BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i222–i229
doi:10.1093/bioinformatics/btm222

Systematic discovery of functional modules and context-specific functional annotation of human genome

Yu Huang[1,†], Haifeng Li[1,†], Haiyan Hu[1], Xifeng Yan[2], Michael S. Waterman[1], Haiyan Huang[3] and Xianghong Jasmine Zhou[1,*]

[1]Molecuolar and Computational Biology, University of Southern California, Los Angeles and [2]IBM T. J. Watson Research Center, Hawthorne, NY and [3]Department of Statistics, University of California, Berkeley, CA, USA

ABSTRACT

Motivation: The rapid accumulation of microarray datasets provides unique opportunities to perform systematic functional characterization of the human genome. We designed a graph-based approach to integrate cross-platform microarray data, and extract recurrent expression patterns. A series of microarray datasets can be modeled as a series of co-expression networks, in which we search for frequently occurring network patterns. The integrative approach provides three major advantages over the commonly used microarray analysis methods: (1) enhance signal to noise separation (2) identify functionally related genes without co-expression and (3) provide a way to predict gene functions in a context-specific way.

Results: We integrate 65 human microarray datasets, comprising 1105 experiments and over 11 million expression measurements. We develop a data mining procedure based on frequent itemset mining and biclustering to systematically discover network patterns that recur in at least five datasets. This resulted in 143 401 potential functional modules. Subsequently, we design a network topology statistic based on graph random walk that effectively captures characteristics of a gene's local functional environment. Function annotations based on this statistic are then subject to the assessment using the random forest method, combining six other attributes of the network modules. We assign 1126 functions to 895 genes, 779 known and 116 unknown, with a validation accuracy of 70%. Among our assignments, 20% genes are assigned with multiple functions based on different network environments.

Availability: http://zhoulab.usc.edu/ContextAnnotation

Contact: xjzhou@usc.edu

1 INTRODUCTION

Systematic functional characterization of genes identified in the genome sequencing projects is urgently needed in the post-genomic era. The rapid increase in large-scale gene expression data provides us unique opportunities to meet this need. A commonly used approach is to cluster genes with similar expression patterns (Beer and Tavazoie, 2004; Gasch and Eisen, 2002; Tamayo *et al.*, 1999), and to predict functions of unknown genes based on their expression similarity to known genes (Gasch and Eisen, 2002; Niehrs and Pollet, 1999). However, there are two problems with such clustering approaches: (*problem 1*) *genes with similar expression profiles may not have the same function*: for example, an experimental condition may perturb multiple biological pathways simultaneously, such that genes from these different functional pathways may show similar and indistinguishable expression patterns; and moreover, experimental noise and outliers may lead to biased and erroneously high estimates of expression similarity. (*Problem 2*) *Genes with similar functions may not have similar expression profiles*: for example, measurements of expression similarity, e.g. Pearson's correlation or Euclidean distance, may not capture the relationship between two expression profiles due to time—shifts (Qian *et al.*, 2001); and genes may be regulated at levels other than transcription. Recently, we have proposed and validated two novel expression relationships, 'transitive expression similarity' (Zhou *et al.*, 2002) and 'second-order expression similarity' (Zhou *et al.*, 2005), which can be used to link functionally related genes without similar expression profiles. No doubt that many other types of expression relationships exist among functionally related genes—some may even be beyond our current knowledge. How to identify such unknown expression relationships *systematically* is one of the major aims of this article. In addition, we will address another important problem (*problem 3*) in functional annotation, which has so far received little attention: how to annotate gene functions in a context-specific manner? An increasing number of examples indicate that in higher organisms, functional plasticity may be the rule rather than the exception (Jeffery, 2003a, b). A gene may acquire different functions under different endogenous or exogenous conditions. However, current functional prediction approaches (Wu *et al.*, 2002; Zhu *et al.*, 2005) and genome databases [such as SGD (Wang *et al.*, 2005), Wormbase (Drysdale *et al.*, 2005) and Flybase (Aggarwal *et al.*, 2006)] all annotate gene functions without specifying the necessary context. Recently, Lussier *et al.* (2006) for the first time systematically addressed this problem by proposing a system, PhenoGO, which extracts phenotypic contextual information from published literatures for existing Gene Ontology functional annotations (Lussier *et al.*, 2006).

In this article, we aim to overcome the above discussed three issues by using information in multiple microarray datasets. We model each microarray dataset with a graph, where a vertex

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

represents a gene, and if two genes show high correlation in their expression profiles, we connect them with an edge. A series of microarray datasets can be modeled as a series of co-expression networks, in which we search for *frequently* occurring network patterns. Such a network pattern consists of gene sets that function as a unit under various conditions, and thus likely represent a functional module. Based on the recurrent network patterns, we perform functional annotation. This approach can address the three problems above: (1) *to separate true functional links from spurious co-expression links*. We suggest that a co-expression link *recurrent* in multiple microarray datasets is more likely to represent a true functional link. (2) *To identify functionally related genes without direct co-expression*. When we combine multiple expression networks, subtle signals may emerge that cannot be identified in any of the individual networks. Such signals include recurrent paths that may extend beyond simple co-expression clusters yet represent functional modules. If we only consider a single co-expression network, it is difficult to stratify functionally important paths from their complex network environment. However, if a path frequently occurs across multiple co-expression networks, it is easily differentiated from the background. (3) *To conditionally annotate gene functions*. Because a gene cannot exert its function by itself but instead does so by interaction with other genes, its functional switch is likely to be caused by or result from the alteration of its interaction partners. Put into a network perspective, a gene's function may be different if placed in different subnetworks. As different external or endogenous conditions result in different topologies of the co-expression network, we can relate a gene's function to the experimental conditions via its network environment, thus leading to the context-specific functional annotation.

