Prediction of Functional Modules Based on Gene Distributions in Microbial Genomes

Hongwei Wu[1,2]
hongweiw@csbl.bmb.uga.edu

Fenglou Mao[1]
fenglou@csbl.bmb.uga.edu

Zhengchang Su[1,2]
zhx@csbl.bmb.uga.edu

Victor Olman[1]
olman@csbl.bmb.uga.edu

Ying Xu[1,2]
xyn@bmb.uga.edu

[1] Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia,
[2] Computational Biology Institute, Oak Ridge National Laboratory

Abstract

We present a computational method for prediction of functional modules that can be directly applied to the newly sequenced microbial genomes for predicting gene functions and the component genes of biological pathways. We first quantify the functional relatedness among genes based on their distribution (i.e., their existences and orders) across multiple microbial genomes, and obtain a gene network in which every pair of genes is associated with a score representing their functional relatedness. We then apply a threshold-based clustering algorithm to this gene network, and obtain modules for each of which the number of genes is bounded from above by a pre-specified value and the component genes are more strongly functionally related to each other than genes across the predicted modules. Particularly, when the module size is bounded by 130, we obtain 167 functional modules covering 813 genes for *Escherichia coli* K12, and 138 functional modules covering 731 genes for *Bacillus subtilis subsp. subtilis* str. 168. We have used the gene ontology (GO) information to assess the prediction results. The GO similarities among the genes of the same functional module are compared with the GO similarities among the genes that are randomly clustered together. This comparison reveals that our predicted functional modules are statistically and biologically significant, and the genes of the same functional module share more commonality in terms of *biological process* than in terms of *molecular function* or *cellular component*. We have also examined the predicted functional modules that are common to both *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168, and provide explanations for some functional modules.

Keywords: functional modules, microbes, comparative genomics, gene ontology, graph clustering
Supplementary Tables:
http://csbl.bmb.uga.edu/~hongweiw/GIW2005_supplementary/GIW2005_supp.htm

1 Introduction

The wealth of genomic sequence data generated through the worldwide sequencing efforts of microbial genomes [http://www.sanger.ac.uk/Projects/Microbes/, http://www.tigr.org/tdb/mdb/mdbcomplete.html, http://microbialgenome.org, http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html] have provided unprecedented opportunities for computational biologists to unveil the enormous amount of hidden information encoded in the genomes about the biological machinery of these micro-organisms. As we understand now, each of the complex biological processes in a microbial cell is carried out through interactions of multiple functional modules at various levels. These functional modules are made of interacting bio-molecules and serve as the basic building blocks of the complex biological machinery in a microbial cell. In general, functional modules at different levels are made of combinations of operons, regulons, modulons and stimulons, many of which might have *conserved* components and structures across multiple microbial organisms [7, 16, 20, 24].

In our previous study [23], we have used comparative genome analysis and gene ontology (GO) information to predict functional modules in microbial genomes, and have observed that the neighborhood profiles alone can provide sufficient information for accurate prediction. So, in this paper, we present a new computational method for the prediction of functional modules in microbial organisms solely based on genes' neighborhood profiles. More specifically, we (1) propose a probabilistic model for a single gene's neighborhood profile (including its existence and order in a given set of microbial genomes); (2) compute the likelihood of any two genes' neighborhood profiles to represent their functional relatedness; and (3) design a clustering algorithm to obtain modules of strongly functionally related genes from the network where every pair of genes has a certain degree of functional relatedness. The focus of this paper is 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 gene functions and biological (metabolic, signaling and regulatory) networks.

Our method is rationalized by the fact that neighboring genes in prokaryotic genomes are likely to be functionally related (as evidenced by the operon structures). The idea of identifying genes that are consistently close to each other in multiple genome has been an effective approach for predicting genes' functional relatedness, and has been used for the prediction of operons [3, 5, 22], gene-teams (or Über-operons) [8, 12, 13, 17] etc. Different than these approaches, we assess the functional relatedness for each pair of genes based on a probabilistic model developed for their neighborhood profiles. In this way all genes in a genome form a network in which every pair of genes are associated with a score representing their functional relatedness[1]. Though it is only used for the prediction of functional modules in this paper, this gene network contains much more information to be further explored. For example, when viewed at different resolution levels, this gene network exhibits different levels of modularity of the biological machinery in a microbial cell.

The rest of this paper is organized as follows. We describe the materials and methods in Section 2. The experiments, results and discussions are provided in Section 3. Section 4 summarizes the work.

2 Materials and Methods

2.1 Query and Reference Genomes

Supplementary Table 1 summarizes the taxonomy lineages of the 224 microbial genomes used in our study that belong to 175 different species (note that some species may have more than one *strains*). Let G_0 denote the query genome for which the functional modules are to be predicted. When choosing the reference genomes, G_k ($k = 1, \cdots, K$), to establish the neighborhood profiles for the genes of G_0, we remove the redundant information, which is reflected as the multiple genomes of the same species, by (1) not choosing the genomes belonging to the same species as G_0 and (2) only choosing one genome per species.

2.2 Orthologous Gene Detection

Let N_0 be the total number of genes, and g_i be the i-th ($i = 1, \cdots, N_0$) gene of G_0. We use the bi-directional best-hit (BDBH) method to search for the orthologous genes of g_i in reference genomes.

