BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 13 2008, pages 1523–1529
doi:10.1093/bioinformatics/btn210

Data and text mining

Microbial genotype–phenotype mapping by class association rule mining

Makio Tamura*,† and Patrik D'haeseleer

Lawrence Livermore National Laboratory, Computing Applications and Research Department/Chemistry, Materials, Earth and Life Sciences Department, Microbial Systems Biology Group, Livermore, CA 94550, USA

Received on October 8, 2007; revised on February 28, 2008; accepted on April 26, 2008

Advance Access publication May 8, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Microbial phenotypes are typically due to the concerted action of multiple gene functions, yet the presence of each gene may have only a weak correlation with the observed phenotype. Hence, it may be more appropriate to examine co-occurrence between sets of genes and a phenotype (multiple-to-one) instead of pairwise relations between a single gene and the phenotype. Here, we propose an efficient class association rule mining algorithm, NETCAR, in order to extract sets of COGs (clusters of orthologous groups of proteins) associated with a phenotype from COG phylogenetic profiles and a phenotype profile. NETCAR takes into account the phylogenetic co-occurrence graph between COGs to restrict hypothesis space, and uses mutual information to evaluate the biconditional relation.

Results: We examined the mining capability of pairwise and multiple-to-one association by using NETCAR to extract COGs relevant to six microbial phenotypes (aerobic, anaerobic, facultative, endospore, motility and Gram negative) from 11 969 unique COG profiles across 155 prokaryotic organisms. With the same level of false discovery rate, multiple-to-one association can extract about 10 times more relevant COGs than one-to-one association. We also reveal various topologies of association networks among COGs (modules) from extracted multiple-to-one correlation rules relevant with the six phenotypes; including a well-connected network for motility, a star-shaped network for aerobic and intermediate topologies for the other phenotypes. NETCAR outperforms a standard CAR mining algorithm, CARAPRIORI, while requiring several orders of magnitude less computational time for extracting 3-COG sets.

Availability: Source code of the Java implementation is available as Supplementary Material at the Bioinformatics online website, or upon request to the author.

Contact: makio323@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The causal relationship between microbial genotype and phenotype can be extrapolated from co-occurrence of genes and phenotypes across a wide range of genomes. A phylogenetic profile (Eisen, 1998; Pellegrini *et al.*, 1999) is a vector encoding the presence and absence of a gene across sequenced genomes. We can likewise construct a phenotype profile (Jim *et al.*, 2004) (Table 1), indicating as to which organisms exhibit the phenotypic trait. Systematic comparison (Goh *et al.*, 2006; Jim *et al.*, 2004; Korbel *et al.*, 2005) can provide us with genotype–phenotype relationships and clues to understand the underlying biological mechanisms. Slonim *et al.* (2006) proposed a method to extract preferentially co-inherited generic *modules*, clusters of genes that have significant pairwise association with the phenotype observation.

Clusters of orthologous groups (COGs) of proteins (Tatusov *et al.*, 1997, 2003) provide a mapping between genes and their orthologs across sequenced genomes, and are an informative abstraction of genes for the construction of phylogenetic profiles. In this research, we compile 11 969 unique phylogenetic profiles of COGs for 155 prokaryotes from the STRING (Von Mering *et al.*, 2007) database. Individual COG phylogenetic profiles may only show a relatively weak correlation with a phenotype profile, even when the corresponding gene is essential for the phenotype. As an example, in Table 1, the profiles of COG_A COG_B, and COG_C have a weak pairwise relationship with the phenotype profile. However, when all of these COGs are present, the phenotype is always observed. Such an association between a set of genes to a phenotype (multiple-to-one) may suggest the importance of co-occurrence of these genes for the phenotype, potentially indicating an epistatic genetic interaction between them (Moore and Williams, 2005).

Class association rule (CAR) mining is a data mining technique to extract sets of items relevant with a class of interest. A standard CAR mining algorithm, CARAPRIORI (Agrawal and Srikant, 1994; Liu, 2006), finds if–then rules: *Set of items* ⇒ *class*; here, the rule may represent a hypothesis relating co-occurrence of a set of COGs and the presence of a phenotype. Bowers *et al.* (2004) suggested a method to derive more general logical rules by exhaustive enumeration, an approach that is computationally intractable for combinations of more than two COGs. Rule induction algorithms such as sequential covering by CN2 (Clark and Boswell, 1991) or simultaneous covering by decision tree algorithms (Quinlan, 1986) can also mine if–then rules, but they only discover small numbers of rules for efficient prediction or classification purposes, while CAR mining comprehensively searches for all rules satisfying some criterion. Here, we present a new CAR mining algorithm NETCAR to extract sets of COGs associated with a target phenotype.

*To whom correspondence should be addressed.

†Present address: Procter & Gamble Co., Miami Valley Innovation Center, Cincinnati, OH 45253-8707.

Table 1. An example phenotype profile and phylogenetic profiles for three COGs (COG_A–COG_C) across six organisms (O_1–O_6)

Organism	O_1	O_2	O_3	O_4	O_5	O_6
Phenotype	0	0	1	1	0	0
COG_A	1	1	1	1	0	0
COG_B	0	0	1	1	1	1
COG_C	1	1	1	1	1	0

2 PROBLEM

Conventionally, a CAR mining algorithm uses *Confidence* and *Support* where *Confidence* is the conditional probability of observation of the class (phenotype) given the set of items (COGs), and *Support* is the fraction of samples (genomes) in which the rule is valid in the data. When we have a rule such as COG_A *and* $COG_B \Rightarrow phenotype_C$ with 100% *Confidence*, $phenotype_C$ is observed whenever COG_A and COB_B are present in a genome. However, this rule may not have much biological relevance unless the converse relation holds as well, i.e. whenever $phenotype_C$ is observed, COG_A and COG_B tend to be present. In order to measure how well the set of COGs approximate a necessary and sufficient condition for the presence of the phenotype: *Set of COGs* $\Leftrightarrow$ *phenotype*, we use mutual information (*MI*) (Cover and Thomas, 1991) to evaluate the association between the COG combination profile (see Section 4) and phenotype profile (Slonim *et al.*, 2006).

