, *Streptococcus pneumoniae* TIGR4, *Streptococcus pyogenes*, *Streptococcus pyogenes* MGAS315, *Streptococcus pyogenes* MGAS8232, *Streptococcus pyogenes* SSI-1, *Thermoanaerobacter tengcongensis*, *Ureaplasma urealyticum*
Bacteroidetes	*Bacteroides thetaiotaomicron* VPI-5482, *Chlorobium tepidum* TLS, *Porphyromonas gingivalis* W83
Actinobacteria	*Bifidobacterium longum* NCC2705, *Corynebacterium diphtheriae*, *Corynebacterium efficiens* YS-314, *Corynebacterium glutamicum* ATCC 13032, *Mycobacterium bovis* subsp. *bovis* AF2122/97, *Mycobacterium leprae*, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium tuberculosis* H37Rv, *Streptomyces avermitilis* MA-4680, *Streptomyces coelicolor* A3(2), *Tropheryma whipplei* TW08/27, *Tropheryma whipplei* str. *Twist*
Spirochaetes	*Borrelia burgdorferi*, *Treponema pallidum*
Chlamydiae	*Chlamydia muridarum*, *Chlamydia trachomatis*, *Chlamydophila caviae* GPIC, *Chlamydophila pneumoniae* AR39, *Chlamydophila pneumoniae* CWL029, *Chlamydophila pneumoniae* J138, *Chlamydophila pneumoniae* TW-183
Fusobacteria	*Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586
Cyanobacteria	*Gloeobacter violaceus*, *Nostoc* sp. PCC 7120, *Prochlorococcus marinus* str. MIT 9313, *Prochlorococcus marinus* subsp. *marinus* str. CCMP1375, *Prochlorococcus marinus* subsp. *pastoris* str. CCMP1378, *Synechococcus* sp. WH 8102, *Synechocystis* sp. PCC 6803, *Thermosynechococcus elongatus* BP-1
Nanoarchaeota	*Nanoarchaeum equitans* Kin4-M
Planctomycetes	*Pirellula* sp.
Thermotogae	*Thermotoga maritima*
Proteobacteria	*Bordetella bronchiseptica*, *Bordetella parapertussis*, *Bordetella pertussis*, *Bradyrhizobium japonicum*, *Buchnera aphidicola* (*Baizongia pistaciae*), *Buchnera aphidicola* str. APS (*Acyrthosiphon pisum*), *Buchnera aphidicola* str. Sg (*Schizaphis graminum*), *Campylobacter jejuni*, *Candidatus Blochmannia floridanus*, *Caulobacter crescentus* CB15, *Chromobacterium violaceum* ATCC 12472, *Coxiella burnetii* RSA 493, *Escherichia coli* CFT073, *Escherichia coli* O157:H7, *Escherichia coli* O157:H7 EDL933, *Haemophilus ducreyi* 35000HP, *Haemophilus influenzae* Rd, *Helicobacter hepaticus* ATCC 51449, *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Mesorhizobium loti*, *Neisseria meningitidis* MC58, *Neisseria meningitidis* Z2491, *Nitrosomonas europaea* ATCC 19718, *Pasteurella multocida*, *Photorhabdus luminescens* subsp. *laumondii* TTO1, *Pseudomonas aeruginosa* PA01, *Pseudomonas putida* KT2440, *Pseudomonas syringae* pv. tomato str. DC3000, *Ralstonia solanacearum*, *Rickettsia conorii*, *Rickettsia prowazekii*, *Salmonella enterica* subsp. *enterica* serovar *Typhi*, *Salmonella enterica* subsp. *enterica* serovar *Typhi* Ty2, *Salmonella typhimurium* LT2, *Shewanella oneidensis* MR-1, *Shigella flexneri* 2a str. 2457T, *Shigella flexneri* 2a str. 301, *Sinorhizobium meliloti*, *Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis*, *Wolinella succinogenes*, *Xanthomonas axonopodis* pv. *citri* str. 306, *Xanthomonas campestris* pv. *campestris* str. ATCC 33913, *Xylella fastidiosa* 9a5c, *Xylella fastidiosa* Temecula1, *Yersinia pestis*, *Yersinia pestis* KIM

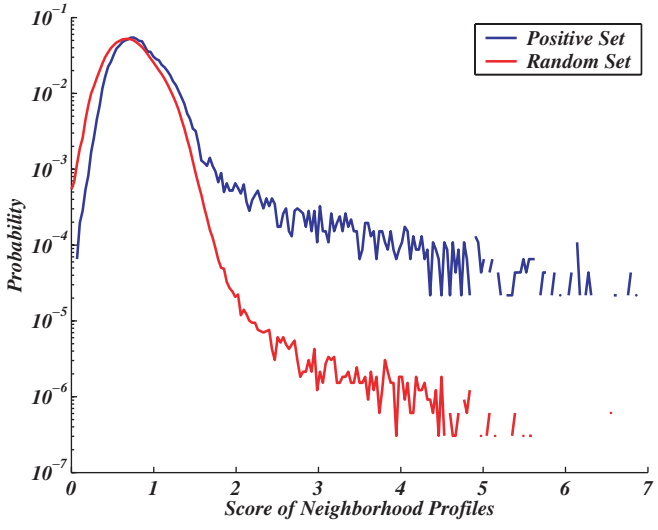


Figure 4. Distribution of $S_N(g_i, g_j)$ for the positive (blue) and the random (red) sets.

them, we consider a pair of genes in the positive set to be in the same module, and a pair in the random set not to be in the same module.

On estimating the joint conditional distributions of $\{d(g_i, g_j), S_N(g_i, g_j)\}$ of (9), we have taken two different approaches. One approach, called the naive Bayesian inference (30), assumes the conditional independence between $d(g_i, g_j)$ and $S_N(g_i, g_j)$, and estimates the conditional distributions of $d(g_i, g_j)$ and $S_N(g_i, g_j)$ separately. The second approach, called the Bayesian inference, directly estimates the joint conditional distribution of $\{d(g_i, g_j), S_N(g_i, g_j)\}$. Each approach has its own strengths and limitations. For example, the naive Bayesian inference approach heavily relies on the assumption of conditional independence; hence, the resulting estimated joint distribution may be far away from the true joint distribution when the assumption is not valid. Whereas, for the Bayesian inference approach, when estimating the distribution of random variables, we need much more observations for a multi-dimensional random vector than for a one-dimensional random variable (the number of needed observations grows exponentially with the dimensionality of the random vector)

to achieve the same level of resolution (31); hence, there may exist a resolution problem with the estimated joint distribution. Especially for the estimation of

$$P(d(g_i, g_j), S_N(g_i, g_j) \mid g_i \text{ and } g_j \in \text{the same module})$$

because there are only a very small percentage of gene pairs (46 009/3 324 331 $\approx$ 1.4%) whose two genes are known to belong to the same functional module based on the current information of Eco Cyc, the Bayesian inference approach cannot achieve a high-resolution level. It is an interesting problem regarding how to find the right tradeoff between these two approaches, but that is out of the scope of this paper.