In this study, we integrate 65 human microarray datasets, comprising 1105 experiments and over 11 million expression measurements. We develop a data mining procedure based on frequent itemset mining (FIM) and biclustering to extensively discover network patterns that recur in at least five datasets. This resulted in 143 401 potential functional modules. Subsequently, we design a network topology statistic based on graph random walk that effectively captures characteristics of a gene's local functional environment. Functional annotations based on this statistic are then assessed using the random forest method with six other attributes of the network modules. We assign 1126 functions to 895 genes, 779 known and 116 unknown, with a validation accuracy of 70%. Note that predictions on known genes were used only for validation in previous studies. Our predictions on known genes, on the other hand, additionally provide the context information of genes' function. Among our assignment, 20% genes are assigned with multiple functions based on different network environments. The functional predictions together with the necessary context information are available at http://zhoulab.usc.edu/ContextAnnotation.

2 MATERIALS AND METHODS

2.1 Microarray data

We collected 65 human microarray datasets including 52 Affymetrix (U133 and U95 platforms) and 13 cDNA datasets (details see Supplementary Material) from the NCBI GEO (Edgar *et al.*, 2002) and SMD (Gollub *et al.*, 2003) databases (version December 2005). The selection criteria are that each dataset contains at least eight experiments and that the percentage of statistically significant co-expressed gene pairs (see the section of Graph Construction for details) is not higher than 3%. The first criterion ensures that the dataset contains enough profiles so that the constructed co-expression graph is reliable while the second criterion filters out the datasets, in which too broad perturbations result in a large number of spurious co-expression estimates. The collected datasets are preprocessed as follows. The datasets generated from Affymetrix chips are log transformed (base e) to place them on the same scale as the cDNA datasets. Note that the original values less than 10 in Affymetrix datasets are set to 10. For each dataset, genes with low expression variation (lowest 10% in terms of the ratio of SD to mean for Affymetrix data and of SD for cDNA data.) are discarded. Finally, genes with more than 30% missing values and arrays with more than 20% missing values are also discarded.

2.2 Gene Ontology function categories

The Gene Ontology file on Biological Processes was downloaded from GO consortium (February 2005). The associated annotation information for human genes was from the NCBI Gene Database (August 2005). With the method proposed in (Zhou *et al.*, 2002), we selected process categories from GO that contain more than 175 genes but each of their children contains <175 genes. The 40 GO categories obtained are called the informative functional categories and will be used in functional annotation later. Note that genes with annotations but not belonging to any of the informative categories were discarded. After the data preprocessing, the 65 datasets comprise in total 8297 genes, of which 5629 have at least one known function and 2668 do not have any known functions.

2.3 Graph construction

Each microarray dataset is modeled as a relation graph where each node represents one gene and two genes are connected if their expression correlation is significant. Here the expression correlation, denoted as r, is taken as the minimum of the absolute value of leave-one-out Pearson correlation coefficients, which is robust against single experiment outliers and sensitive to overall similarities in expression patterns (Zhou *et al.*, 2002). We then use the statistic $t = \sqrt{(n-2)r^2/(1-r^2)}$ to determine if the expression correlation is significant. More precisely, the quantity t is modeled as a t-distribution with $n-2$ degrees of freedom, where n is the number of measurements used in the computation of the correlation. In our study, an expression correlation significant at $P \leq 0.01$ level is included as an edge in the relation graph.

2.4 Mining recurrent network modules

Given 65 graphs, each of which contains (at most) 8297 nodes, we attempt to identify connected network patterns that comprise at least 4 nodes and that occur in at least five graphs. This is computationally very difficult due to the exponential number of potential patterns. In our approach, we first search for frequent edge sets that are not necessarily connected and then extract connected components from them. Conceptually, we represent the 65 graphs as a matrix where each row represents an edge (i.e. a gene pair), each column represents a graph, and each entry (0 or 1) indicates whether the edge appears in that graph. Clearly, our problem of discovering frequent edge sets can be formulated as a typical biclustering problem that searches for submatrices with high density of 1s, which is a well-known NP-hard problem.

We developed a biclustering algorithm based on simulated annealing to discover frequent edge sets. More precisely, we employ simulated annealing to maximize the objective function $c'/(mn+\lambda c)$, where c is the number of 1s in the input matrix, c', m and n are the numbers of 1s, rows and columns of the bicluster, respectively, and λ is a regularization factor. Clearly, such an objective function is in favor of biclusters with a high density of 1 and with large size. Note that the density is maximized to 1 when $c'=mn$, while the size of bicluster is maximized when $c'=c$ (i.e. the pattern is as large as the input matrix). The regularization parameter λ controls the compromise between the density and the size. However, there is no theoretic result on selecting optimal λ. In the study, we tried many heuristic choices of λ and the reported results are based on $\lambda = 0.2 / \max(1, \log_{10} n_1)$, where n_1 is the number of edges of the initial configuration (i.e. seed).