The main problem with CAR mining is how to narrow down the hypothesis space. The space of all possible sets of s COGs is $O(m^s)$ where m is the total number of COGs, hence brute-force search becomes intractable for large m or s. In our case, there are more than 285 billion possible 3-COG sets out of 11 969 unique COG profiles, so we need to use heuristics to focus on a subset of the most promising candidate sets. CARAPRIORI uses the downward closure property based on minimum *Support* (Agrawal and Srikant, 1994), and it guarantees to exhaust all candidate sets that satisfy the minimum *Support*. However, when we use small minimum *Support* to avoid missing interesting rules, the size of hypothesis space can become exponential in the size of the set, making it very expensive to extract sets of 3 or more COGs by CARAPRIORI. What is worse, our dataset has many more items (COGs) than samples (genomes), contrary to the type of market basket data for which CARAPIORI was originally developed, and CARAPRIORI may tend to generate many irrelevant rules (Liu *et al.*, 2006).

Instead of the exhaustive selection, we propose a new algorithm to narrow down hypothesis space restricted by a COG connectivity graph (Butte and Kohane, 2000; Klus *et al.*, 2001; Moriyama *et al.*, 2003) where each node is a COG and edges are assigned between COGs for which the mutual information between their phylogenetic profiles exceeds a certain threshold value. This has a 2-fold benefit: to reduce computation time, and to generate rules consisting of genes which are more likely to have some biologically relevant functional association with each other. COGs which have similar phylogenetic profile are likely to be functionally associated (Overbeek *et al.*, 1999; Pellegrini *et al.*, 1999). Supplementary Figure S1 in supporting information shows the distribution of mutual information between pairs of phylogenetic profiles of COGs corresponding to enzymes that are connected via a compound in a KEGG pathway map (Kanehisa and Goto, 2000; Ogata *et al.*, 1999), compared to those of randomly selected COG pairs. The figure illustrates that a pair of COGs with a high mutual information is likely to have some functional connection (Von Mering *et al.*, 2003).

3 ALGORITHM

The basic algorithm is as follows: (1) select *Parent* COGs whose profile shows strong pairwise association with a phenotype profile of interest and *Child* COGs that are no more than $s-1$ steps away from a *Parent* on the COG connectivity graph (where s is the size of the rules we want to construct); (2) generate candidate COG sets containing at least one *Parent*, which form a connected subgraph on the COG connectivity graph; and (3) evaluate mutual information between the combined phylogenetic profile of each set with the phenotype profile.

Algorithm 1 shows pseudocode of an implementation of the NETCAR algorithm. It takes a profile matrix $\boldsymbol{M}$, a phenotype profile ***phe***, the size of the set in a rule s and three mutual information thresholds, *mMIp*, *mMIc* and *mMIr*. GETCONNECTIVITYGRAPH at line 11 constructs a connectivity graph $\boldsymbol{G}$ of COGs, where an edge is assigned if the mutual information between two COG profiles exceeds the threshold value *mMIc*. SELECTPARENT at line 12 returns a list of indexes of *Parents*, ***pa***, whose mutual information with the phenotype ***phe*** exceeds the user-defined threshold value *mMIp*. SELECTCHILD at line 13 returns a list of *Children*, ***ch***, which are within $s-1$ steps from a *Parent* on the connectivity graph $\boldsymbol{G}$. ***pa*** and ***ch*** are combined into one target COG index array $\boldsymbol{t}$: $p_1,p_2,\dots,p_{size(pa)},c_1,c_2,\dots,c_{size(ch)}$ (at line 14). For sets up to size 4, we can check that the COG set forms a connected subgraph by requiring that the sum of pairwise distances between COGs be smaller than or equal to *maxPL* (line 15), the sum of distances for a linear s-node path (for $s \geq 5$, this heuristic may yield some unconnected subgraphs). The all-to-all shortest distance matrix $\boldsymbol{D}$ is precomputed by the FLOYD-WARSHALL algorithm at line 16. Starting from a single *Parent*, successively larger sets up to size s are generated by adding additional *Parents* or *Children*, inside the while loop (line 23 onwards). Intermediate sets that would exceed a valid sum of distances, *maxPL*, are pruned early (line 28); otherwise, they are pushed back onto the stack for further expansion. When the size of a set reaches s (at line 32), a combined vector $\boldsymbol{v}$ is constructed by taking the intersection (AND rule) of the phylogenetic profiles of all the COGs in the set (line 33), and the mutual information between $\boldsymbol{v}$ and the phenotype profile ***phe*** is computed. If the value is larger than the user-defined *mMIr*, then the set is added to the collection of rules ***rules*** (lines 36–39).

4 MATERIALS AND METHODS

4.1 Data

We constructed the phylogenetic profile matrix $\boldsymbol{M}$ from the STRING version 7.0 database (Von Mering *et al.*, 2007), an extension of the original COG database (Tatusov *et al.*, 2003). The mapping table contains presence/absence of 4873 COGs, plus an extended set of 33 858 non-supervised orthologous groups (NOGs) to cover genes that are not included