We have computed $C_{\text{ombined}}(g_i, g_j)$ by using both the naive Bayesian and Bayesian inferences. We have performed the Kolmogorov–Smirnov test and the χ^2-test to check if the distribution of $C_{\text{ombined}}(g_i, g_j)$ is different for the positive and the negative tests. The Kolmogorov–Smirnov tests reject the hypothesis that the distribution of $C_{\text{ombined}}(g_i, g_j)$ is the same for the positive and the random sets with a significance level less than 10^{-4} for both approaches; and, from the χ^2-tests we have obtained the values of χ^2-statistics as 11 591 and 4866, both corresponding to the P-values less than 10^{-4}, for the naive Bayesian and Bayesian approaches, respectively. [The χ^2-tests have been performed on the normalized $C_{\text{ombined}}(g_i, g_j)$ for both approaches. Although seven bins have been used, bins have been located differently for the naive Bayesian and Bayesian approaches. The bins for the naive Bayesian approach are located at 0.35, 0.36, . . . , 0.41, so that there are at least 8% of positive gene pairs and 8.92% of random gene pairs in each bin. The bins for the Bayesian approach are located at 0.25, 0.26, . . . , 0.31, so that there are at least 5% positive gene pairs and 5.8% random gene pairs in each bin.] All these tests demonstrate that the distribution of $C_{\text{ombined}}(g_i, g_j)$ is different for the positive and the random sets, for both the naive Bayesian and the Bayesian approaches. To compare these two inference approaches, we have normalized $C_{\text{ombined}}(g_i, g_j)$ so that the normalized $C_{\text{ombined}}(g_i, g_j)$ for each approach ranges from 0 to 1. Figure 5 shows the distributions of the normalized $C_{\text{ombined}}(g_i, g_j)$ for both the positive and the random sets, and for both the naive Bayesian and the Bayesian inference approaches. As shown in the figure that (i) for both approaches, a pair of genes of the same functional module are more likely to have a large $C_{\text{ombined}}(g_i, g_j)$ than a random pair; and (ii) the naive Bayesian approach discriminates the positive pairs out of the random pairs more accurately than the Bayesian approach. So, in the rest of this paper, we focus on the naive Bayesian inference approach.

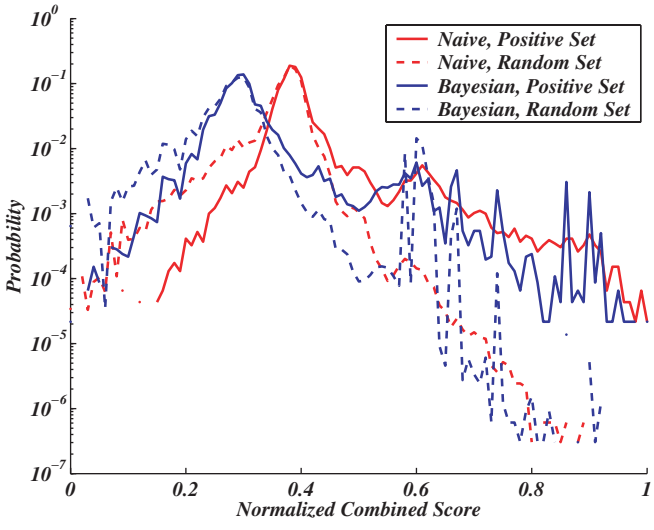


Figure 5. The normalized $C_{\text{ombined}}(g_i, g_j)$ of both naive Bayesian (red) and Bayesian (blue) inference approaches for both the positive (solid) and the random (dashed) sets.

Threshold-based module prediction and evaluation

Every two genes will have a score $C_{\text{ombined}}(g_i, g_j)$ measuring their functional relationship. The higher the score, the stronger their functional relationship is. Note that a negative value of $C_{\text{ombined}}(g_i, g_j)$ does not necessarily mean that the g_i and the g_j are less likely to belong to the same functional module, because the positive and the random sets we have used to estimate the ratios and conditional distributions in (9) are not complete. More specifically, the positive set we have used is only part of the true positive set, and the random set we have used contains some true positives as well as true negatives, where by true positive we mean a set consisting of all pairs whose two genes belong to the same module, and by true negative we mean a set consisting of all pairs whose two genes do not belong to the same module.

The genes and their functional connections can be interpreted at different levels. At the lowest resolution level, all genes are functionally connected to form one large network that is responsible for all activities of a cell; at a higher resolution level, genes with stronger functional relationship stand out and form smaller and densely interacted modules that are responsible for some specific activities of a cell. At the highest resolution level, each gene forms a functional module by itself.

To predict biologically meaningful functional modules of smaller sizes, we apply a simple thresholding method described as follows. We first compute for each gene g_i the mean (m_i) and standard deviation (σ_i) of its functional connection scores $C_{\text{ombined}}(g_i, g_j)$ with all other genes g_j ($j \neq i$), keep the connection between g_i and g_j if and only if

$$C_{\text{ombined}}(g_i, g_j) \geq m_i + \alpha\sigma_i \quad \text{and} \quad C_{\text{ombined}}(g_i, g_j) \geq m_j + \alpha\sigma_j$$

with $\alpha \geq 0$ being a threshold parameter, and call a group of genes that are directly or indirectly linked as a predicted functional module.

We choose the value of α to make our predicted modules consistent as much as possible with the known functional modules in Eco Cyc, where the consistency is measured by using the matching degrees defined below.

Matching degree between a pair of known and predicted gene modules. Let K_m be the set of genes in the m-th known functional module, and C_n be the set of genes in the n-th predicted module, the matching degree between K_m and C_n, t_{mn}, is defined as follows:

$$t_{mn} = \frac{|K_m \cap C_n|}{|K_m \cup C_n|}, \qquad \textbf{11}$$

where $|\cdot|$ represents for the cardinality of a set, and $\cap$ and $\cup$ represent the intersection and union operations between two sets.

As we demonstrate below, the matching degree t_{mn} defined in (11) is actually a combination of the measures of sensitivity and specificity of C_n, where sensitivity = $|K_m \cap C_n|/|K_m|$ and specificity = $|K_m \cap C_n|/|C_n|$.

$$\begin{aligned} t_{mn} &= \frac{|K_m \cap C_n|}{|K_m| + |C_n| - |K_m \cap C_n|} = \left(\frac{|K_m|}{|K_m \cap C_n|} + \frac{|C_n|}{|K_m \cap C_n|} - 1 \right)^{-1} \\ &= (\text{sensitivity}^{-1} + \text{specificity}^{-1} - 1)^{-1} \\ &\approx 2 \times \text{sensitivity} \times \text{specificity} \\ &\quad - (\text{sensitivity} + \text{specificity}) + 1. \end{aligned} \tag{12}$$

In the last row of (12) are the first- and second-order terms of the Taylor expansion [MathWorld—a Wolfram Web Resource, http://mathworld.wolfram.com/TaylorExpansion.html] of the original non-linear function of t_{mn} in (11) around sensitivity = 1 and specificity = 1. Note that one can add coefficients before sensitivity or specificity in (12) to put different weights on these two measures.