Although this method performs well in our experiments, the search space has to be restricted in order to discover hundreds of thousands patterns in reasonable time because of the huge size of matrix (more than 1 million rows and 65 columns). As an attempt to solve this problem, we employ the FIM technique (Grahne and Zhu, 2003) to restrict the search space and also provide seeds for our biclustering algorithm. In what follows, we briefly describe FIM and related concepts. Let $I=\{i_1, i_2, \ldots, i_n\}$ be a set of items and database D be a set of m transactions, where each transaction T is a set of items such that $T \subseteq D$. Let X be a set of items. A transaction T is said to contain X if and only if $X \subseteq T$. X is frequent if at least s transactions in the database contain X. In our case, we regard an edge as a transaction and its occurrence in a particular graph as an item. For our purpose, we include only edges occurring in at least five graphs in the transaction dataset. The output of FIM algorithms is the set of all possible item sets which occur in at least s transactions. Note that the submatrices of frequent itemsets and their supporting transactions are actually biclusters full of 1s. These clusters with perfect density can then serve as seeds for our biclustering algorithm to search for larger biclusters that permit holes (i.e. 0s). We ended up with $\sim$ 1.8 millions frequent itemsets which contain at least four edges and occur in at least five graphs. These FIM patterns, however, should not be used as seeds directly because they are highly overlapping, which is due to the nature of frequent itemset's definition. This is well known in data mining community. In order to improve these seeds and also reduce unnecessary computation in final biclustering, we first remove FIM patterns whose supporting transactions/edges are the subset of those of other patterns. Second, we also merge two patterns if their union has a density larger than 0.8. This procedure is repeated until no additional merge can happen.

After the post-processing, we finally have about half million merged FIM patterns to feed our biclustering algorithm. Given a FIM pattern with v member genes, we will use all possible $v(v-1)/2$ edges among these v genes and all datasets as the input matrix for our biclustering algorithm. The FIM pattern is also used as the initial configuration of simulated annealing. From the output biclusters of our algorithm, we extract connected components as the final output patterns.

2.5 Network topology score for each function category

Given a network pattern, the most popular gene function prediction method involves the use of the hyper-geometric distribution to model the probability of genes function based on neighborhood. This method however ignores the network topology, which is probably the most important information in the network patterns. To avoid this problem, we developed a new method based on graph random walk to fully explore the topology of network patterns. Our method is still based on the principle of 'guilt by association'. In terms of network topology, the association between genes is measured by how close they are (i.e. the length of path between them) and how tightly connected they are (i.e. how many paths between them). Statistically, this translates to how likely it is to reach one gene starting from another gene in a random walk. This probability can be approximately calculated by matrix multiplication.

Given a network pattern consisting of v genes, let P be a stochastic matrix of size $v \times v$, of which the element P_{ij} is $1/n_i$ if genes i and j are connected, or 0 otherwise, where n_i is the number of neighbors of gene i. If we regard genes as states and P_{ij} as the probability of transformation from genes/states i to j, then the random walk on the graph can be thought as a Markov process. Therefore, it is easy to see that the element of P^k is the probability that gene i reaches gene j in k steps of the walk. The intuition behind our method is that genes with similar function are more likely to be well connected (i.e. gene i can reach gene j with high probability in a random walk). Simply put, the probability P^k_{ij} would be large if genes i and j share the same function. Let o be the Gene Ontology binary matrix of which element o_{ij} is 1 if gene i belongs to category j and 0 otherwise. Thus, the matrix $M = P^k\, o$ gives the scores of genes relating to functional categories. The higher the score, the more likely a gene has that function. In practice, we choose $k=3$ because we would like to confine our prediction to a local area of network patterns. With the score matrix M, the function of each gene is estimated by finding the functional category with the maximum score in the corresponding row of M.

2.6 Assessment of function assignments with random forest

A random forest is a classifier consisting of a collection of tree-structured classifiers (Breiman, 2001). In an attempt to improve our method, we wanted to include other attributes of network patterns in the final prediction beside the network topology score. In particular, we include recurrence, density, size, average node degree, percentage of unknown genes and functional enrichment of network modules. To take those factors into account, we use a random forest to determine whether the function assignments based on the network topology score is robust. Note that the purpose of the random forest here is to determine whether to accept or reject the functional assignment made based on the network topology score. The random forest was trained using the assignments of known genes. The trained model was then applied to assignments of unknown genes. Finally, we keep only the function assignments that the random forest classified as 'accept'.

3 RESULTS

3.1 Systematic identification of functional modules in human genome

We constructed 65 co-expression networks from 65 microarray datasets. In total, the graphs contain 8297 genes. Using the graph-based mining approach as described in Methods Section, we obtained 1 823 518 network patterns which occur in at least five graphs. We further designed a bi-clustering approach (see Methods Section) to merge patterns similar in both their network topology and their dataset recurrence. This drove down the number of network patterns to 143 400, which covers 2769 known and 1054 unknown genes and varies in size from 4 to 180. The whole pipeline took 254 CPU hours (2 GHz AMD Opteron Processor 270). In general, the size of a module is inversely proportional to its recurrence. Among those modules, 45% of modules each contain more than 90% known genes, which allow us to assess the functional homogeneity of the module. We define a module to be functionally homogenous if the hyper-geometric P-value, after Bonferroni correction,