Algorithm 1 Pseudocode of a NETCAR implementation

```
Input:
$\boldsymbol{M} \in [0,1]^{m\times n}$                    ▷ Phylogenetic profile matrix
$\boldsymbol{phe} \in [0,1]^{1\times n}$                   ▷ Phenotype profile
$s$                                       ▷ size of COG set in a rule
$mMI_r$                                   ▷ minimum MI to select rules
$mMI_p$                                   ▷ minimum MI to select parent COGs
$mMI_c$                                   ▷ minimum MI to select child COGs
Output:
$\boldsymbol{rules}$                                   ▷ array of association rules

procedure NETCAR($\boldsymbol{M}$, $\boldsymbol{phe}$, $s$, $mMI_r$, $mMIr_p$, $mMI_c$)
    $\boldsymbol{G} \leftarrow$ GETCONNECTIVITYGRAPH($\boldsymbol{M}$, $mMI_c$)
    $\boldsymbol{pa} \leftarrow$ SELECTPARENT($\boldsymbol{M}$, $\boldsymbol{phe}$, $mMI_p$)
    $\boldsymbol{ch} \leftarrow$ SELECTCHILD($\boldsymbol{G}$, $s$, $\boldsymbol{pa}$)
    $\boldsymbol{t} \leftarrow \boldsymbol{pa} \bigcup \boldsymbol{ch}$
    $maxPL \leftarrow$ GETMAXSUMALLTOALLPATHLENGTH($s$)
    $\boldsymbol{D} \leftarrow$ SHORTESTDISTANCE($\boldsymbol{G}$, $\boldsymbol{t}$)
    $\boldsymbol{S}$                                   ▷ Stack
    for $i \leftarrow 1$, SIZE($\boldsymbol{pa}$) do
        $\boldsymbol{c} \leftarrow \boldsymbol{pa}[i]$                    ▷ item combination
        $\boldsymbol{S}$::PUSH($\boldsymbol{c}$)
    end for

    while $\boldsymbol{S}$ is not empty do
        $\boldsymbol{c} \leftarrow \boldsymbol{S}$::POP
        if SIZE($\boldsymbol{c}$) $<$ $s$ then
            for $j \leftarrow i+1$, SIZE($\boldsymbol{t}$) do
                $\boldsymbol{c_{next}} \leftarrow \boldsymbol{c} \bigcup \boldsymbol{t}[j]$
                if CHECK($\boldsymbol{c_{next}}$, $\boldsymbol{D}$, $maxPL$) then
                    $\boldsymbol{S}$::PUSH($\boldsymbol{c_{next}}$)
                end if
            end for
        else if SIZE($\boldsymbol{c}$) equal $s$ then
            $\boldsymbol{v} \leftarrow$ GETJOINTARRAY($\boldsymbol{c}$)
            if CHECK($\boldsymbol{c}$, $\boldsymbol{D}$, $maxPL$) then
                $MI \leftarrow$ GETMI($\boldsymbol{v}$, $\boldsymbol{phe}$)
                if $MI \geq mMI_r$ then
                    $\boldsymbol{r}$::$\boldsymbol{comb} \leftarrow \boldsymbol{c}$        ▷ rule class instance
                    $\boldsymbol{r}$::$MI \leftarrow MI$
                    $\boldsymbol{rules} \leftarrow \boldsymbol{r} \bigcup \boldsymbol{rules}$
                end if
            end if
        end if
    end while
    return $\boldsymbol{rules}$
end procedure
```

in the original COG database. Here we use only the 155 representative core prokaryotic organisms (out of 337 prokaryotes in STRING 7.0), in order to mitigate the sequencing bias among lineages. There are only 11 969 unique phylogenetic profiles out of original 38 731 COG and NOG profiles (for the remainder, we ignore the distinction between COG and NOGs). We used the dataset generated by Slonim *et al.* (2006) for six binary phenotype profiles, combined with additional data from NCBI's GenomeProject database, the Joint Genome Institute's Integrated Microbial Genomics (IMG) system (Markowitz *et al.*, 2006), the Genomes Online Database (Kyrpides, 1999) and literature. There are 62 aerobic, 31 anaerobic, 42 facultative, 11 endospore forming, 76 motile and 95 Gram-negative organisms in our dataset. The enzyme connections used in the COG pair analysis is extracted from the KEGG LIGAND database as of July 9, 2007.

4.2 Phylogenetic profile and phenotype profile

The phylogenetic profiles is represented by a binary matrix $\boldsymbol{M} \in [0,1]\ m\times n$, and a phenotype profile is represented by a binary vector $\boldsymbol{phe} \in [0,1]\ 1\times n$ where m is the total number of COGs and n is the number of genomes. The i-th row vector of $\boldsymbol{M}$ represents the phylogenetic profile of the i-th COG, where the j-th element indicates presence or absence of this COG in the j-th genome. Likewise, the j-th element of $\boldsymbol{phe}$ shows presence or absence of a phenotype in the j-th organism.

Multiple phylogenetic profiles are combined into a single vector $\boldsymbol{v} \in [0,1]$ $1\times n$ where the j-th element is equal to 1 only if the j-th elements of all the COG profiles are equal to 1 (i.e. the combined profile vector is the AND function of the individual COG phylogenetic profiles).

4.3 Mutual information

Mutual information MI(Cover and Thomas, 1991) between two binary vectors $\boldsymbol{u},\boldsymbol{v} \in [0,1]^{1\times n}$ is calculated as follows; $MI(\boldsymbol{u};\boldsymbol{v}) = \sum_{y\in[0,1]}\sum_{x\in[0,1]} P(x,y)\log P(x,y)/P(x)P(y)$ where x and y are the values of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively, and $P()$ is the probability function.

4.4 Connectivity graph for a phylogenetic profile

For the phylogenetic profiles $\boldsymbol{M} \in [0,1]\ ^{m\times n}$ where m is the total number of COGs and n is the number of genomes, the i-th row vector $\boldsymbol{m}_i$ represents the phylogenetic profile of the i-th COG. We can construct an adjacency matrix of the connectivity graph $\boldsymbol{G}=(g_{i,j} \in [0,1])_{i,j}^{m,m}$ as follows: $g_{i,j}=1$ if $MI_{i,j} \geq mMI_c$ or otherwise 0, where $MI_{i,j}$ is mutual information between $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$, and mMI_c is a threshold value.