Highest matching degree for a known module. Let $\mathbf{C}$ be the collection consisting of all the N predicted modules $C_1, \ldots, C_N$, i.e. $\mathbf{C} \equiv \{C_1, C_2, \ldots, C_N\}$. Since t_{mn} is monotonically increasing as the sensitivity and/or specificity of C_n increases, by maximizing t_{mn} with respect to all the predicted modules in $\mathbf{C}$, we pick one particular predicted module for K_m that best balances the measures of sensitivity and specificity. We define the highest matching degree (HMD) for K_m, t_m which measures the matching capability of all the predicted modules of $\mathbf{C}$ to K_m, as follows:

$$t_m = \max_{1 \leq n \leq N} t_{mn} = \max_{1 \leq n \leq N} \left(\frac{|K_m \cap C_n|}{|K_m \cup C_n|} \right). \tag{13}$$

The so-defined HMD has the following properties:

(i) If the genes in $\mathbf{C}$ does not cover any gene of K_m, then the HMD achieved by $\mathbf{C}$ for K_m is 0.
(ii) If there exist several predicted modules in $\mathbf{C}$ each of which can be perfectly matched with a fraction of K_m, then the HMD achieved by $\mathbf{C}$ for K_m is the ratio of the maximum size of these predicted modules to the size of K_m.
(iii) If there exists one predicted module C_n one of whose fractions is perfectly matched with the entire K_m, then the HMD achieved by $\mathbf{C}$ for K_m is the ratio of the size of K_m to the size of C_n.
(iv) If there exists one predicted module C_n that is perfectly matched with the entire K_m, then the HMD achieved by $\mathbf{C}$ for K_m is 1.

Let $\mathbf{K}$ be the collection consisting of all the M known functional modules, i.e. $\mathbf{K} \equiv \{K_1, K_2, \ldots, K_M\}$. We choose the value of α so that the average HMD (AHMD) over $\mathbf{K}$ [defined in the following equation] is maximized.

$$\text{AHMD} \equiv \frac{1}{M} \sum_{m=1}^{M} t_m = \frac{1}{M} \sum_{m=1}^{M} \max_{1 \leq n \leq N} \left(\frac{|K_m \cap C_n|}{|K_m \cup C_n|} \right). \tag{14}$$

Statistical analysis on the highest matching degree. Let $t_1, t_2, \ldots, t_m$ be the HMDs for $\mathbf{K} \equiv \{K_1, K_2, \ldots, K_M\}$ achieved by our predicted modules $\mathbf{C} \equiv \{C_1, C_2, \ldots, C_N\}$. To evaluate the statistical significance of $\{t_1, t_2, \ldots, t_M\}$, we first estimate the probability distribution of the HMDs for the same $\mathbf{K}$ achieved by a collection of randomly predicted modules $\mathbf{C}' \equiv \{C'_1, C'_2, \ldots, C'_N\}$ where each C'_n is of the same size as C_n, and then estimate the Z-score of $\{t_1, t_2, \ldots, t_M\}$. If the Z-score is high, then the HMDs achieved by $\mathbf{C}$ are statistically significant. Here by randomly predicted modules we mean a module that is predicted by randomly picking out the genes with equal probability and without replacement from the pool of the N_0 genes of *E.coli* K12.

Given a known functional module K_m, because of the nature of randomness of the functional modules in $\mathbf{C}'$, the matching degree achieved by each C'_n in $\mathbf{C}'$ and the HMD achieved by $\mathbf{C}'$ are all random variables; hence, in the following analysis, we use T'_{mn} and T'_m to denote the matching degree and the HMD for K_m achieved by C'_n and $\mathbf{C}'$, respectively.

We first focus on the statistical model of T'_{mn}. The distribution of T'_{mn} can be approximated by using a Binomial distribution (32) when the number of genes in the known functional module, $|K_m|$, is small compared with N_0, i.e.

$$P(T'_{mn} = t) \approx \binom{|C_n|}{z_{mn}} p_m^{z_{mn}} (1 - p_m)^{(|C_n| - z_{mn})}, \tag{15}$$

where $p_m \equiv |K_m|/N_0$ and

$$z_{mn} \equiv \left[\frac{t}{1+t} (|K_m| + |C_n|) \right] \tag{16}$$

with $[\cdot]$ representing a round-off.

We then turn our attention to the statistical model of T'_m. Because the modules $C'_1, C'_2, \ldots, C'_N$ are disjoint, the distribution of T'_m can be approximated based on (15) as follows:

$$\begin{aligned} P(T'_m \leq t) &= P\left(\max_{1 \leq n \leq N} T'_{mn} \leq t \right) = P(T'_{m1} \leq t, T'_{m2} \leq t, \ldots, T'_{mN} \leq t) \\ &= \prod_{n=1}^{N} P(T'_{mn} \leq t) \approx \prod_{n=1}^{N} \sum_{u=0}^{z_{mn}} \binom{|C_n|}{u} p_m^u (1 - p_m)^{(|C_n| - u)} \end{aligned} \tag{17}$$

with z_{mn} being given as in (16).

Let μ_m and s_m be the mean and standard deviation of T'_m, associated with K_m, then $(T'_m - \mu_m)/s_m$ is called the standardized HMD of K_m. By using the Central Limit Theorem (32), the sum of the standardized HMDs over all $\mathbf{K} \equiv \{K_1, K_2, \ldots, K_M\}$ asymptotically complies to a normal distribution, i.e.

$$\frac{1}{\sqrt{M}} \sum_{m=1}^{M} \frac{T'_m - \mu_m}{s_m} \sim \aleph(0, 1). \tag{18}$$

The Z-score of $\{t_1, t_2, \ldots, t_M\}$, which are the HMDs for $\mathbf{K} \equiv \{K_1, K_2, \ldots, K_M\}$ achieved by our predicted modules $\mathbf{C} \equiv \{C_1, C_2, \ldots, C_N\}$, is then computed as follows:

$$Z_{\text{score}} = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \frac{t_m - \mu_m}{s_m}. \tag{19}$$

A high Z-score means that $\{t_1, t_2, \ldots, t_M\}$ and, consequently, our predicted modules in $\mathbf{C}$ as well as our prediction method, are statistically significant.

Table 3. The maximum AHMDs and their associated α-values for the 10 experiments, each of which corresponds to one repeat of the procedure of forming the training set, computing the combined score and predicting modules

Experiment	Pathway		Regulon		Operon	
	α	AHMD	α	AHMD	α	AHMD
1	6.75	0.265	6.5	0.168	6.5	0.164
2	6.25	0.257	5.25	0.171	5.25	0.165
3	7	0.259	5.75	0.192	5.5	0.172
4	6.25	0.260	5.25	0.184	5	0.157
5	6	0.269	5.5	0.194	6	0.181
6	6.25	0.268	5.75	0.183	6	0.171
7	5.25	0.249	5.25	0.190	4.75	0.171
8	6.25	0.288	6	0.217	6	0.188
9	5.75	0.261	4.75	0.191	4.75	0.165
10	6	0.267	5.25	0.200	5.25	0.176

Table 4. The maximum AHMD values, the associated values of α, the number (N) of predicted modules, the total number (|**C**|) of genes in all the predicted modules and the associated Z-scores, for the known pathways, regulons and operons achieved by using different sources of information