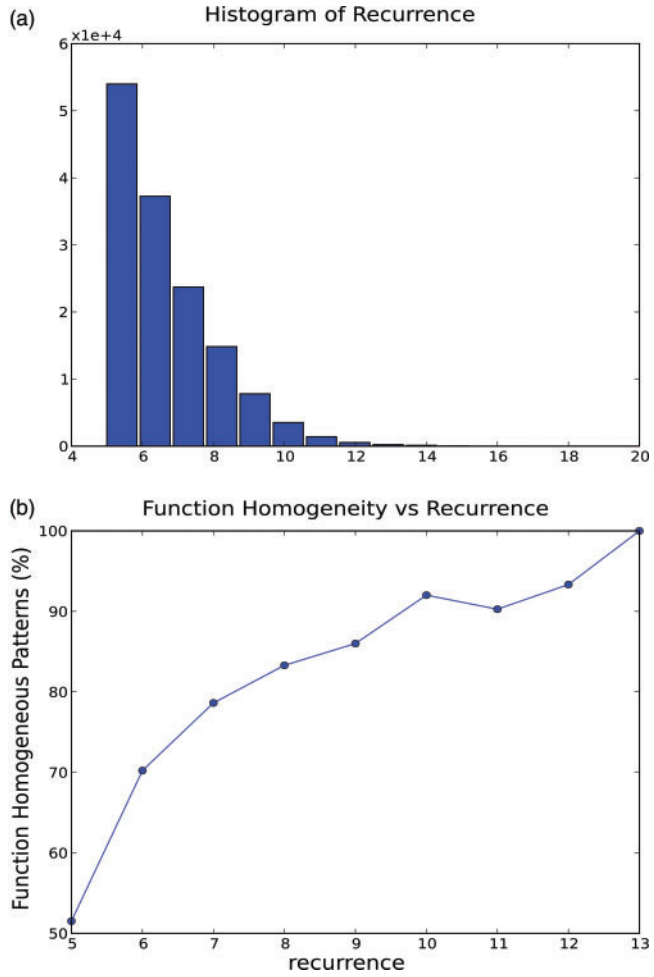

Fig. 1. (**a**) Distribution of recurrence of identified modules approximates an exponential distribution. (**b**) For all modules containing seven genes, the percentage of function-homogenous modules increases with recurrence. Modules of different sizes show similar trend (see Supplementary Material).

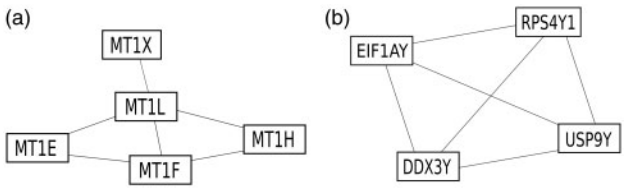

Fig. 2. Two network modules with high recurrences.

is <0.01. Among the identified network modules, 77.0% of the patterns are functionally homogenous.

Figure 1a shows the histogram of network recurrence across the 65 datasets, which approximates an exponential distribution. We define a module to be active in a dataset if 80% of its edges appear in that dataset. The recurrence of those modules ranges from 5 to 20. The most frequently occurring module (with recurrence 20) contains five genes MT1E, MT1F, MT1H, MT1L and MT1X (the network topology is shown in Fig. 2a), all of which are metallothionein and only MT1X has known annotations in the GO database as 'nitric oxide mediated signal transduction' and 'response to metal ion'. Multiple experimental studies have revealed concurrent activities of the MT genes in intracellular defense against reactive oxygen and nitrogen species. For example, substances causing oxidative stress and agents involved in inflammatory processes induce the synthesis of metallothionein (Chun *et al*., 2004; Chung *et al*., 2006; Izmailova *et al*., 2003). This evidence is consistent with their observed tight expression regulation in our study. Another frequent occurring module, comprising four genes RPS4Y1, USP9Y, DDX3Y and EIF1AY as a clique (the network topology is shown in Fig. 2b), occurs in 18 datasets. Interestingly, USP9Y, DDX3Y and EIF1AY are all located in the chromosomal region Yq11 and considered to be involved in spermatogenesis (Vogt, 2005). In addition, RPS4Y1 shares high sequence similarity with RPS4Y2, which also resides in Yq11 and is linked to spermatogenesis. These examples demonstrate that recurrent network modules are highly likely to be involved in a specific biological process.

In general, the higher the recurrence, the more likely the modules are to be functionally homogenous (Fig. 1b). This lays the foundation for using multiple microarray datasets to enhance the functional inferences. In fact, when the recurrence is high, even loosely connected network patterns, or paths, can represent functional modules. We define the connectivity of a graph g to be $2m/[n(n-1)]$, where m is the number of edges and n the number of vertices in g. Figure 3a shows an example. All seven genes are involved in 'immune response', though extremely loosely connected; they are identified through their occurrence in six graphs. Most current algorithms identify network modules by looking for densely connected subnetworks in a single network (Bader and Hogue, 2003; Shannon *et al*., 2003; Spirin and Mirny, 2003). Here, by considering pattern recurrence across many networks, we are able to identify network modules of most topologies. In fact, 24% of the identified modules have connectivity <0.5. Figure 3b shows the network connectivity distribution of the modules.

3.2 Enrichment of protein–protein interaction in network modules

To explore the types of interaction relations among the network members beyond co-expression, we resort to the only available large-scale interaction information, protein interaction data. We retrieved the human protein interaction information from EBI(European Bioinformatics Institute)/IntAct (Hermjakob *et al*., 2004) (version 2006-10-13), and test for each of the 143 400 modules whether protein interaction was over-represented among its member genes compared to all human genes based on the hyper-geometric test. Total 60556 (22.44%) network modules were found to be enriched in protein interaction at P-value 0.001 level. This shows that genes in our network modules are more likely to encode interacting proteins. Interestingly, many of the protein interaction-enriched network modules fall into functional categories, such as protein biosynthesis, DNA metabolism, etc., and many interacting protein pairs are not necessarily co-expressed. Figure 4 shows an example of such network modules with two edge colors indicating co-expression and protein interaction, respectively.