Though the BDBH method has certain limitations (e.g., it may fail for both phylogenetically very close and very distant genomes), it has long been used for orthologous gene prediction. In our previous study, we have used both the BDBH method and the reciprocal smallest distance algorithm (which is an improved version of the BDBH method utilizing both sequence alignment and evolutionary

[1] The probability model used for computing the likelihood of any two genes' neighborhood profiles is slightly different than that of [23] due to the introduction of the concept of *directon*, as detailed in Section 2.

distances) [21] for the detection of orthologous genes, based on which we have then predicted functional modules for *Escherichia coli* K12. These two methods have been shown to lead to similar results [23]. So, in this paper we only use the BDBH method; and we do not distinguish between g_i and its orthologous genes throughout the rest of the paper.

2.3 Neighborhood Profile of One Gene

The neighborhood profile of a gene g_i consists of $K = 175$ in our study entries, each of which represents one reference genome and includes the following information: (1) presence or absence of g_i in the reference genome and (2) the order of g_i in the reference genome if it exists.

The neighborhood profile of g_i can be viewed as an observation of a *random process* that governs the distribution of g_i across different genomes. We make the following assumptions about this random process:

- A gene's behavior (i.e., presence and order) is independent among different reference genomes. Equivalently, the entries of the neighborhood profile of g_i are independent of each other.

- A gene g_i is present in a reference genome G_k with a probability p_{ik}, and p_{ik} is the same among the reference genomes that belong to the same *phylum*. This means that p_{ik} can be estimated by using the maximum likelihood estimation method as:

$$p_{ik} = \frac{\text{number of genomes having } g_i \text{ in the phylum } G_k \text{ belongs to}}{\text{total number of genomes in the phylum } G_k \text{ belongs to}}$$

- If g_i is present in G_k, then it can be located anywhere in G_k with equal probability. Let L be the number of directons (A *directon* consists of genes transcribed in the same direction with no intervening gene transcribed in the opposite direction [14]) of G_k, $N_{k1}, \cdots, N_{kL}$ be the numbers of genes in the L directons, respectively, and $N_k \equiv \sum_{l=1}^{L} N_{kl}$ be the total number of genes of G_k. The probability of g_i belonging to the L-th directon is $p_{ik}N_{kl}/N_k$, and the probability of g_i at a particular position is p_{ik}/N_k.

One may consider a more sophisticated model for the distributions of genes across different genomes, for instance, by taking into account of the lengths (in terms of base pairs) of directons and genes. However, as detailed below, since our method relies more on the model for each pair of genes than for each single gene, this simple model about the single gene's distribution is adequate to provide sufficient information for the assessment of functional relatedness between any pair of genes.

2.4 Neighborhood Profiles of a Pair of Genes

Now consider two genes g_i and g_j. There are five possible cases for their presence and orders in any reference genome G_k:

Case 1. $(\bar{g}_i, \bar{g}_j, G_k)$: neither g_i nor g_j is present in G_k.

Case 2. $(g_i, \bar{g}_j, G_k)$: only g_i is present in $G_k$but g_j is not.

Case 3. $(\bar{g}_i, g_j, G_k)$: only g_j is present in G_k but g_i is not.

Case 4. $(g_i, g_j, |g_i - g_j| = d, G_k)$: both g_i and g_j are present in G_k, and belong to the same directon with a distance of d genes.

Case 5. (g_i, g_j, NA, G_k): both g_i and g_j are present in G_k, but do not belong to the same directon.

If g_i and g_j are not functionally related to each other in any way, then their distributions in any reference genome G_k can be viewed as independent, and therefore the probability for the above five cases[2] can be computed as:

[2] Strictly speaking, instead of computing the probability of Case 4, we actually compute the probability that g_i and g_j are present in the same directon of G_k with a distance no more than d genes.

Case 1. the probability that neither g_i nor g_j is present in G_k can be computed as:

$$P(\bar{g}_i, \bar{g}_j, G_k) = (1 - p_{ik})(1 - p_{jk})$$

Case 2. the probability that only g_i is present in G_k but g_j is not can be computed as:

$$P(g_i, \bar{g}_j, G_k) = p_{ik}(1 - p_{jk})$$

Case 3. the probability that only g_j is present in G_k but g_i is not can be computed as:

$$P(\bar{g}_i, g_j, G_k) = (1 - p_{ik})p_{jk}$$

Case 4. the probability that both g_i and g_j are present in G_k, and belong to the same directon with a distance no more than d genes can be computed as follow:

$$P(g_i, g_j, |g_i - g_j| \le d, G_k) = p_{ik}p_{jk} \sum_{l=1}^{L} \Big(\frac{N_{kl}}{N_k}\Big)^2 \min\Big\{\frac{N_{kl} + (2N_{kl} - 1)d - d^2}{N_{kl}^2}, 1\Big\}$$

Case 5. the probability that both g_i and g_j are present in G_k, but do not belong to the same directon can be computed as:

$$P(g_i, g_j, NA, G_k) = p_{ik}p_{jk}\Big(1 - \sum_{l=1}^{L}\Big(\frac{N_{kl}}{N_k}\Big)^2\Big)$$