	AHMD	α	N	|**C**|	Z-score
Pathways (M = 207)					
$C_{\text{ombined}}(g_i, g_j)$	0.265	6.75	185	654	62.293
$S_{\text{N}}(g_i, g_j)$	0.236	3.75	189	998	58.474
$d(g_i, g_j)$	0.0364	3	28	221	4.753
$S_{\text{GO}}(g_i, g_j)$	0.224	4	106	796	70.103
Regulons (M = 132)					
$C_{\text{ombined}}(g_i, g_j)$	0.168	6.5	194	717	37.908
$S_{\text{N}}(g_i, g_j)$	0.182	3.5	189	1099	40.576
$d(g_i, g_j)$	0.0200	2.75	26	431	0.769
$S_{\text{GO}}(g_i, g_j)$	0.117	3.75	115	959	31.591
Operons (M = 745)					
$C_{\text{ombined}}(g_i, g_j)$	0.164	6.5	194	717	32.572
$S_{\text{N}}(g_i, g_j)$	0.176	5	188	702	39.406
$d(g_i, g_j)$	0.0147	3	28	221	0.502
$S_{\text{GO}}(g_i, g_j)$	0.0708	3.75	115	959	13.868

The phylogenetic and neighborhood profiles are obtained by using the BDBH method.

EXPERIMENTS AND RESULTS

Performance by using the combined score $C_{\text{ombined}}(g_i, g_j)$

We have performed the following procedure to predict functional modules based on the combined score $C_{\text{ombined}}(g_i, g_j)$:

(i) Randomly choose 1250 genes (≈50%) out of the gene pool (consisting of 2579 *E.coli* K12 genes) to form the training set, and use the pairs of these training genes to estimate the ratios and conditional distributions in (9).
(ii) Compute $C_{\text{ombined}}(g_i, g_j)$ for all pairs (including the pairs of the training genes).
(iii) Segment the large interaction network with various α values to predict functional modules that maximize the AHMD [defined in (14)].

Owing to the nature of randomness in choosing genes to form the training set in the first step, the outputs of the following

Table 5. The maximum AHMD values, the associated values of α, the number (N) of predicted modules, the total number (|**C**|) of genes in all the predicted modules and the associated **Z**-scores, for the known pathways, regulons and operons achieved by using the combined information, neighborhood profiles and phylogenetic profiles, respectively

	AHMD	α	N	|**C**|	Z-score
Pathways (M = 207)					
$C_{\text{ombined}}(g_i, g_j)$	0.248	7.25	191	700	66.694
$S_{\text{N}}(g_i, g_j)$	0.212	3.5	173	920	52.832
$d(g_i, g_j)$	0.0416	3.25	61	416	2.647
Regulons (M = 132)					
$C_{\text{ombined}}(g_i, g_j)$	0.176	6	171	1006	45.653
$S_{\text{N}}(g_i, g_j)$	0.170	3.25	165	1008	37.033
$d(g_i, g_j)$	0.0317	3.25	61	416	−1.406
Operons (M = 745)					
$C_{\text{ombined}}(g_i, g_j)$	0.164	7.25	191	700	32.959
$S_{\text{N}}(g_i, g_j)$	0.157	4.5	173	669	36.274
$d(g_i, g_j)$	0.0246	3.25	61	416	−0.988

The neighborhood and phylogenetic profiles are obtained by using the reciprocal smallest distance algorithm (33).

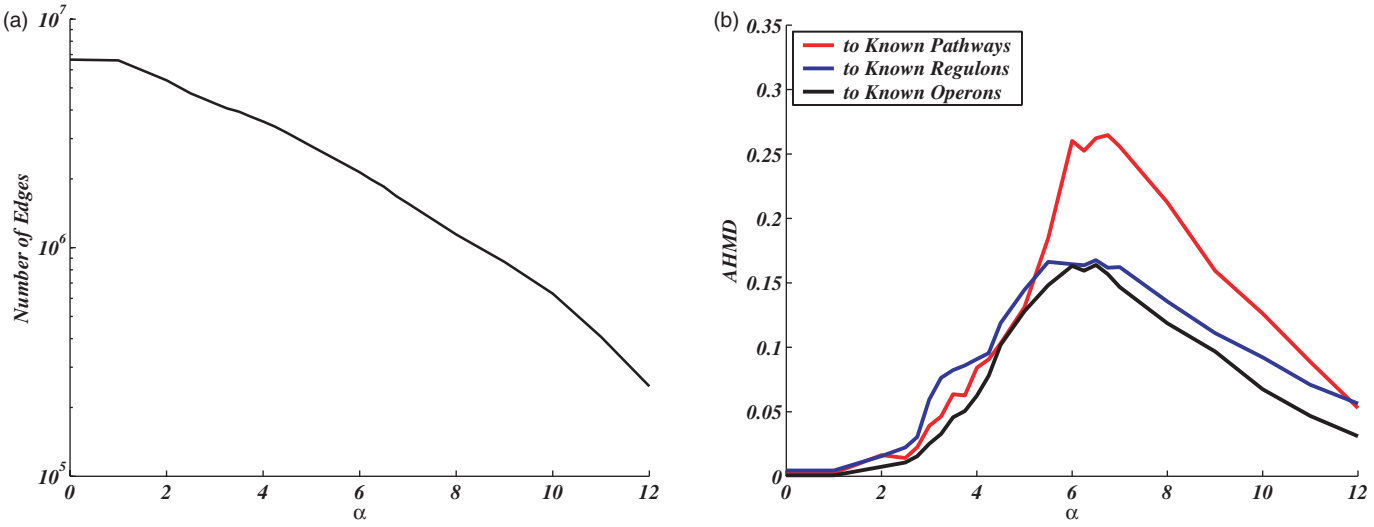


Figure 6. (**a**) The number of edges as a function of α; and (**b**) AHMD values for the known pathways, regulons and operons as functions of α.