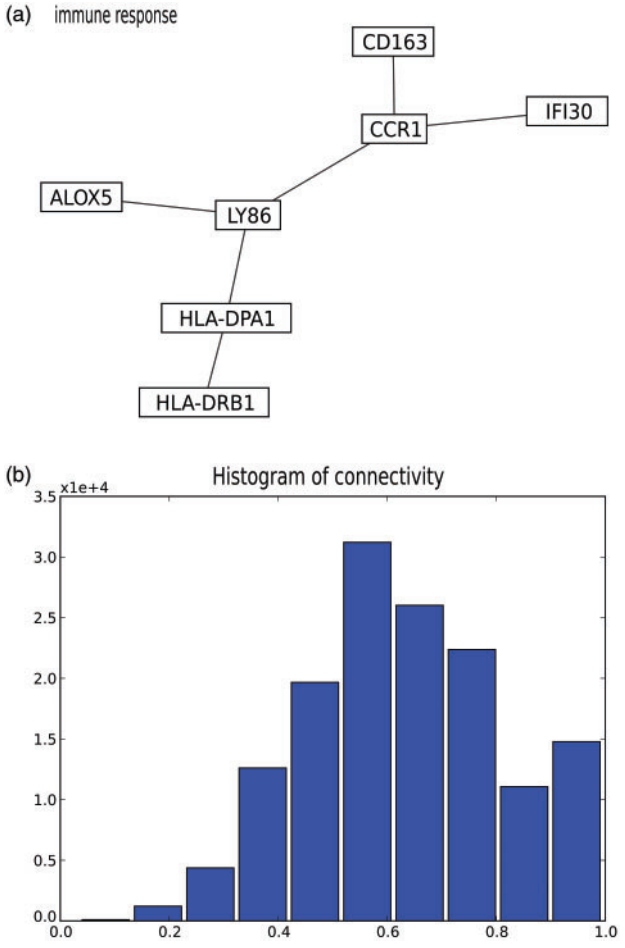

Fig. 3. (**a**) A loosely connected network module (connectivity = 0.28) enriched with the function 'immune response' ($P < 10^{-7}$). (**b**) Distribution of network connectivity among identified modules.

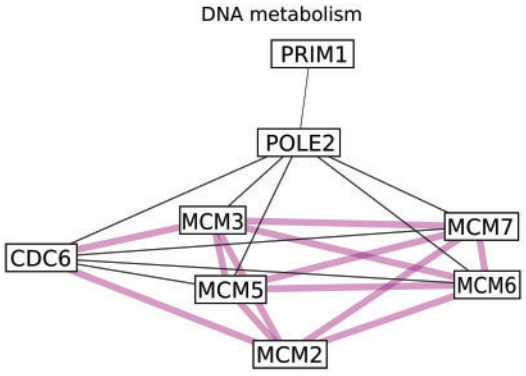

Fig. 4. A network pattern enriched with protein interaction. Edges representing protein interaction are colored in red.

3.3 Function prediction with random forest classifier

Based on the 143 400 recurrent network patterns, we assigned to each gene the function with maximum network topology score in each network pattern and made functional predictions for 779 known and 116 unknown genes by random forest with 70.5% accuracy. In our random forest model, there are seven explanatory variables: functional enrichment *P*-value, network topology score, network connectivity, network size, average node degree, unknown gene ratio and the pattern recurrence numbers.

In the application of random forest, a small number k of the variables are randomly selected to split the node for each tree. We proved that the performance of random forests is not sensitive to the choice of k (details on Supplementary Material), which is also an expected result based on Breiman's comments (Breiman and Cutler, 2003).

The predicted gene functions cover a wide range of functional categories, e.g. protein biosynthesis, electron transport, vesicle-mediated transport and immune response. Each prediction is made conditionally on the gene's network environment and specific perturbations. Some functions, such as protein biosynthesis, occur universally under almost all perturbations. Others, such as cell cycle, are activated predominately in conditions related to cancer and development. The comprehensive prediction results, together with the necessary context information, are available at http://zhoulab.usc.edu/ContextAnnotation. Many of our predictions are supported by experimental studies in the literature. For example, we predicted NCF4 to participate in 'immune response'. According to a study (Wientjes *et al.*, 1993), NCF4 is important for immunity and its deficiency leads to chronic granulomatous disease (CGD). We assigned the function 'mitotic cell cycle' to AURKB; and AURKB is known to be responsible for mitotic arrest in the absence of aurora A (Yang *et al.*, 2005). We predicted RPS8 to be involved in 'protein biosynthesis', and RPS8 has been shown to participate in translation (Yu *et al.*, 2005). Spc25 was predicted to be involved in 'mitotic cell cycle', which is supported by the evidence that SPC25 is an essential kinetochore component that plays a significant role in proper execution of mitotic events (Bharadwaj *et al.*, 2004).

It should be noted that the prediction accuracy of 70% is an underestimate due to the sparse nature of human GO annotations. Since GO annotation is based only on positive biological evidence, many annotated genes may still have other undiscovered functions. Furthermore, the GO directed acyclic graph structure is not perfect. For example, we predicted ch-TOG to have 'mitotic cell cycle' function. Based on recent evidence (Cassimeris and Morabito, 2004), the updated (2006 December) GO classifies it as 'RNA transport', 'centrosome organization and biogenesis', 'spindle pole body organization and biogenesis' and 'establishment and/or maintenance of microtubule cytoskeleton polarity'. Since none of these four GO nodes is a child of 'mitotic cell cycle', we have to classify the prediction to be wrong, while the original paper clearly documented its important involvement in mitotic cell cycle (Cassimeris and Morabito, 2004).