When all reference genomes are considered together, the log-likelihood of the neighborhood profiles of g_i and g_j, $L(g_i, g_j)$, which supports the hypothesis that g_i and g_j are not functionally related, is computed as:

$$\begin{aligned} L(g_i, g_j) \;\; = \;\; & \sum_{k=1}^{K} \{ I(\bar{g}_i, \bar{g}_j, G_k) \log P(\bar{g}_i, \bar{g}_j, G_k) + I(g_i, \bar{g}_j, G_k) \log P(g_i, \bar{g}_j, G_k) \\ & + I(\bar{g}_i, g_j, G_k) \log P(\bar{g}_i, g_j, G_k) + I(g_i, g_j, NA, G_k) \log P(g_i, g_j, NA, G_k) \\ & + I(g_i, g_j, |g_i - g_j| \le d_k^{ij}, G_k) \log P(g_i, g_j, |g_i - g_j| \le d_k^{ij}, G_k)\}, \end{aligned}$$

where $I(\,\cdot\,)$ is an indicator, which is 1 if and only if the criteria within the parentheses are met, and d_k^{ij} is the observed distance between g_i and g_j in G_k. The larger $L(g_i, g_j)$ is, the more supportive the neighborhood profiles of g_i and g_j are for this hypothesis; and the smaller $L(g_i, g_j)$ is, the more supportive the neighborhood profiles of g_i and g_j are for the alternative hypothesis, which is that g_i and g_j are functionally related. So, in the rest of this paper, we use $S(g_i, g_j) = -L(g_i, g_j)$ to denote the score for the functional relatedness between g_i and g_j.

2.5 Clustering Algorithm

Every pair of genes g_i and g_j in the query genome G_0 has a score $S(g_i, g_j)$ measuring their functional relatedness, as defined above. The query genome G_0 can now be viewed as a graph, where nodes represent genes and edges with weights $S(g_i, g_j)$ represent the functional relatedness between the corresponding genes. This graph of functional relatedness can be interpreted at different levels. At the coarsest resolution level, all genes are functionally related so that they together are responsible for all activities of a cell. At a finer resolution level, genes with stronger functional relatedness stand out and form smaller and strongly interacted modules that are responsible for specific activities of a cell. At the finest resolution level, each gene forms a functional module by itself.

To predict biologically meaningful functional modules of smaller sizes, we apply a simple threshold-based method to partition the graph into intra-connected clusters. The principle behind this algorithm

```
Clustering Algorithm (MaxSize, initial value of $\alpha$)
Initialization
    There is only one module $C$ consisting of all genes
    If $|C| \geq$ MaxSize ($|C|$ is the number of genes in $C$)
        $C$ is added into the temporary module collection $\mathbf{C}_{\text{temp}}$
    Else
        $C$ is added into the final module collection $\mathbf{C}_{\text{final}}$
    End if
Loop until $\mathbf{C}_{\text{temp}}$ is empty
    Take one module $C$ out of $\mathbf{C}_{\text{temp}}$
    Loop for each gene $g_i$ in $C$
        Compute $\mu_i$ and $\sigma_i$ of $S(g_i, g_j)$ between $g_i$ and all the other genes $g_j$ in $C$
    End loop
    Loop for every two genes $g_i$ and $g_j$ in $C$
        Keep the connection between $g_i$ and $g_j$ if and only if $S(g_i, g_j)$ stands out
    End loop
    Loop for each newly generated module $C'$
        If $|C'| \geq$ MaxSize
            $C'$ is added into the temporary collection $\mathbf{C}_{\text{temp}}$
        Else
            $C'$ is added into the final collection $\mathbf{C}_{\text{final}}$
        End if
    End loop
    $\alpha$ is increased by $\Delta\alpha$
End loop
```

Figure 1: Pseudo-code of the clustering algorithm.

is that two genes being clustered together at a particular resolution level will be further clustered together at a finer resolution level if and only if their functional relatedness *stands out* (compared to the functional relatedness of other gene pairs) in the current module. More specifically, let μ_i and σ_i be the mean and standard deviation of the functional relatedness between g_i and all the other genes of the same module at the current resolution level, then $S(g_i, g_j)$ is considered to *stand out* in the current module if and only if $S(g_i, g_j) \geq \mu_i + \alpha\sigma_i$ and $S(g_i, g_j) \geq \mu_j + \alpha\sigma_j$ being a threshold. In our clustering algorithm,

- each module corresponds to a connected (sub-)graph.
- the initial value of α is provided to the algorithm by the user.
- a module will be stopped from further decomposition if the number of its genes is no greater than a pre-specified value *MaxSize*; otherwise, it will be further decomposed after the current value of α is increased by a small amount $\Delta\alpha(= 0.05$ in our study).

Our clustering algorithm is depicted in Figure 1.