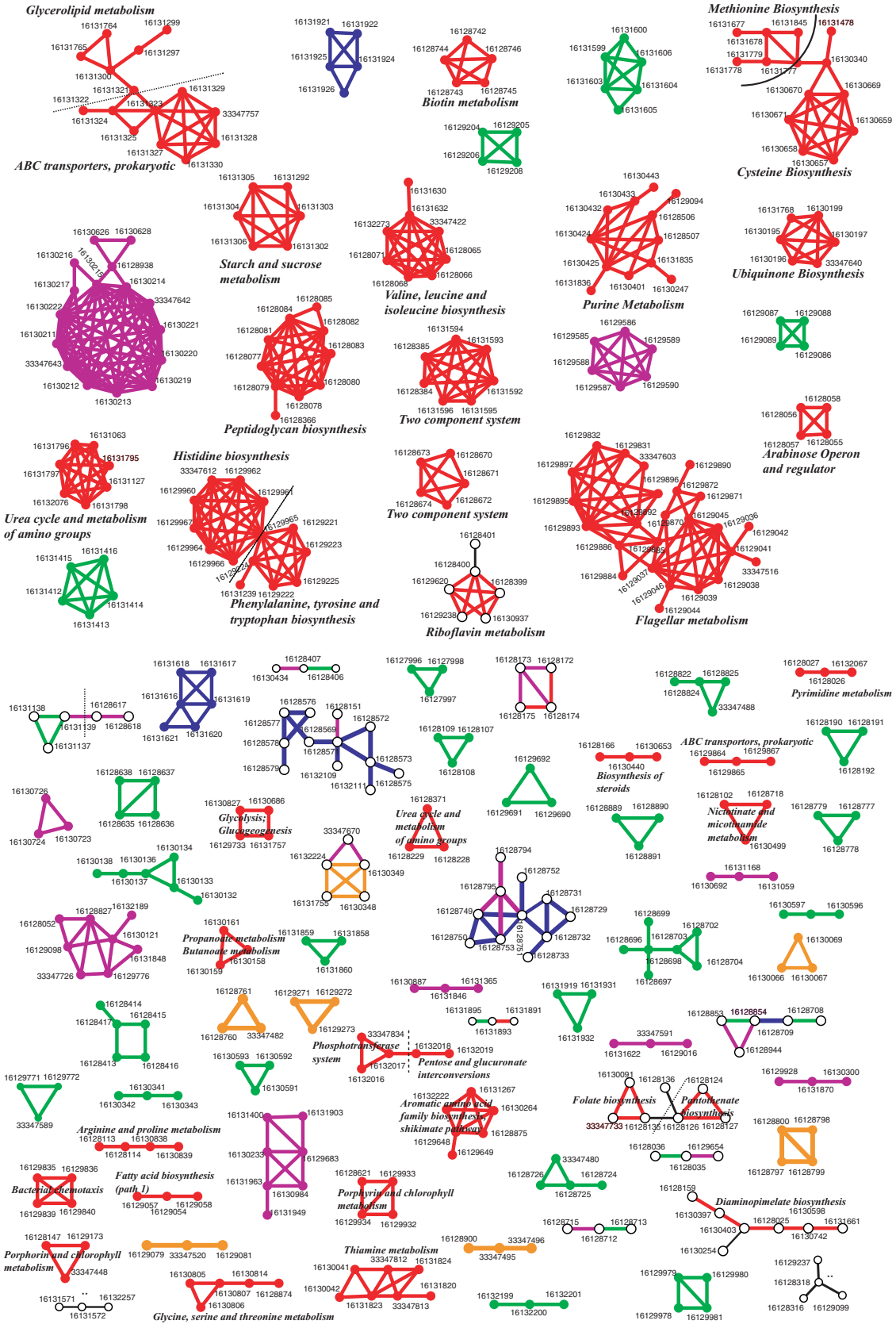


Figure 7. Predicted modules consisting of at least three genes obtained by using $\alpha = 6.75$, where edges of red represent for belonging to the same known pathways, edges of blue represent for belonging to the same known regulons, edges of green represent for belonging to the same known operons, edges of orange represent for transporter unit, edges of purple represent for having similar GO assignments and edges of black represent for having not been experimentally verified.

steps, including $C_{\text{ombined}}(g_i, g_j)$, the maximum AHMD and the associated α-value, are all random in essence. Therefore, we have repeated the above procedure 10 times. Table 3 summarizes the maximum AHMDs and their associated α-values for these 10 experiments.

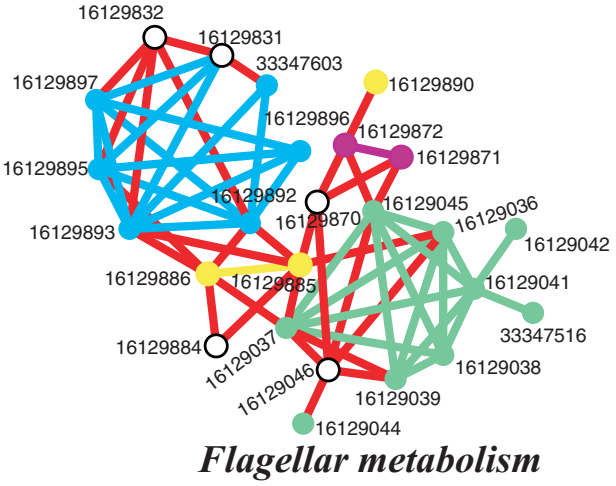


Figure 8. Predicted module corresponding to the flagellar metabolism pathway, where according to Eco Cyc the genes of blue belong to the same regulon, and the genes of yellow, green and dark red belong to three different operons, respectively.

As shown in Table 3, all experiments have achieved similar AHMD for the known pathways, regulons and operons; and have shown that our predicted functional modules are matched with the known pathways better than with the known regulons or operons. Because all the experiments have shown similar performance in terms of AHMD, without loss of generality, we focus on the first experiment for the rest of the analysis.

Comparisons among different sources of information

To see whether the combined score can better describe functional relationships among genes than individual scores, we have also performed experiments on each individual source of information, i.e. we have used $S_{\text{GO}}(g_i, g_j)$, $d(g_i, g_j)$ or $S_{\text{N}}(g_i, g_j)$ alone as a measure of functional relationship between genes to predict functional modules. Table 4 summarizes the maximum AHMD values, the associated values of α, the number (N) of predicted modules, the total number (|**C**|) of genes in all the predicted modules and the associated Z-score for the experiments based on $C_{\text{ombined}}(g_i, g_j)$, $S_{\text{N}}(g_i, g_j)$, $d(g_i, g_j)$ and $S_{\text{GO}}(g_i, g_j)$, respectively.

From Table 4, we can see that for the individual information sources,

(i) The AHMD values and the associated Z-scores achieved by using the phylogenetic profile dissimilarity measure $d(g_i, g_j)$ are much smaller than those achieved by