4.5 FDR

We used a random permutation method proposed by Zhang and Padmanabhan (2004) to measure an FDR. If N_{o} and N_{r} are the number of rules that have a mutual information MI or higher with respect to the original and randomly permutated phenotype profile, respectively, then $N_{\mathrm{o}}/N_{\mathrm{r}}$ is a simple estimated positive FDR (Storey and Tibshirani, 2003) for the given mutual information MI. We calculated the median value of $N_{\mathrm{o}}/N_{\mathrm{r}}$ from 200 random permutation experiments. We can scan all pairwise associations, but it takes too much time to scan all 2-COG and 3-COG associations. Therefore, we randomly selected as many 2- and 3-COG sets from the COG phylogenetic profile as possible within 15 min computational time; 3.6×10^5 (0.5% of all possible 2-COG combinations) and 5.7×10^5 (0.0002% of all 3-COG combinations), respectively. This process is repeated 500 times, and FDR for the mutual information MI is calculated as an average of the median value. Supplementary Figure S11 of supporting information shows the relationships between mutual information and FDR for pairwise, 2-COG and 3-COG rules for all six phenotypes.

4.6 Experimental parameters

Experiments were performed under version 1.6 Java runtime environment on a 64 bit Linux machine with 7.0 Gb memory and 3.00 GHz CPU power. For rule mining by NETCAR, the mutual information to select *Parent* COGs, *mMIp*, is adjusted to FDR level of 0.1% The threshold mutual information, *mMIc*, is set to 1.1 × average mutual information among *Parents* for 2- and 3-COG rule mining.

5 RESULTS AND DISCUSSION

5.1 Controlling for multiple hypothesis testing using false discovery rate

Because of the large number of rules evaluated by CAR mining; we controlled the statistical significance of the resulting rules by false discovery rate (FDR) (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003; Zhang and Padmanabhan, 2004). FDR is the expected proportion of true null hypotheses among all rejected hypotheses in a multiple hypotheses evaluation, i.e. the expected proportion of false positives among our results, and we can compare the efficiency of different approaches by the number of relevant rules and COGs they are able to extract at the same FDR level. Previous work on pairwise genotype–phenotype associations (Goh *et al.*, 2006; Jim *et al.*, 2004) estimated significance based on a conservative Bonferroni correction on *P*-values derived from a hypergeometric distribution. A traditional family-wise error rate (FWER) (Shaffer, 1995) approach such as Bonferroni essentially calculates the probability of making even a single error (or more) among all of the evaluated rules under the null hypothesis. In cases where a very large number of hypotheses are to be tested, and multiple positive results are expected, FWER is far too conservative, and correcting for FDR is more appropriate. For this same reason, FDR is also the method of choice for extracting multiple significantly expressed genes out of the thousands represented on a gene expression microarray (Pounds, 2006; Storey and Tibshirani, 2003). We adjusted threshold mutual information by FDR level for each rule mining, and compare the pairwise and multiple-to-one association rule mining by using NETCAR. The detailed experimental procedure to measure FDR is given earlier in Section 4.

For 2- and 3-COG rule generation from our dataset, NETCAR starts with *Parents* of 0.1% FDR level. In the case of 3-COG rule mining, the NETCAR algorithm eventually evaluates about 10^5–10^9 candidate sets when we set the final rule threshold mutual information (*mMIr*) of 0.1% FDR level. Overall, NETCAR takes less than 5 min to mine 3-COG rules for one phenotype, the majority of which is taken up by precomputing the COG connectivity graph and distance matrix. In comparison, it takes at least 10 h to explore all 3-COG sets with 0.1% FDR level of *Parents* and the same size of *Children*. Hence, the selection strategy using the connectivity graph significantly reduces the amount of computational time. We estimate that it would take more than 1 month to mine 3-COG rules by CARAPRIORI to extract rules with the same level of FDR.

5.2 Multiple-to-one association rules and genetic module

Figure 1a shows the number of unique COGs found in pairwise, 2-COG and 3-COG association rules within various FDR levels where broken line corresponds to the number of uncharacterized COGs. The number of unique COGs is roughly linear to the number of extracted rules (Supplementary Fig. S2) and up to around 15% are uncharacterized COGs by any associations. With the same FDR, the multiple-to-one association reveals substantially larger number of relevant COGs than pairwise association does, except for motility. Figure 1b shows the number of COGs that are extracted by multiple-to-one association with FDR of 0.1% , but are not covered by pairwise association with much relaxed FDR level. Except for motility and endospore, the pairwise method fails to capture a large number of COGs, including many uncharacterized ones, even when the FDR level is relaxed up to 10%. We find 38 uncharacterized COGs that are mined in 3-COG association rules for the aerobic phenotype with a FDR level <0.1%, but are not covered by pairwise association with very relaxed FDR level of 10%. With the same condition, we also find 52, 55, 2, 52 uncharacterized COGs for anaerobic, facultative, endospore and Gram-negativity phenotype.