When *MaxSize* is set as 1, all the functional modules obtained throughout this clustering process, including all of those belonging to $\mathbf{C}_{\text{final}}$ or any $\mathbf{C}_{\text{temp}}$ during the process, form a *tree*, wherein (1) a node corresponds to a module, (2) the root module consists of all the genes in G_0, (3) each leaf module consists of a single gene, (4) the modules at the same level are associated with the same value of α, (5) a module contains only a subset of the genes that belong to its parent module, and (6) two modules do not share any common genes if they are not ancestor-descendent. This tree structure of functional modules is solely determined by the initial value of α and the strategy that α is varied during the clustering process (α is linearly increased in our study). This tree reflects the hierarchical structure of our functional module. Given a particular value of *MaxSize* (other than 1), the final collection $\mathbf{C}_{\text{final}}$ contains those modules with size no greater than *MaxSize* that are first encountered while traversing the tree from the root to leaves. Generally, different values of *MaxSize* correspond to viewing the functional modules at different resolution levels. Given two values of *MaxSize*, $MaxSize_1 \leq MaxSize_2$, each of the modules obtained at $MaxSize_1$ must be the descendent of some module obtained at $MaxSize_2$. Therefore, a larger value of *MaxSize* can be interpreted to correspond to a coarser resolution level.

Through the experiments where *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168 are used as the query genome, respectively, we have observed that the value of *MaxSize* affects the clustering result more significantly than the initial value of α. In this study, the initial value of α is set to be 0, which means that the connection between g_i and g_j will be immediately cut out at the beginning of the clustering process if either $S(g_i, g_j) < \mu_i$ or $S(g_i, g_j) < \mu_j$.

3 Experiments and Discussion

3.1 Functional Modules at Different Resolution Levels

To see how the value of *MaxSize* affects the clustering results, we consider those predicted functional modules that have at least three genes. Figure 2 shows the number of such modules, the number of covered genes, and the average number of genes per module as functions of *MaxSize* for both *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168. Observe from the figure that for both genomes the number of predicted functional modules and the number of covered genes become relatively stable starting from *MaxSize* = 130. Therefore, in the following detailed study of the predicted functional modules, we choose *MaxSize* = 130 for both genomes.

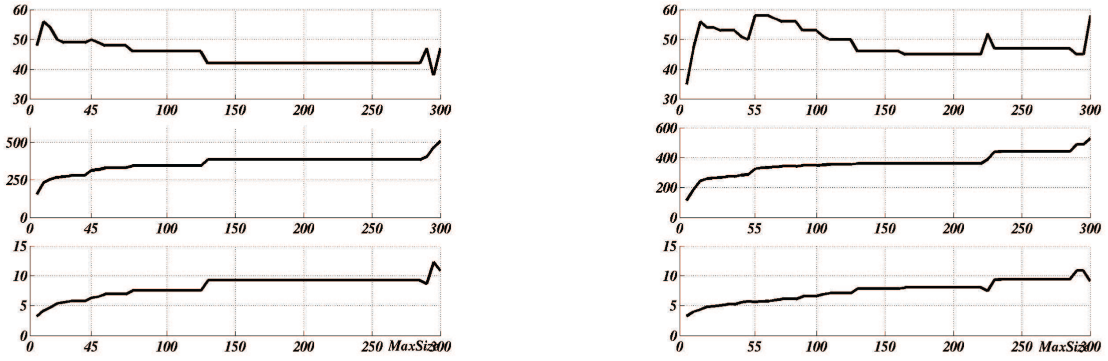

Figure 2: *Escherichia coli* K12 (left) and *Bacillus subtilis subsp. subtilis* str. 168 (right). From top to bottom, the number of predicted functional modules, the number of covered genes, and the average number of genes per module as functions of *MaxSize* (horizontal axis).

3.2 Evaluation of Prediction Results by Using the GO Information

With *MaxSize* being set at 130, we have obtained 167 functional modules covering 813 genes for *Escherichia coli* K12, and 138 functional modules covering 731 genes for *Bacillus subtilis subsp. subtilis* str. 168, where each functional module has at least two genes in. These modules are summarized in Supplementary Tables 2 and 3, respectively.

To evaluate our prediction results, we have applied the GO information. The three gene ontologies — *molecular function*, *biological process*, and *cellular component* — describe the attributes of gene products from different perspectives, 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 [19]. A directed acyclic graph can be induced from each GO term V, wherein:

- at the root level is the term *Gene_Ontology*, at the bottommost level is V itself, and in between are the ancestor GO terms of V.

- the relationship between the child and parent terms is interpreted as that the child term is either an *instance* or a *component* of the parent term, which means that the child term is always more specific than its parent terms in describing the attributes of the gene product.

We use the same method as [23] to measure the similarity between GO terms and between genes. Let V_1 and V_2 be the directed acyclic graphs of two GO terms, then their similarity, $s^{\mathrm{GO}}(V_1, V_2)$, is defined as:

$$s^{\mathrm{GO}}(V_1, V_2) \equiv \max \ \mathrm{depth}(\mathrm{node}|\mathrm{node} \in V_1 \cap V_2),$$

where $V_1 \cap V_2$ refers to the nodes common to both V_1 and V_2, and the depth of a node is defined as the number of nodes along the longest path from the node to the root. Let $V(g)$ be the collection of all the GO terms assigned to the gene g, then the GO similarity s^{gene} between two genes g_i and g_j is defined as:

$$s^{\mathrm{gene}}(g_1, g_2) \equiv \max_{V_i \in \mathbf{V}(g_i), V_j \in \mathbf{V}(g_j)} s^{\mathrm{GO}}(V_i, V_j)$$

The above defined GO similarity is very similar to the information-content based semantic similarity defined in [11] in that both definitions consider specificity of the common attributes of two genes. We have used the GO annotations provided by the GO Annotation Project [1] for *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168.