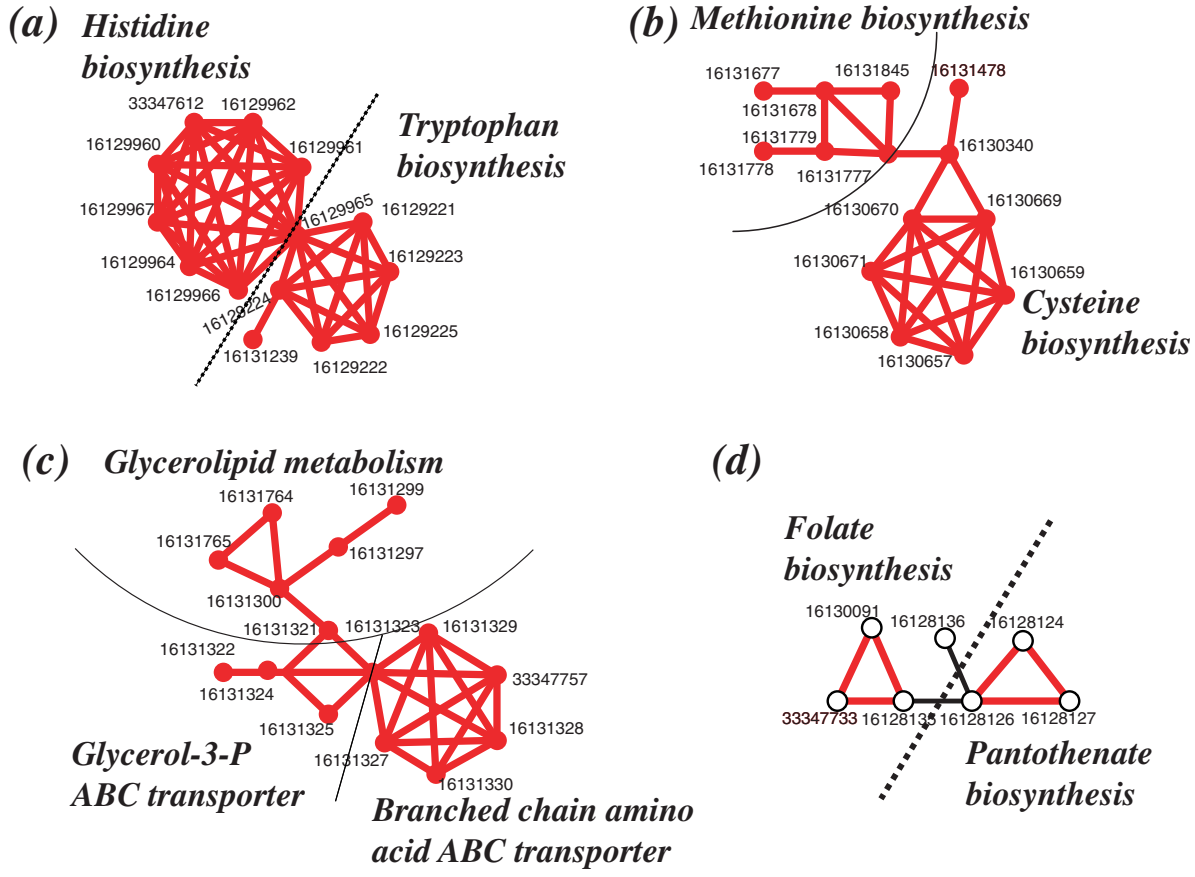


Figure 9. Predicted modules involving more than one pathways.

using the other information sources, which demonstrates that $d(g_i, g_j)$ alone cannot provide sufficient information to predict functional modules that are reasonably consistent with the known functional modules.

(ii) Although the AHMD values for both the GO assignments and the neighborhood profiles are moderate, their associated *Z*-scores are very high, which demonstrates that either $S_{GO}(g_i, g_j)$ or $S_N(g_i, g_j)$ alone already provides sufficient information to achieve sound consistency with the known functional modules. This observation regarding the neighborhood profiles makes it very promising as a measure for the prediction of functional modules for those microbial genomes that have been sequenced but do not have much other information (e.g. GO assignments) available.

(iii) For the GO assignments, the AHMD value and the associated *Z*-score for the known pathways are larger than those for either the known regulons or known operons, which demonstrate that the functional modules predicted by using our approach are more consistent with the known pathways than with the known regulons or known operons. We have made a similar observation for the neighborhood profiles.

and for the combined information,

(i) For the known pathways, the AHMD value achieved by using the combined information $C_{ombined}(g_i, g_j)$ is larger than that of each individual information source, although its *Z*-score is smaller than that of the GO assignments. Because the *Z*-scores for both $C_{ombined}(g_i, g_j)$ and $S_{GO}(g_i, g_j)$ are already very high, and their values of N and $|\mathbf{C}|$ (consequently the sizes of all the predicted modules) are different, the fact that the former is smaller than the latter does not necessarily mean that $C_{ombined}(g_i, g_j)$ is worse than $S_{GO}(g_i, g_j)$; rather, the obvious difference among their AHMD values demonstrates that information fusion can achieve a higher degree of consistency with the known pathways than individual information sources.

(ii) For the known operons, though the neighborhood profiles alone already provide sufficient information to achieve sound consistency, the GO assignments do not. The incapabilities of $S_{GO}(g_i, g_j)$ greatly undermine the capabilities of $S_N(g_i, g_j)$ during the information fusion; hence, either the AHMD value or *Z*-score for the combined information $C_{ombined}(g_i, g_j)$ cannot even be as high as that of $S_N(g_i, g_j)$ alone. We have made similar observations about known regulons.

The observation that phylogenetic profiles do not seem to contribute much to the identification of functional modules, while surprising, is consistent with the observation made by other authors (19). To exclude the possibility that this might be an artifact of the specific prediction method for orthologous genes using BDBH, we have also used the reciprocal smallest distance algorithm (33) for the prediction of orthologous genes, and have then performed the same experiment of comparing different information sources. The results are summarized in Table 5, for which we have made the same observations as above.

Modules predicted using $C_{ombined}(g_i, g_j)$ for a particular choice of α

For the combined information $C_{ombined}$, Figure 6 shows the total number of edges and the AHMD values for the known pathways, regulons and operons as functions of α. Observe from the figure that (i) the number of edges decreases rapidly as the value of α increases; (ii) the tendencies of the three AHMD values all first increase and then decrease as the value of α increases; and (iii) these three AHMD values achieve their maximum around $\alpha \in [6, 7]$.

When using $\alpha = 6.75$, we have obtained 185 predicted functional modules covering 654 genes. Figure 7 shows those predicted modules consisting of at least three genes, where genes are identified by using their PIDs. For most cases, genes within the same predicted modules belong to the same functional modules according to Eco Cyc, or have similar GO assignments. This means that we can predict functional units based on our approach. For example, all genes in the module of Figure 8 are involved in the flagellar metabolism pathway according to KEGG (22), even though only the colored genes and edges are confirmed by Eco Cyc to belong to the same operons or regulons. We have also made the following interesting observations:

(i) Genes within the same predicted module belong to several different pathways, as shown in Figure 9. For the predicted module (a) involving the histidine and the tryptophan biosyntheses, the two pathways are connected through the gene *hisA* (16129965), which is assigned to be involved in both pathways by GO Annotation (16). For the predicted module (b) involving the methionine and the cysteine biosyntheses, the two pathways are connected through the gene *metB* (16131777), which is assigned to be involved in both pathways by KEGG (22). For the predicted module (c) involving the glycerolipid metabolism, glycerol-3-P ABC transporter and branched amino acid ABC transporter, the first two parts are connected because they are both related to glycerol, and the last two parts are connected

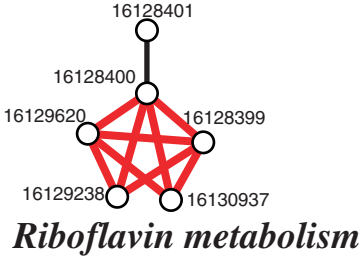


Figure 10. Genes are connected mainly because they are conserved neighboring genes along the same strand of DNA.