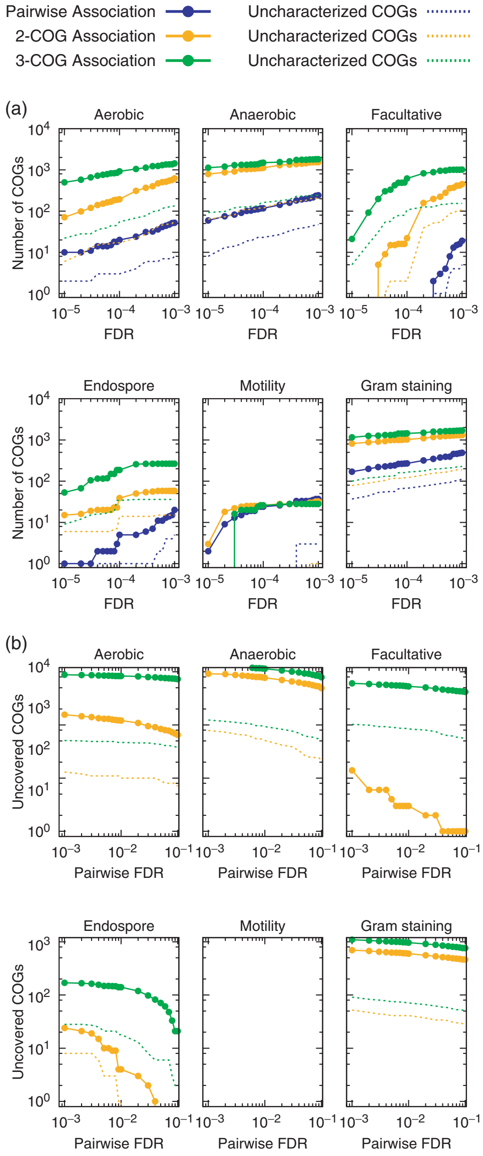

Fig. 1. (**a**) Number of unique COGs in extracted rule within FDR levels. Blue, orange and green lines are 1-COG, 2-COG and 3-COG rules, respectively, and broken lines are number of uncharacterized COGs in the same colored association rules. (**b**) Number of unique COGs that are in 2- and 3-COG association rules with FDR of 1.0×10^{-4} but are not in pairwise association rules with much relaxed FDR level ranging between 1.0×10^{-3} and 1.0×10^{-1}. (Values for motility are 0, hence the missing curves for that phenotype.)

A list of these COGs in 3-COG association rules, but are not covered by pairwise association with relaxed FDR level of 10%, are also available in supporting information.

We compile association networks from 3-COG positively associated rules with a stringent FDR level (Fig. 2), indicating which COGs occur in the top rules, how frequently and in which combination. The list of these COGs are available in supporting information. We observed two distinctive types of association network topology: (1) *Clique* type networks, such as the motility module, where genes are well associated with each other and (2) *Star* type networks, typified by the aerobic module, where a few COGs have links to many other COGs in a star-like topology. It is interesting to note that even with a relaxed FDR, there are many green nodes that have a weak association with the target phenotype, with the exception of the motility phenotype, for which pairwise association may be appropriate to extract associated COGs. What is more, pairs of COGs with similar profiles (darker blue edge) do not always form the most frequently observed COG combinations (wider edges) in these networks. This type of COG association can be considered as a functional *module*, even though the individual phylogenetic profiles may be quite divergent. This concept is complementary to the concept of a *module* used by Slonim *et al.* (2006), in which each COG profile has a significant pairwise correlation to a phenotype profile, and profiles in the same module are similar to each other. Indeed, the well-connected components in the *Clique*-type networks are similar to the *modules* defined by Slonim *et al.*; however, for the other phenotypes, the relevant COGs do not always have a strong pairwise association, but the module can be understood by a *Star* or mixed-type network. Previous work (Ravasz *et al.*, 2002) suggests that the metabolic network is more consistent with a fractal structure model, a mixture of a scale-free (Jeong *et al.*, 2000) and modular network. The pairwise method may have an intrinsic problem discovering relevant genes in such a structure. In contrast, the multiple-to-one method can explore the associated genes with weak pairwise associations. The biological reason behind the presence of individual COGs in the rule will require further investigation, but a brief overview of the resulting networks is given subsequently.

5.3 Genetic module and COG functions

5.3.1 Aerobic Both the aerobic and anaerobic networks contain redox-active proteins. Mutual information captures both positive and negative correlations. The aerobic phenotype is the only one we tested for which the negatively correlated rules are dominant. We found many oxygen-sensitive enzymes in these negatively associated rules, while enzymes that detoxify active oxygen derivatives occur in positively associated rules. The capability to extract both positive and negative association is one of the advantages of CARApriori, compared to the standard CAR mining algorithm, CARApriori, which can only extract positively correlated rules. Oxidase complex and dehydrogenase enzymes are frequent strong pairwise association COGs (orange node) while Dinucleotide (FAD)/ flavin mononucleotide (FMN)-containing dehydrogenase, oxidoreductases, the pyruvate dehydrogenase complex, catalyzing oxidative decarboxylation of pyruvate to form acetyl-CoA, and peroxiredoxin, a family of multifunctional antioxidant enzymes, are dominant COGs with a weak pairwise association (green node) in the aerobic network. Citrate synthase, one of the major green nodes, catalyzes the reaction to produce citrate from acetyl-CoA and oxyalacetate at the first step of the citric acid cycle in aerobic respiration.

5.3.2 Anaerobic Oxygen toxicity in anaerobes was thought to be due to the absence of superoxide dismutase, although recent genomic studies suggest that this may not be the discriminative factor, and that the mechanisms behind microbial sensitivity to oxygen may as yet be unknown (Madigan *et al.*, 2000). Oxygen-sensitive and redox-related proteins, such as 6Fe–6S cluster proteins, activator of 2-hydroxyglutaryl-CoA dehydratase, cobalamin biosynthesis protein CbiD or pyruvate-formate lyase-activating enzyme, etc., including many weak pairwise association proteins, appear in the anaerobic network, but many of these proteins may be also found in aerobes.

5.3.3 Facultative Enzymes in the fermentation pathway such as the phosphotransferase system IIA, IIB and IIC components, form a connected component in the facultative network, in combination with COG3091, an uncharacterized protein conserved primarily within the Bacilli. A second major uncharacterized protein, COG3094, is conserved mainly within the betaproteobacteria, and is associated with several other membrane-associated COGs. Facultative microbes may contain both genes of aerobic respiration and genes for fermentation or anaerobic respiration, as well as genes for detoxifying active oxygen derivatives.

5.3.4 Endospore A regulator of polyketide synthase expression, an AT-rich DNA-binding protein, cell division initiation protein, transcription regulator and uncharacterized COG1799 and COG1302 are the main COGs. It has been reported that more than 200 genes may be involved in the endospore formation process (Madigan *et al.*, 2000), but the large population of weak pairwise association COGs in the extracted rules suggest that many of endospore-relevant proteins may be used in other biological functions as well.