Let $\mathbf{C} \equiv \{C_1, C_2, \cdots, C_M\}$ be the collection of the predicted modules with C_m being a predicted module and M being the number of predicted modules, where $m = 1, 2, \cdots, M$. For each C_m, we first compute the GO similarity between any two genes that both belong to C_m so that there are totally $n_m \equiv |C_m|(|C_m| - 1)/2$ such GO similarity measures; we then compute the average ($\bar{s}_m^{\mathrm{gene}}$) of these n_m GO similarity measures. In this way, the collection C is associated with a collection of numbers $\{\bar{s}_1^{\mathrm{gene}}, \bar{s}_2^{\mathrm{gene}}, \cdots, \bar{s}_M^{\mathrm{gene}}\}$, each of which is the average GO similarity within a predicted module. To assess the statistical significance of $\{\bar{s}_1^{\mathrm{gene}}, \bar{s}_2^{\mathrm{gene}}, \cdots, \bar{s}_M^{\mathrm{gene}}\}$ (as well as of $\{C_1, C_2, \cdots, C_M\}$), we first estimate[3] the means $\{\mathrm{mean}_1, \mathrm{mean}_2, \cdots, \mathrm{mean}_M\}$ and standard deviations $\{\mathrm{std}_1, \mathrm{std}_2, \cdots, \mathrm{std}_M\}$ of the same measures (i.e., the average GO similarities within modules) for randomly generated functional modules, and then compute the Z-score of $\{\bar{s}_1^{\mathrm{gene}}, \bar{s}_2^{\mathrm{gene}}, \cdots, \bar{s}_M^{\mathrm{gene}}\}$ as:

$$Z\text{-score} = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \frac{\bar{s}_m^{\mathrm{gene}} - \mathrm{mean}_m}{\mathrm{std}_m}$$

When $\{\bar{s}_1^{\mathrm{gene}}, \bar{s}_2^{\mathrm{gene}}, \cdots, \bar{s}_M^{\mathrm{gene}}\}$ is for a collection of randomly generated functional modules, the Z-score asymptotically follows a normal distribution [2]. Therefore, a high Z-score means that the collection of the predicted functional modules $\{C_1, C_2, \cdots, C_M\}$ is statistically significant. Table 1 summarizes the Z-scores of $\{\bar{s}_1^{\mathrm{gene}}, \bar{s}_2^{\mathrm{gene}}, \cdots, \bar{s}_M^{\mathrm{gene}}\}$ when different GOs are used for the GO similarity measures. Observe from the table that the Z-scores corresponding to the *biological process* GO are the highest for both *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168. So, statistically speaking, the genes that are predicted to belong to the same functional module share more commonality in terms of *biological process* than in terms of *molecular function* or *cellular component*, which is clearly consistent with our intention to capture genes working in the same biological processes.

[3] 1000 iterations have been run for the estimation. In each iteration, a collection of M functional modules are generated by (1) first randomly choosing without replacement $\sum_{m=1}^{M} |C_m|$ genes out of the pool, and (2) randomly clustering the chosen genes into M clusters with the m-th cluster of size of $C_m (m = 1, 2, \cdots, M)$.

Table 1: The Z-scores of $\{\bar{s}_1^{\text{gene}}, \bar{s}_2^{\text{gene}}, \cdots, \bar{s}_M^{\text{gene}}\}$, which correspond to the predicted functional modules $\{C_1, C_2, \cdots, C_M\}$, for both *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168, when different gene ontologies are used for the GO similarity measures.

	Escherichia coli K12	*Bacillus subtilis subsp. subtilis* str. 168
Biological process	25.3277	27.0246
Molecular function	21.0444	16.2324
Cellular component	4.34677	2.74972

Also, Figure 3 shows the distribution of the *biological process* GO similarity among the genes that are predicted to belong to the same functional module *versus* that among all genes, and Table 2 summarizes the mean and standard deviations of these distributions. Observe from the figure and the table that (1) the distribution of the *biological process* GO similarity among the genes that are predicted to belong to the same functional module is very different than the one among all genes, and (2) a pair of genes with *biological process* GO similarity ≥ 4 is more likely to belong to the same predicted functional module than not.

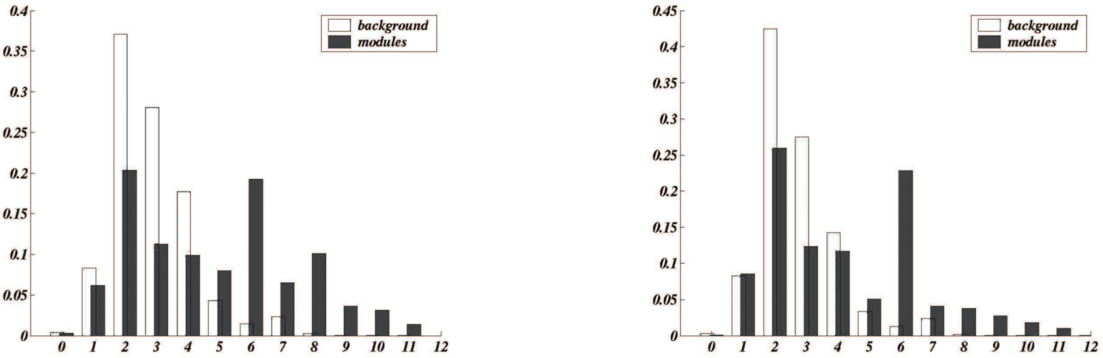

Figure 3: *Escherichia coli* K12 (left) and *Bacillus subtilis subsp. subtilis* str. 168 (right). The distribution of the GO similarities among the genes that are predicted to belong to the same functional module (modules, in red) versus the distribution of the GO similarities among all genes (background, in blue).