3.4 Context-specific function annotation

One of the surprises of the human genome project is that we have far fewer genes than expected. A possible explanation to relate the limited number of genes to the high degree of complexity is that many genes perform multiple functions.

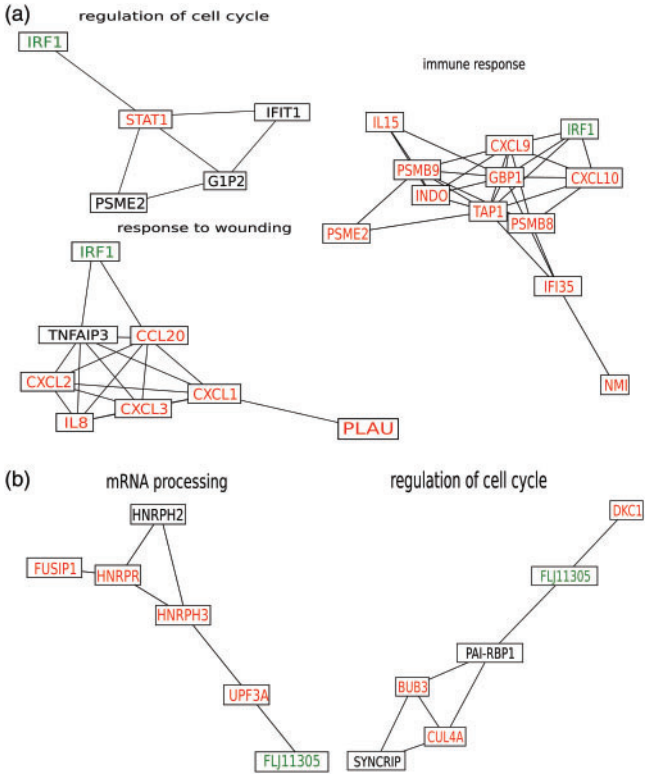

Fig. 5. Functional predictions for (**a**) IRF1 and (**b**) FLJ11305 upon different network environments. Nodes labeled in red are annotated with the titled functions by GO consortium. Nodes labeled in green are the genes with predicted functions.

Since our approach allows one gene to appear in more than one network module, we are able to perform context-sensitive functional annotation. That is, we can assign to a gene multiple functions as well as the network environments in which the gene exerts those functions. This is valuable even if a gene's function is already known. Among our predictions, 20% of genes are assigned multiple functions. This is certainly an under-estimate, since for each network module we only picked the functional category associated with the highest network topology score. Of course, some of the different assignments for the same gene are relevant, such as 'response to pest, pathogen or parasite' and 'immune response', or 'regulation of cell cycle' and 'mitotic cell cycle'. However, the dramatic difference in the network environment associated with those functional assignments indicates different functional involvement of this gene, which is beyond the rough classification of GO functional categories. In our predictions, among the genes assigned with multiple functions, 72% are in network modules that differ at least 50% of member genes and 57% are in network modules that differ at least 70% of member genes.

Figure 5 shows two examples of genes predicted to have multiple functions. In Figure 5a, IRF1 appears in three different network modules, and annotated with the functions 'immune response', 'regulation of cell cycle', and 'response to wounding', respectively. The first pattern appears in six datasets (details on Supplementary Material). The dataset conditions include cancer, infection and inflammatory responses, which is consistent with IRF1's role in 'immune response'. The second pattern appears also in six datasets, measuring exercise effect, infections and cancer. Since cell cycle may also be accelerated upon inflammatory responses, and conditions such as 'cancer' may impact various pathways, it is hard to conceptually separate those two datasets into two types of strictly different conditions. In fact, in two of those datasets measuring infection (GDS260 and SMD dataset with Category = Infection, Subcategory = PBMC, experimenter = Cheryl Hemingway), the two network modules merge into one, indicating a potential role of IRF1 in mediating cross-pathway communication. The third pattern annotated with the function 'response to wounding' occurs in five datasets. Since the process 'response to wounding' is highly related to the process 'immune response', and it may also initiate the acceleration of 'cell cycle', the dataset conditions are similar to those previously described, except the condition 'osmotic stress reaction', which is in agreement with the specific process 'response to wounding'. The first two functions 'immune response' and 'regulation of cell cycle' agree with the known annotation of IRF1, and we believe that the function 'response to wounding' is also likely to be true due to strong evidence (the hyper-geometric *P*-value measuring the module functional homogeneity is 10^{-5}). As another example, Figure 5b shows that the unknown gene FLJ11305 occurs in two different network modules, and is annotated with two functions 'mRNA processing' and 'regulation of cell cycle'. Each module is activated in five datasets, including drug treatment, dyslipidemia, Huntingtons disease, exercise effect and cancer, with the conditions dyslipidemia and ovarian tumor being shared between the two modules. The fact, that often different network modules involving the same gene can be merged together under some conditions, indicates that many genes with multiple functions may participate in related pathways, and they are likely to serve as cellular process communicator.