5.3.5 Motility Flagellar apparatus proteins form a well-connected graph. This network forms a genetic module with a clear boundary, in which each gene has a strong pairwise association with the phenotype observation. Therefore, the number of extracted COGs both in the pairwise and multiple-to-one association are the same, and its upper bound is well limited within a certain number, about 20–30 COGs for the motility phenotype. However, only motility phenotype has this shape in our six phenotypes. Figure 2 only shows flagellar apparatus COGs. Another well-connected network is formed by chemotaxis-related COGs.

5.3.6 Gram negative The biological relationship of Gram negativity with cell wall structure is well understood; a membrane with a thick peptidoglycan layer stains Gram positive while organisms with a periplasm with thin peptidoglycan and outer membrane stain Gram negative. Indeed, outer-membrane proteins and various transport system proteins that may be used for the membrane proteins (Jedrzejas and Huang, 2003) form a graph in the Gram-negative network. However, both pairwise and multiple-to-one association extracted relatively large number of COGs with Gram-negative bacteria, but may not relevant with the Gram-stain mechanism.

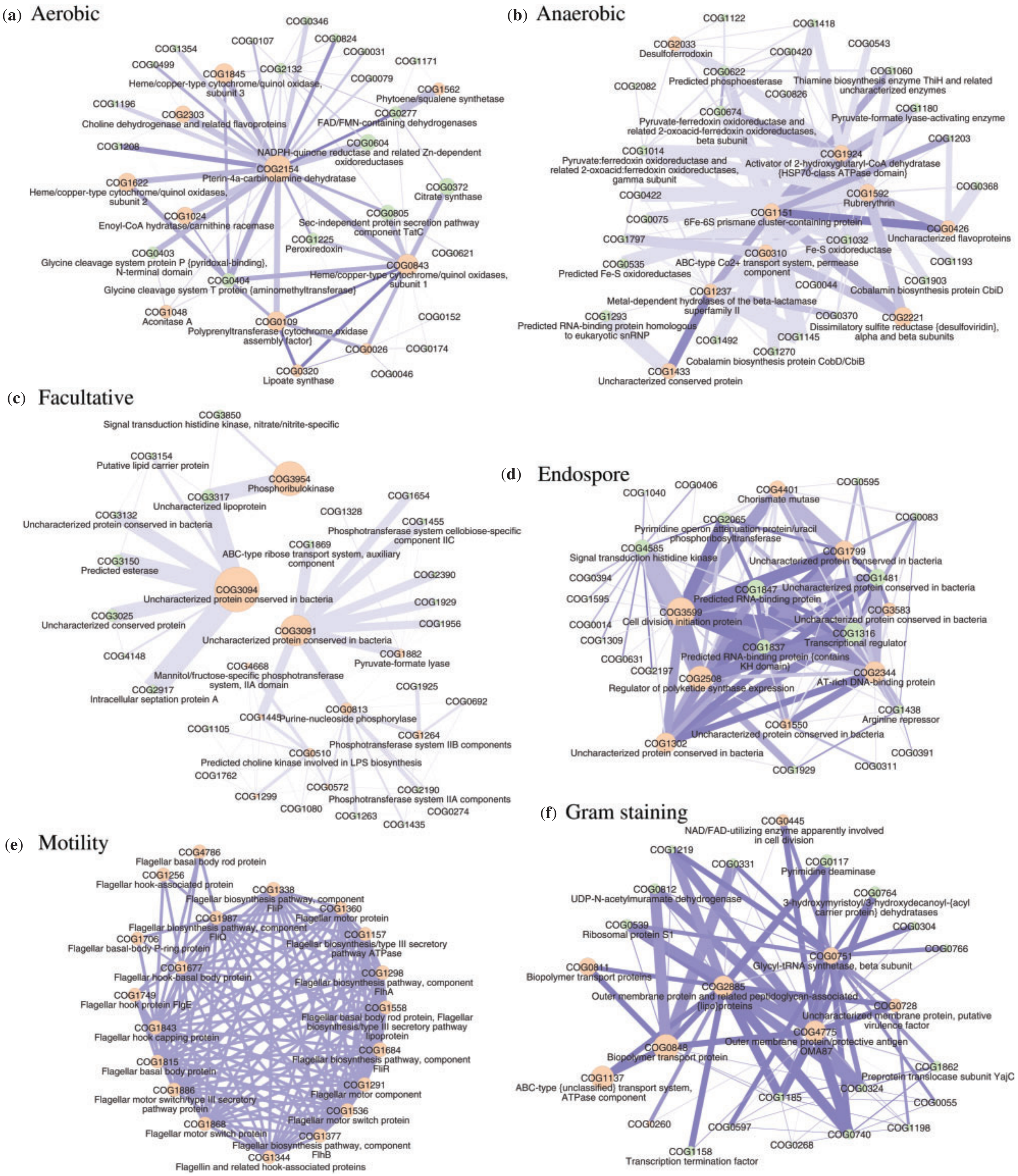
(a) Aerobic
COG0346
COG0824
COG0107
COG1354
COG0031
COG1171
COG0499
COG1845
COG2132
COG0079
Heme/copper-type cytochrome/quinol oxidase, subunit 3
COG1562
Phytoene/squalene synthetase
COG1196
COG2303
COG0277
Choline dehydrogenase and related flavoproteins
FAD/FMN-containing dehydrogenases
COG0604
COG1208
NADPH-quinone reductase and related Zn-dependent oxidoreductases
COG2154
Pterin-4a-carbinolamine dehydratase
COG0372
COG1622
Citrate synthase
Heme/copper-type cytochrome/quinol oxidases, subunit 2
COG1024
COG0805
Enoyl-CoA hydratase/carnithine racemase
Sec-independent protein secretion pathway component TatC
COG1225
Peroxiredoxin
COG0403
COG0621
Glycine cleavage system protein P (pyridoxal-binding), N-terminal domain
COG0843
Heme/copper-type cytochrome/quinol oxidases, subunit 1
COG0404
Glycine cleavage system T protein (aminomethyltransferase)
COG1048
Aconitase A
COG0152
COG0109
Polyprenyltransferase (cytochrome oxidase assembly factor)
COG0026
COG0174
COG0320
Lipoate synthase
COG0046
(b) Anaerobic
COG1122
COG1418
COG2033
Desulfoferrodoxin
COG0420
COG0543
COG0622
COG1060
COG2082
Predicted phosphoesterase
Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes
COG0826
COG0674
COG1180
Pyruvate-ferredoxin oxidoreductase and related 2-oxoacid-ferredoxin oxidoreductases, beta subunit
Pyruvate-formate lyase-activating enzyme
COG1203
COG1014
COG1924
Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, gamma subunit
Activator of 2-hydroxyglutaryl-CoA dehydratase (HSP70-class ATPase domain)
COG0368
COG0422
COG1592
Rubrerythrin
COG1151
COG0075
6Fe-6S prismane cluster-containing protein
COG0426
COG1797
COG1032
Uncharacterized flavoproteins
Fe-S oxidoreductase
COG0310
COG1193
COG0535
ABC-type Co2+ transport system, permease component
Predicted Fe-S oxidoreductases
COG1903
COG0044
Cobalamin biosynthesis protein CbiD
COG1237
Metal-dependent hydrolases of the beta-lactamase superfamily II
COG1293
COG0370
COG2221
Predicted RNA-binding protein homologous to eukaryotic snRNP
Dissimilatory sulfite reductase (desulfoviridin), alpha and beta subunits
COG1492
COG1145
COG1270
Cobalamin biosynthesis protein CbiD/CbiB
COG1433
Uncharacterized conserved protein
(c) Facultative
COG3850
Signal transduction histidine kinase, nitrate/nitrite-specific
COG3154
Putative lipid carrier protein
COG3954
Phosphoribulokinase
COG3317
COG1654
Uncharacterized lipoprotein
COG3132
COG1328
Uncharacterized protein conserved in bacteria
COG1455
Phosphotransferase system cellobiose-specific component IIC
COG1869
ABC-type ribose transport system, auxiliary component
COG3150
Predicted esterase
COG2390
COG3094
Uncharacterized protein conserved in bacteria
COG1929
COG3025
COG1956
Uncharacterized conserved protein