Table 2: Means and standard deviations (SD) of the GO similarities among the genes that are predicted to belong to the same functional module (module) and among all genes (background), for *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168.

	Escherichia coli K12		*Bacillus subtilis subsp. subtilis* str. 168	
	Mean	SD	Mean	SD
Background	2.8672	1.2899	2.7548	1.2492
Module	3.6105	2.0270	3.2692	1.8069

3.3 Functional Modules Mapped between *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168

Supplementary Tale 4 summarizes the 86 predicted modules that are common (i.e., including at least one common gene) to both *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168 when

MaxSize is set at 130. Among them, there are 76 modules having at least two genes in common, and 28 modules having at least three genes in common. Since *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168 belong to different phylum (proteobacteria and firmicutes, respectively) of bacteria, their commonality may represent (at least part of) the commonality within bacteria.

Among these common functional modules, there are 38 in *Escherichia coli* K12, and 35 in *Bacillus subtilis subsp. subtilis* str. 168, each of which includes a subset of genes of a transcriptional unit. For example, in *Escherichia coli* K12, all the five genes of a five-gene transcription unit cyoABCDE are predicted to belong to the same functional module; and, in *Bacillus subtilis subsp. subtilis* str. 168, six genes purEKCSQL of a twelve-gene transcription unit purEKBCSQLFMNHD are predicted to belong to the same functional module [`http://biocyc.org/`]. This demonstrates, on one hand, that genes of the same transcriptional unit tend to be together during evolution so that their co-transcription and co-expression relationships are preserved; and, on the other hand, that genes of the same transcription unit undergo re-arrangements during evolution so that not all of them are together all the time.

We have observed from Supplementary Table 4 that a functional module can consist of genes that are not in the same transcription unit but are involved in the same biological process. For example, in *Escherichia coli* K12, seven genes purKECMNDH, which are in four different transcription units (purEK, purC, purMN, and purHD, respectively) but are all involved in the same pathway *IMP biosynthesis*, are predicted to belong to the same functional module; and in *Bacillus subtilis subsp. subtilis* str. 168, four genes, thiC-yjbV-thiEM, which belong to at least three different transcription units but are all involved in the same pathway *thiamine-PP biosynthesis*, are predicted to belong to the same functional module [`http://biocyc.org/`, `http://www.ebi.uniprot.org/`]. Particularly,

- [Supplementary Table 5] We have predicted 53 genes of *Escherichia coli* K12 that belong to at least five different transcription units into one functional module; and predict 33 genes of *Bacillus subtilis subsp. subtilis* str. 168 that belong to at least three different transcription units into one functional module. These two modules of the two genomes share 33 common genes that encode ribosomal proteins, elongation factors, secretory proteins, adenylate kinase[4], and RNA polymerases, respectively. This indicates that these two functional modules are related to the process of *DNA-transcription and translation* in both genomes. Interestingly, 19 genes of *Escherichia coli* K12 that seem to be less relevant to *DNA-transcription and translation* based on their NCBI annotations are also included in this functional module. We have found the following evidence to support the inclusions of infA-lysSU-hlsR-ispFD-rsmC into this functional module.

a. infA encodes the translation initiation factor IF-1.

b. lysS is in an operon (prfB-lysS) that encodes the peptide-chain-release factor 2 [6]; and, lysS and lysU are the two lysyl-tRNA ligases, where lysS is expressed constitutively and lysU is heat inducible [9].

c. The protein that hlsR encodes interacts with 15 proteins encoded by rplIMNSV-rpmB-rpsBCEGJ-sbcC-yagG-ycfS-ydhA, respectively [`http://www.ebi.uniprot.org/`], where rplNV and rpsCEGJ, encoding ribosomal proteins, are already included in this functional module.

d. The protein encoded by ispF interacts with three proteins encoded by flgD, hldD and rpsJ, respectively [`http://www.ebi.uniprot.org/`], where rpsJ is already included in this functional module; and, the proteins encodes by ispD and ispF catalyze step 2 and 4 of the process *isopentenyl-PP from 1-deoxy-D-xylulose 5-phosphate*, respectively.

e. rsmC encodes the ribosomal RNA small subunit methyltransferase. Also, we have found the following evidence to support the prediction that trpABCDE, cysKE and pabAB are clustered into the same functional module.

[4] The protein that adk encodes, adenylate kinase, interacts with two other proteins encoded by rpoC and ybdL, respectively, where rpoC, encoding the RNA polymerase beta subunit, is involved in the transcription process. This justifies why adk is included into this functional module in both genomes.

f. trpABCDE and cysKE are involved in the biosynthesis of two particular amino-acids, tryptophan and cystine, respectively; additionally, cysK is homologous to trpB in several genomes citebib21.

g. The pathway *tryptophan biosynthesis* of which trpABCDE are part and the pathway *tetrahydrofolate biosynthesis* of which pabAB are part both start with chorismate. Particularly, the process of *anthranilate from chorismate* is catalyzed by trpDE in *Escherichia coli* K12, and by pabA-trpE in *Bacillus subtilis subsp. subtilis* str. 168 [`http://www.ebi.uniprot.org/`].