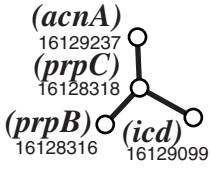
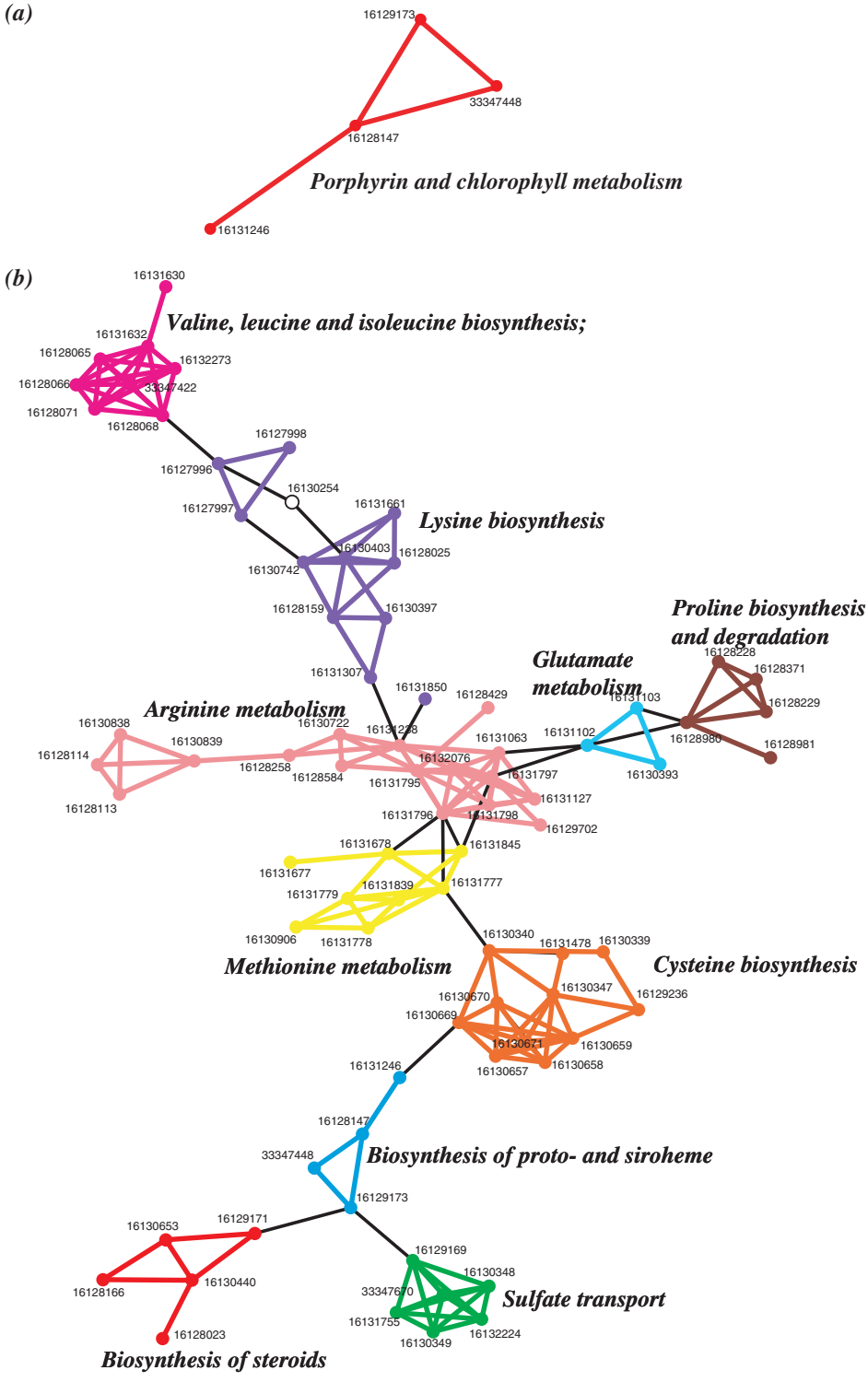


Figure 11. Predicted modules that have not been experimentally verified.

because they have similar GO assignments (i.e. transport). So far, we have not been able to find experimental evidence to support our predicted module in (d) which involves the pantothenate and folate biosyntheses, but this prediction may be due to that both pathways are related to the precursors for co-enzyme biosynthesis (34).

(ii) Genes are connected in a predicted module mainly because they are conserved neighboring genes on the same strand of the genome, as *nusB* (16128401) and *ribH* (16128400) in the module of Figure 10 and other five modules each consisting of one pair of genes. These genes are highly likely to be functionally related, and deserve further experimental investigations.

(iii) Genes are connected via their paralogous genes. For example, we have predicted two modules each consisting of two genes that do not have any obvious commonality.



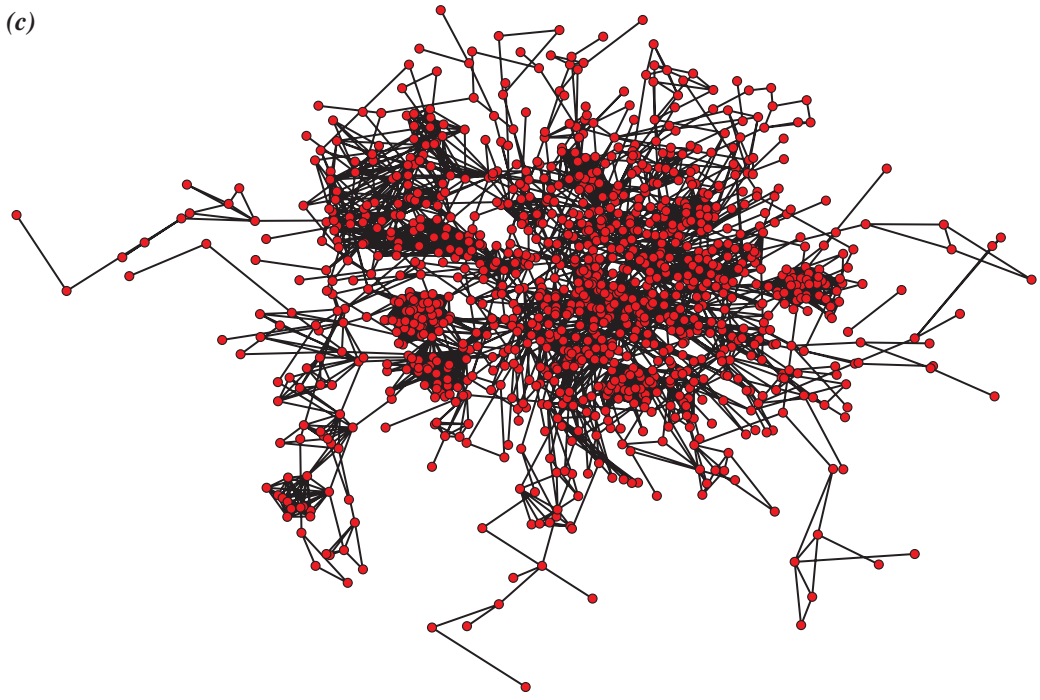


Figure 12. Predicted modules that *hemL* (16128147) is involved for different values of α: (**a**) α = 6.75, (**b**) α = 5.5 and (**c**) α = 4.

For the pair *amtB* (16128436) and *glnB* (16130478), one of the paralogous genes of *glnB*, *glnK* (16128435), belongs to the same operon as *amtB* according to Eco Cyc. For the other pair *purU* (16129193) and *add* (16129581), one of the paralogous genes of *purU*, *purN* (16130425), is involved the same pathway (purine metabolism) as *add*.

(iv) So far, we have not been able to find evidences in Eco Cyc, KEGG or GO to support our predicted modules in Figure 11 and the other five modules each consisting of two genes. They deserve further experimental investigations.