COG3091
Uncharacterized protein conserved in bacteria
COG4148
COG1882
Pyruvate-formate lyase
COG4668
Mannitol/fructose-specific phosphotransferase system, IIA domain
COG1925
COG2917
Intracellular septation protein A
COG0692
COG0813
COG1445
Purine-nucleoside phosphorylase
COG1105
COG1264
Phosphotransferase system IIB components
COG0510
Predicted choline kinase involved in LPS biosynthesis
COG1762
COG0572
COG2190
COG1299
Phosphotransferase system IIA components
COG1080
COG1263
COG0274
COG1435
(d) Endospore
COG4401
COG0595
COG1040
COG0406
Chorismate mutase
COG2065
COG0083
Pyrimidine operon attenuation protein/uracil phosphoribosyltransferase
COG4585
COG1799
Signal transduction histidine kinase
Uncharacterized protein conserved in bacteria
COG0394
COG1481
COG1847
Uncharacterized protein conserved in bacteria
COG1595
Predicted RNA-binding protein
COG3583
COG3599
Uncharacterized protein conserved in bacteria
COG0014
Cell division initiation protein
COG1316
COG1309
Transcriptional regulator
COG1837
COG0631
Predicted RNA-binding protein (contains KH domain)
COG2197
COG2344
COG2508
AT-rich DNA-binding protein
Regulator of polyketide synthase expression
COG1438
Arginine repressor
COG1550
Uncharacterized protein conserved in bacteria
COG1302
COG0391
Uncharacterized protein conserved in bacteria
COG0311
COG1929
(e) Motility
COG4786
Flagellar basal body rod protein
COG1256
Flagellar hook-associated protein
COG1338
Flagellar biosynthesis pathway, component FliP
COG1360
COG1987
Flagellar motor protein
Flagellar biosynthesis pathway, component FliQ
COG1157
COG1706
Flagellar biosynthesis/type III secretory pathway ATPase
Flagellar basal-body P-ring protein
COG1677
COG1298
Flagellar hook-basal body protein
Flagellar biosynthesis pathway, component FlhA
COG1749
COG1558
Flagellar hook protein FlgE
Flagellar basal body rod protein, Flagellar biosynthesis/type III secretory pathway lipoprotein
COG1843
Flagellar hook capping protein
COG1684
Flagellar biosynthesis pathway, component FliR
COG1815
Flagellar basal body protein
COG1291
COG1886
Flagellar motor component
Flagellar motor switch/type III secretory pathway protein
COG1536
Flagellar motor switch protein
COG1868
Flagellar motor switch protein
COG1377
Flagellar biosynthesis pathway, component FlhB
COG1344
Flagellin and related hook-associated proteins
(f) Gram staining
COG0445
NAD/FAD-utilizing enzyme apparently involved in cell division
COG1219
COG0331
COG0117
Pyrimidine deaminase
COG0812
COG0764
UDP-N-acetylmuramate dehydrogenase
3-hydroxymyristoyl/3-hydroxydecanoyl-(acyl carrier protein) dehydratases
COG0539
COG0304
Ribosomal protein S1
COG0751
COG0766
Glycyl-tRNA synthetase, beta subunit
COG0811
Biopolymer transport proteins
COG2885
Outer membrane protein and related peptidoglycan-associated (lipo)proteins
COG0728
Uncharacterized membrane protein, putative virulence factor
COG4775
Outer membrane protein/protective antigen OMA87
COG0848
Biopolymer transport protein
COG1862
Preprotein translocase subunit YajC
COG1137
ABC-type (unclassified) transport system, ATPase component
COG0324
COG1185
COG0055
COG0260
COG0597
COG1198
COG0740
COG0268
COG1158
Transcription termination factor

Fig. 2. (**a–f**) COG association graphs for the six phenotypes. The nodes are COGs involved in the rules within FDR level of 1.5×10^{-5}, 5.0×10^{-8}, 1.5×10^{-5}, 1.0×10^{-5}, 5.0×10^{-5} and 5.0×10^{-14} for aerobic, anaerobic, facultative, endospore, motility and Gram negativity phenotypes, and edges show that the linked COGs are used in the same rule. The orange nodes are COG covered by pairwise association with 100 times relaxed FDR except for anaerobic and Gram negativity with FDR of 0.01, while the green nodes represent the other COGs with a weaker pairwise correlation. The size of each node and the width of each edge are proportional to the frequencies of the corresponding COG and link in the extracted rules, respectively. Darker edges indicate a closer profile similarity between the linked COGs.