• [Supplementary Table 6] We have predicted 23 genes of *Escherichia coli* K12 that belong to at least four different transcription units into one functional module, and predicted 27 genes of *Bacillus subtilis subsp. subtilis* str. 168 that belong to at least six different transcription units into one functional module. These two modules of the two genomes share 21 common genes that encode 17 flagellar-related proteins, and four chemotaxis-related proteins. Among the other eight genes that are unique to either *Escherichia coli* K12 or *Bacillus subtilis subsp. subtilis* str. 168, at least seven also encode flagellar- or chemotaxis-related proteins. The prediction that the flagellar- and chemotaxis-related genes are in the same functional module can be supported by the works of [4, 15].

a. the chemotaxis-related proteins, which are capable of detecting the changes in the concentration of attractants and repellents, regulate the movement of a cell.

b. the movement of a cell is accomplished through flagellar rotations.

• [Supplementary Table 7] We have predicted 14 genes of *Escherichia coli* K12 that belong to at least three different transcription units into one functional module, and predicted 14 genes of *Bacillus subtilis subsp. subtilis* str. 168 that belong to at least four different transcription units into one functional module. Eleven of the 12 common genes, and the two remaining genes that are unique to either *Escherichia coli* K12 or *Bacillus subtilis subsp. subtilis* str. 168, are all known to be related to the formation of the cell envelope and cell division. So, it is interesting to investigate the roles of the other uncharacterized genes, yfiH-yggS of *Escherichia coli* K12 and ylmE-ylmG of *Bacillus subtilis subsp. subtilis* str. 168, in the same biological process.

As we have mentioned earlier, different values of *MaxSize* correspond to viewing the functional modules at different resolution levels. The reason that we choose $MaxSize = 130$ for our detailed analysis is that for this particular choice some quantitative measures (i.e., the number of predicted modules, the number of covered genes, and the average number of genes per functional module) are relatively stable for both genomes. As demonstrated by the following example, this criterion is not necessarily universally true or biologically motivated.

• [Supplementary Table 8] When *MaxSize* is set at 130, we have predicted 92 genes of *Escherichia coli* K12 that belong to more than ten different transcription units into one functional module, and predicted 127 genes of *Bacillus subtilis subsp. subtilis* str. 168 that belong to more than 16 transcription units into one functional module. These two functional modules in the two genomes share 37 common genes. However, even these common genes are involved in so diverse biological processes, which might indicate that both predicted modules may not have common biological goals. When *MaxSize* is set at 10 (which corresponds to viewing the functional relatedness among genes at a much finer resolution level), 49 of these 92 *Escherichia coli* K12 genes are predicted to belong to 15 different functional modules; 95 of these 127 *Bacillus subtilis subsp. subtilis* str. 168 genes are predicted to belong to 30 different functional modules; and all the other genes become isolated. Among these functional modules predicted at $MaxSiz = 10$, there are 11 common to both genomes, which are related to *zinc/manganese transportation*, *arginine biosynthesis*, *oligopeptide transportation*, *molybdate transportation*, *dTDP-rhamnose biosynthesis*, *glycogen biosynthesis/degradation*, *phosphate transportation*, *glycine cleavage*, *sn-glycerol 3-phosphate transportation*, and *ribose transportation*, respectively. It will be interesting to investigate the functional relatedness among these smaller-sized modules (e.g., predicted at $MaxSize = 10$) to unveil the unified biological goals of the larger-sized modules (e.g., predicted at $MaxSize = 130$).

As we have understood about the biological machinery of microbial genomes, the functional modules are organized in a hierarchical way. At the root level is the single module consisting of all the genes in the query genome; and at the leaf level are the modules each of which consists of only one gene. Biological processes of different degrees of complexity may involve different numbers of genes; hence, their corresponding functional modules may be of different sizes. This may suggest that using a uniform value of *MaxSize* is not appropriate for all biological processes, and some biological meaningful guidance should be used during the clustering.

4 Conclusions

In this paper we have developed a computational method for the prediction of functional modules in microbial genomes. Since our proposed method is purely computational, it can be directly applied to the newly sequenced microbial genomes to predict gene functions and/or the component genes of biological pathways. These predictions could possibly be used to guide experimental designs for investigating particular biological processes.

We have first quantified the functional relatedness among genes based on their distributions across multiple microbial genomes, and have then applied a threshold-based clustering algorithm to obtain modules from the gene network in which every pair of genes is associated with a score representing their functional relatedness. We have used GO information to assess the prediction results, and have looked into the predicted functional modules that are common to both *Escherichia coli* K12 and *Bacillus subtilis subsp. subtilis* str. 168.