Modules predicted using $C_{ombined}(g_i, g_j)$ for a particular gene

We have also focused on one particular gene *hemL* (16128147) to see how its involved module changes as α is changed. Figure 12 shows three predicted modules that *hemL* is involved for α = 6.75, α = 5.5 and α = 4, respectively. When α = 6.75, the predicted module consists of only four genes, all of which are involved in the porphyrin and chlorophyll metabolism pathway. When α decreases to 5.5, the predicted module consists of 79 genes, which are involved in valine, leucine and isoleucine biosynthesis, lysine biosynthesis, arginine metabolism, glutamate metabolism, proline biosynthesis and degradation, methionine metabolism, cysteine biosynthesis, biosynthesis of proto- and siroheme, biosynthesis of steroids, and sulfate transport pathways, respectively. When α decreases further to 4, the predicted module consists of 1116 out of all 2579 genes.

As we mentioned earlier, the functional relationships among genes can be viewed at different levels. At a very high resolution level (as shown in Figure 12a), only a small number of genes are grouped together so that the group is responsible for one specific activity; at a lower resolution level (as shown in Figure 12b), a relatively larger number of genes are grouped together so that the group is responsible for more general activities; and, at the lowest resolution level (as shown in Figure 12c) most of the genes are connected directly or indirectly so that the group is responsible for most activities of a cell. Consequently, by varying the threshold values, we can predict the hierarchical structure of the functional modules.

CONCLUSIONS

We have presented a new computational method to predict functional modules by combining the information from the comparative genome analysis and the GO in the framework of the Bayesian inference. In this work, we have also developed a formal measure to quantify the degree of consistency between the predicted and the known modules, and provided analysis of the statistical significance for such consistency degrees. We have applied our method to the genome of *E.coli* K12, and have observed that (i) the predicted modules are more consistent with the known pathways than to the known regulons or operons; (ii) neighborhood profiles or GO assignments alone can provide sufficient information for predicting modules that are fairly consistent with the known functional modules, but phylogenetic profiles cannot; (iii) by fusing the information from the GO, phylogenetic and neighborhood profiles using the naive Bayesian inference and using the combined information for module prediction, even higher degrees of consistency can be achieved for the known functional modules; (iv) most of the predicted modules can be verified by Eco Cyc, KEGG or GO, and the unverified predicted modules reveal interesting gene functional relationships that deserve further experimental investigations; and (v) different threshold values can be used to predict functional modules at different resolution levels.

In our future study, we will apply the method presented in this paper to other microbial genomes. Particularly, since we have observed that the neighborhood profiles alone can provide sufficient information for the prediction, we will use the neighborhood profiles to evaluate the gene functional relationships for those microbial genomes that have already been completely sequenced but do not have much other information (e.g. GO) available. We will also generalize the current method of information fusion based on the Bayesian inference to incorporate more sources of information, e.g. microarray data. And finally, we will apply more sophisticated methods to gene clustering so that even higher degrees of consistency can be achieved for the known functional modules.

ACKNOWLEDGEMENTS

This research was supported in part by National Science Foundation (#NSF/DBI-0354771, #NSF/ITR-IIS-0407204) and by the US Department of Energy's Genomes to Life program (http://doegenomestolife.org/) under project 'Carbon Sequestration in *Synechococcus* sp.: From Molecular Machines to Hierarchical Modeling' (www.genomes2life.org). Funding to pay the Open Access publication charges for this article was provided by the University of Georgia.

Conflict of interest statement. None declared.

REFERENCES

1. Kremling,A., Jahreis,K., Lengeler,J.W. and Gilles,E.D. (2000) The organization of metabolic reaction networks: a signal-oriented approach to cellular models. *Metab. Eng.*, **2**, 190–200.
2. Wagner,R. (2000) *Transcription Regulation in Prokaryotes*. Oxford University Press, Oxford, UK.
3. Stephanopoulos,G.N., Aristidou,A.A. and Nielsen,J. (1998) *Metabolic Engineering: Principles and Methodologies*. Academic Press, San Diego, CA.
4. Zhou,J., Thompson,D.K., Xu,Y. and Tiedje,J.M. (2004) *Microbial Functional Genomics*. John Wiley & Sons, Inc., Hoboken, NJ.
5. Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
6. Strauss,E.J. and Falkow,S. (1997) Microbial pathogenesis: genomics and beyond. *Science*, **276**, 707–712.
7. Whittam,T.S. and Bumbaugh,A.C. (2002) Inferences from whole-genome sequences of bacterial pathogens. *Curr. Opin. Genet. Dev.*, **12**, 718–725.
8. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
9. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
10. 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.
11. Manson McGuire,A. and Church,G.M. (2000) Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nucleic Acids Res.*, **28**, 4523–4530.
12. Müller,F., Blader,P. and Strähle,U. (2002) Search for enhancers: teleost models in comparative genomic and transgenic analysis of *cis* regulatory elements. *Bioessays*, **24**, 564–572.
13. Chen,X., Su,Z., Dam,P., Palenik,B., Xu,Y. and Jiang,T. (2004) Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. *Nucleic Acids Res.*, **32**, 2147–2157.
14. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**(Suppl. 1), S329–S336.
15. Gelfand,M.S., Koonin1,E.V. and Mironov,A.A. (2000) Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucleic Acids Res.*, **28**, 695–705.
16. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen1,J., Cox,A. and Apweiler,R. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in Swiss-Prot, Trembl, and Interpro. *Genome Res.*, **13**, 662–672.
17. Chen,Y. (2004) Biological knowledge discovery through mining multiple sources of high-throughput data. PhD thesis. The University of Tennessee, Knoxville, TN.
18. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. and Gerstein,M. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.
19. Lee,I., Date,S.V., Adai,A.T. and Marcotte,E.M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
20. von Mering,C., Zdobnov,E.M., Tsoka,S., Ciccarelli,F.D., Pereira-Leal,J.B., Ouzounis,C.A. and Bork,P. (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA*, **100**, 15428–15433.
21. Karp,P.D., Riley,M., Paley,S.M., Pellegrini-Toole,A. and Krummenacker,M. (1999) Eco Cyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **27**, 55–58.
22. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
23. Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.
24. Yamada,T., Goto,S. and Kanehisa,M. (2004) Extraction of phylogenetic network modules from prokaryote metabolic pathways. *Genome Informatics*, **15**, 249–258.
25. Yanai,I. and DeLisi,C. (2002) The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol.*, **3**, 0064.1–0064.12.
26. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
27. Lord,P.W., Stevens,R.D., Brass,A. and Goble,C.A. (2002) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.
28. Press,W.H., Flannery,B.P., Teukolsky,S.A. and Vetterling,W.T. (1992) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.
29. 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.
30. Duda,R.O., Hart,P.E. and Stork,D.G. (2001) Pattern Classification. John Wiley & Sons, Inc., 2nd edition.
31. Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer-Verlag, NY.
32. Casella,G. and Berger,R.L. (2001) *Statistical Inference, 2nd edn*. Duxbury Press, CA.
33. Wall,D.P., Fraser,H.B. and Hirsh,A.E. (2003) Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1711.
34. Voet,D. and Voet,J.G. (1995) *Biochemistry, 2nd edn*. John Wiley & Sons, Inc, NY.