6 CONCLUSION

We developed a new class association rule mining algorithm, NETCAR that extracts multiple-to-one relationships between COGs and a phenotype of interest, from a COG phylogenetic and the phenotype profile. NETCAR is much more efficient than a standard CAR mining algorithm, CARAPRIORI in computational time. The multiple-to-one association rules with stringent FDR level for aerobic, anaerobic, facultative, endospore and Gram-negative phenotype contain significantly larger numbers of COGs than those by pairwise methods. We compiled association network from extracted 3-COG rules and revealed that the network cannot only have a *Clique* structure, as implicitly assumed by previous pairwise methods, but also a *Star*-type topology that contains large number of COGs whose occurrence is only weakly correlated with a phenotype observation. These results indicate that a gene module can be a combination of genes that span some depth in a biological network, from a layer where we can see strong pairwise association. The NETCAR algorithm is a powerful CAR mining algorithm that can be used to extract relevant genes (COGs) associated with a phenotype observation, that cannot be elucidated by simple pairwise comparisons. We also discuss the phenotype prediction capability of extracted rules in the supporting material. It is often the case that the dimensionality of large-scale biological data (in our case, number of COGs) is much larger than the number of samples (genomes), and the NETCAR algorithm may be appropriate to extract associations from other such data types. For example, NETCAR may be able to mine co-regulatory gene network modules relevant with a target physiological observation, from microarray data with many more genes than expression arrays.

ACKNOWLEDGEMENTS

The authors would like to thank Bing Liu for many helpful discussions and Yiming Ma for allowing us to use their CBA system (Liu *et al.*, 1998). We also thank Eivind Almaas and Cheol-Min Ghim for many productive discussions about the network analysis, and Yongqin Jiao for valuable suggestions about the functional explanation of COGs.

Funding: This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory (LLNL) in part under Contract No. W-7405-Eng-48 and in part under Contract DE-AC52-07NA27344. The project (05-ERD-065) was funded by the Laboratory Directed Research and Development Program at LLNL.

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. and Srikant,R. (1994) Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487–499.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, **57**, 289–300.

Bowers,P. *et al.* (2004) Use of logic relationships to decipher protein network organization. *Science*, **306**, 2246–2249.

Butte,A. and Kohane,I. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **5**, 427–439.

Clark,P. and Boswell,R. (1991) Rule induction with CN2: some recent improvements. *Proceedings of the Fifth European Working Session on Learning*, **482**, 151–163.

Cover,T. and Thomas,J. (1991) *Elements of Information Theory*. Wiley, New York.

Eisen,J. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.

Goh,C.-S. *et al.* (2006) Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics*, **7**, 257–257.

Jedrzejas,M. and Huang,W. (2003) Bacillus species proteins involved in spore formation and degradation: from identification in the genome, to sequence analysis, and determination of function and structure. *Crit. Rev. Biochem. Mol. Biol.*, **38**,173–198.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Jim,K. *et al.* (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Res.*, **14**, 109–115.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Klus,G. *et al.* (2001) Mutual information analysis as a tool to assess the role of aneuploidy in the generation of cancer-associated differential gene expression patterns. *Pac. Symp. Biocomput.*, **42**, 51.

Korbel,J. *et al.* (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.*, **3**, 134–134.

Kyrpides,N. (1999) Genomes online database (Gold 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, **15**, 773–774.

Liu,B. (2006) *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*. Springer, Berlin, Heidelberg, New York.

Liu,B. *et al.* (1998) Integrating classification and association rule mining. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. ACM SIGKDD, New York, pp. 80–86.

Liu,B. *et al.* (2006) Rule interestingness analysis using OLAP operations. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and Data Mining*, pp. 297–306.

Madigan,M. *et al.* (2000) *Brock Biology of Microorganisms*. Prentice Hall, Upper Saddle River, NJ.

Markowitz,V. *et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.*, **34**, 344–348.

Moore,J. and Williams,S. (2005) Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*, **27**, 637–646.

Moriyama,M. *et al.* (2003) Relevance network between chemosensitivity and transcriptome in human hepatoma cells 1. *Mol. Cancer Ther.*, **2**, 199–205.

Ogata,H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.

Overbeek,R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Pounds,S. (2006) Estimation and control of multiple testing error rates for microarray studies. *Brief. Bioinform.*, **7**, 25–36.

Quinlan,J. (1986) Induction of decision trees. *Mach. Learn.*, **1**, 81–106.

Ravasz,E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Shaffer,J. (1995) Multiple hypothesis testing. *Ann. Rev. Psychol.*, **46**, 561–584.

Slonim,N. *et al.* (2006) Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Mol. Syst. Biol.*, **2**.

Storey,J. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440.

Tatusov,R. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov,R. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics [electronic resource]*, **4**, 41–41.

Von Mering,C. *et al.* (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.

Von Mering,C. *et al.* (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, 358–362.

Zhang,H. and Padmanabhan,B. (2004) Using randomization to determine a false discovery rate for rule discovery. *Proceedings of the Fourteenth Workshop On Information Technologies And Systems; December 11–12, 2004*. WITS, Washington, DC, pp. 140–145.