Because our method predicts functional relatedness among genes based on their distributions across multiple microbial genomes, it heavily depends on the accuracy of the detection of orthologous genes. The BDBH method, which has been used in our study and performs well for most cases, may still fail for both phylogenetically very close and very distant genomes. For phylogenetically very close genomes, it is possible that the true orthologous genes are missed due to small and insignificant differences in sequence alignment scores among paralogous genes. For phylogenetically very distant genomes, it is possible that the false orthologous genes are predicted due to the (relatively low) similarity between genes that contain similar domains. We have found a number of cases where a functional module is formed due to the inclusion of paralogous genes (e.g., trpABCDE and cysKE are clustered together due to the paralogy between trpB and cysK [Supplementary Table 5]). So, in our future study, we plan to assess the functional relatedness of any two genes based on their COG annotations [18]. In this way, the sensitivity level for the detection of orthologous genes is increased (because most of time orthologous genes belong to the same COG) at the price that the specificity level is decreased (because each COG also contains in-paralogous genes). However, an extra bonus of using COG is that it is possible to predict "universal" functional modules that can be mapped to all microbial genomes.

Clustering is also a key step during the prediction of functional modules. The clustering algorithm used in our study, though simple and effective, may lack in a mathematical and biological basis. We plan to use the known modules (e.g., pathways) to guide the clustering in our future study.

Acknowledgments

This research was supported in part by National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204), by 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`] and also by a "Distinguished Cancer Scholar" grant from the Georgia Cancer Coalition.

References

[1] Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., and Apweiler, R., The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro, *Genome Res.*, 13:662–672, 2003.

[2] Casella, G. and Berger, R. L., *Statistical Inference, 2nd ed*, Duxbury Press, 2001.

[3] Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y., and Jiang, T., Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome, *Nucleic Acids Res.*, 32:2147–2157, 2004.

[4] Dahlquist, F. W., Amplification of signaling events in bacteria, *Sci. STKE*, 2002, PE24, 2002.

[5] Ermolaeva, M. D., White, O., and Salzberg, S. L., Prediction of operons in microbial genomes, *Nucleic Acids Res.*, 29:1216–1221, 2001.

[6] Kawakami, K., Jonsson, Y. H., Bjork, G. R., Ikeda, H., and Nakamura, Y., Chromosomal location and structure of the operon encoding peptide-chain-release factor 2 of Escherichia coli, *Proc. Natl. Acad. Sci. USA*, 85:5620–5624, 1988.

[7] Kremling, A., Jahreis, K. J., Lengeler, W., and Gilles, E. D., The organization of metabolic reaction networks: a signal-oriented approach to cellular models, *Metab. Eng.*, 2:190–200, 2000.

[8] Lathe, 3rd, W. C., Snel, B., and Bork, P., Gene context conservation of a higher order than operons, *Trends Biochem. Sci.*, 25:474–479, 2000.

[9] Leveque, F., Plateau, P., Dessen, P., and Blanquet, S., Homology of lysS and lysU, the two Escherichia coli genes encoding distinct lysyl-tRNA synthetase species, *Nucleic Acids Res.*, 18:305–312, 1990.

[10] Levy, S. and Danchin, A., Phylogeny of metabolic pathways: O-acetylserine sulphydrylase A is homologous to the tryptophan synthase beta subunit, *Mol. Microbiol.*, 2:777–783, 1988.

[11] Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A., Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics*, 19:1275–1283, 2003.

[12] Luc, N., Risler, J. L., Bergeron, A., and Raffinot, M., Gene teams: a new formalization of gene clusters for comparative genomics, *Comput. Biol. Chem.*, 27:59–67, 2003.

[13] Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A., and Koonin, E. V., Connected gene neighborhoods in prokaryotic genomes, *Nucleic Acids Res.*, 30:2212–2223, 2002.

[14] Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J., Operons in Escherichia coli: genomic analyses and predictions, *Proc. Natl. Acad. Sci. USA*, 97:6652–6657, 2000.

[15] Simon, M. I., Borkovich, K. A., Bourret, R. B., and Hess, J. F., Protein phosphorylation in the bacterial chemotaxis system, *Biochimie.*, 71:1013–1019, 1989.

[16] Stephanopoulos, A. A. A. G. N. and Nielsen, J., *Metabolic Engineering: Principles and Methodologies.* Academic Press, 1998.

[17] Sun Kim, J.-H. C. and Jiyoung Yang, Gene Teams with Relaxed Proximity Constraint, *IEEE Computational Systems Bioinformatics (CSB'05)*, 44–55, 2005.

[18] Tatusov, R. L., Koonin, E. V., and Lipman, D. J., A genomic perspective on protein families, *Science*, 278:631–637, 1997.

[19] The Gene Ontology Consortium, Creating the gene ontology resource: design and implementation, *Genome Res.*, 11:1425–1433, 2001.

[20] Wagner, R., *Transcription Regulation in Prokaryotes*, Oxford University Press, 2000.

[21] Wall, D. P., Fraser, H. B., and Hirsh, A. E., Detecting putative orthologs, *Bioinformatics*, 19:1710–1711, 2003.

[22] Westover, B. P., Buhler, J. D., Sonnenburg, J. L., and Gordon, J. I., Operon prediction without a training set, *Bioinformatics*, 21:880–888, 2005.

[23] Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y., Prediction of functional modules based on comparative genome analysis and Gene Ontology application, *Nucleic Acids Res.*, 33:2822–2837, 2005.

[24] Zhou, D. K. T. J., Xu, Y. and Tiedje, J. M., *Microbial Functional Genomics*, John Wiley and Sons, 2004.