3.5 Discovery of uncharacterized cellular systems

The comprehensive functional modules generated in this study can facilitate the discovery of uncharacterized cellular systems. Date and Marcotte defined 'uncharacterized cellular systems' as discrete subgraph in reconstructed protein interaction networks in which 50% or more member proteins lack functional assignments (Date and Marcotte, 2003). Among the identified modules, we identified 2206 such modules, varying in size from 4 to 68 member genes. Among those, 204 modules contain only unannotated genes. Figure 6a shows an example. The module contains five genes, DRE1, C15orf25, FLJ11029, FLJ12151 and FLJ14346, that form a densely connected subgraph. The complete subgraph appears in eight datasets (GDS1062, 1312, 1321, 505, 564, 760, 858, 914). Notably, 5 out of the 8 datasets are cancer datasets. Although cancer datasets are enriched in our collected data (21 out of 65), the ratio is still marginally significant at ($P = 0.06$) level. This suggests the potential involvement of the module in cancer. Interestingly, the homolog of DRE1 in *Drosophila* plays an important role in regulating DNA replication-related genes

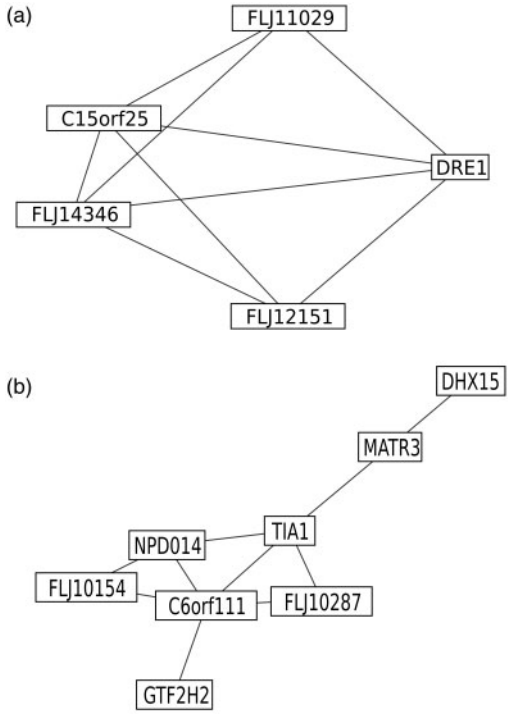

Fig. 6. Two examples of uncharacterized cellular systems.

(Okudaira *et al*., 2005), which may suggest its potential role in cell cycle or cell proliferation—a hypothesis consistent with its activation in cancer. This example demonstrates that even for the poorly uncharacterized modules, our method may provide useful information based on the experimental conditions under which the module is activated. Furthermore, 251 uncharacterized modules have connectivity <0.5, which can hardly be identified from a single graph. Figure 6b shows such an example. Among the eight member genes, three have annotated functions: DHX15 is involved in mRNA processing, GTF2H2 participate in regulation of transcription and TIA1 is a member of a RNA-binding protein family. The exact function of MATR3 is unknown, but it is known to encode a nuclear matrix protein, which may play a role in transcription. These evidences point to a possible involvement of the module in transcription.

4 CONCLUSION AND DISCUSSION

We have presented a generic approach to integrate many microarray datasets to identify functional modules and to perform functional annotation in human genome. To our knowledge, this is the first study to systematically annotate human gene functions based on multiple microarray datasets. Compared to current approaches based only on a single microarray dataset, our method provides: (1) higher specificity: the identified functional modules are more likely to be functionally homogenous; (2) higher sensitivity: we can identify functional modules beyond co-expression clusters.

Our approach is based on pattern mining across co-expression networks. It is known that absolute expression values of a gene cannot be compared across datasets. However, the expression correlations of a gene pair in different datasets are comparable because they are unitless measures each derived from a single dataset. As our co-expression networks are constructed from expression correlations of gene pairs, their comparisons are not affected by inter-dataset variations. Thus, our approach provides an effective way to integrate a large number of microarray experiments conducted in different laboratories, at different times, and using different technology platforms. There are large numbers of public microarray datasets available for model organisms, such as *Homo sapiens, Mus musculus, Arabidopsis thaliana, Drosophila melanogaster, Caenorhabditis elegans and Saccharomyces cerevisiae*. Using our approach, we are in a position to extract orders of magnitude more information for any genome, for which large amount of microarray data exists. A natural extension will be to compare co-expression networks across species. Several studies along this direction (Bergmann *et al*., 2004; Oldham *et al*., 2006; Stuart *et al*., 2003) have already been performed on two or several species. To extend those studies to many species is likely to require efficient algorithm design.

In our studies, 20% of genes received multiple functions in different network contexts. This is certainly an underestimate for at least three reasons: (1) we only assigned one function to a gene based on its top network topology score; (2) the whole functional module may perform multiple functions in different contexts of activities of other modules and (3) our data source only includes 65 datasets, that mostly represent human pathological conditions and cover a small proportion of human dynamical functional landscape. We note that incorporating the concept of dynamics is especially important in charactering human gene functions due to its high temporal and spatial complexity. However, relating specific conditions to particular gene functions is not an easy task due to the subtle difference in experimental conditions, and the difficulties in systematically characterize them.

The principle of our approach, integrating multiple networks for functional studies, can be extended beyond microarray analysis. For example, a popular approach to functional modules is to identify dense subgraphs on protein interaction networks (Chen and Yuan, 2006; Hwang *et al*., 2006; Koyuturk *et al*., 2006; Luo *et al*., 2007; Spirin and Mirny, 2003; Tornow and Mewes, 2003). However, as discussed above, functional modules often occur as non-dense subgraphs, e.g. metabolic and signal pathways. Furthermore, since current protein interaction network are static networks, edges in such a network may not occur together if considering temporal or spatial parameters. Thus, such identified functional modules may not truly represent a functional unit. In the future, given protein interaction networks generated under different conditions, our approach can further facilitate the identification of condition-specific functional modules or dynamic protein complex assembly. In fact, if different species are conceptualized to represent manifestations of different conditions of life forms, several recent studies on conservation and evolution of protein interaction networks across species can be regarded as a first attempt to characterize network dynamics (Flannick *et al*., 2006; Kelley *et al*., 2003; Koyuturk *et al*., 2006; Sharan and Ideker, 2006; Sharan *et al*., 2005). In that context, due to the

NP-hard graph isomorphism problem, how to perform large-scale pattern mining across protein interaction networks of many species is still a challenging problem.

ACKNOWLEDGEMENTS

We thank Juan Nunez-Iglesias for his comments on the manuscript. This work is partially supported by the NIH grants R01GM074163, P50HG002790, U54CA112952 and the NSF grant 0515936. X.J.Z. is an Alfred P. Sloan Research Fellow.

Conflict of Interest: none declared.

REFERENCES

Aggarwal,A. *et al.* (2006) Topological and functional discovery in a gene coexpression meta-network of gastric cancer. *Cancer Res.*, **66**, 232–241.

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Bergmann,S. *et al.* (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.*, **2**, E9.

Bharadwaj,R. *et al.* (2004) Identification of two novel components of the human NDC80 kinetochore complex. *J. Biol. Chem.*, **279**, 13076–13085.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Breiman,L. and Cutler,A. (2003) Manual – Setting Up, Using, And Understanding Random Forests V4.0.

Cassimeris,L. and Morabito,J. (2004) TOGp, the human homolog of XMAP215/Dis1, is required for centrosome integrity, spindle pole organization, and bipolar spindle assembly. *Mol. Biol. Cell*, **15**, 1580–1590.

Chen,J. and Yuan,B. (2006) Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, **22**, 2283–2290.

Chun,J.H. *et al.* (2004) Increased expression of metallothionein is associated with irinotecan resistance in gastric cancer. *Cancer Res.*, **64**, 4703–4706.

Chung,M.J. *et al.* (2006) Cytotoxicity of nitric oxide is alleviated by zinc-mediated expression of antioxidant genes. *Exp. Biol. Med. (Maywood)*, **231**, 1555–1563.

Date,S.V. and Marcotte,E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.*, **21**, 1055–1062.

Drysdale,R.A. *et al.* (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–D395.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Flannick,J. *et al.* (2006) Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res.*, **16**, 1169–1181.

Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**, 0059.

Gollub,J. *et al.* (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94–96.

Grahne,G. and Zhu,J. (2003) Efficiently using prefix-trees in mining frequent itemsets. In *Proceedings of the ICDM Workshop on Frequent Itemset Mining Implementations*.

Hermjakob,H. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, 452.

Hwang,W. *et al.* (2006) A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol. Biol.*, **1**, 24.

Izmailova,E. *et al.* (2003) HIV-1 Tat reprograms immature dendritic cells to express chemoattractants for activated T cells and macrophages. *Nat. Med.*, **9**, 191–197.

Jeffery,C.J. (2003a) Moonlighting proteins: old proteins learning new tricks. *Trends Genet.*, **19**, 415–417.

Jeffery,C.J. (2003b) Multifunctional proteins: examples of gene sharing. *Ann. Med.*, **35**, 28–35.

Kelley,B.P. *et al.* (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA*, **100**, 11394–11399.

Koyuturk,M. *et al.* (2006) Detecting conserved interaction patterns in biological networks. *J. Comput. Biol.*, **13**, 1299–1322.

Luo,F. *et al.* (2007) Modular organization of protein interaction networks. *Bioinformatics*, **23**, 207–214.

Lussier,Y. *et al.* (2006) PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac. Symp. Biocomput.*, 64–75.

Niehrs,C. and Pollet,N. (1999) Synexpression groups in eukaryotes. *Nature*, **402**, 483–487.

Okudaira,K. *et al.* (2005) Transcriptional regulation of the Drosophila orc2 gene by the DREF pathway. *Biochim. Biophys. Acta*, **1732**, 23–30.

Oldham,M.C. *et al.* (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl Acad. Sci. USA*, **103**, 17973–17978.

Qian,J. *et al.* (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol.*, **314**, 1053–1066.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.*, **24**, 427–433.

Sharan,R. *et al.* (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tornow,S. and Mewes,H.W. (2003) Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res.*, **31**, 6283–6289.

Vogt,P.H. (2005) Azoospermia factor (AZF) in Yq11: towards a molecular understanding of its function for human male fertility and spermatogenesis. *Reprod. Biomed. Online*, **10**, 81–93.

Wang,W. *et al.* (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl Acad. Sci. USA*, **102**, 1998–2003.

Wientjes,F.B. *et al.* (1993) p40phox, a third cytosolic component of the activation complex of the NADPH oxidase to contain src homology 3 domains. *Biochem. J.*, **296**, 557–561.

Wu,L.F. *et al.* (2002) Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nat. Genet.*, **31**, 255–265.

Yang,H. *et al.* (2005) Mitotic requirement for aurora A kinase is bypassed in the absence of aurora B kinase. *FEBS Lett.*, **579**, 3385–3391.

Yu,Y. *et al.* (2005) Mass spectrometric analysis of the human 40S ribosomal subunit: native and HCV IRES-bound complexes. *Protein Sci.*, **14**, 1438–1446.

Zhou,X. *et al.* (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.

Zhou,X.J. *et al.* (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol.*, **23**, 238–243.

Zhu,Z. *et al.* (2005) Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.*, **15**, 848–855.