Extracting functional modules from biological pathways

Sihai Dave Zhao[1,2,*], Yong Li,[1,*]

[1]Molecular Discovery Informatics, GlaxoSmithKline R&D, 709 Swedeland Road, UMW2230, King of Prussia, PA 19406, [2]Present address: Department of Biostatistics, Harvard School of Public Health, Huntington Ave., Boston, MA 02115

*To whom correspondence should be addressed

ABSTRACT

It has been proposed that functional modules are the fundamental units of cellular function (Hartwell *et al*., 1999). Methods to identify these modules have thus far relied on gene expression data or protein-protein interaction (PPI) data, but have a few limitations. We propose a new method, using biological pathway data to identify functional modules, that can potentially overcome these limitations. We also construct a network of these modules using functionally relevant PPI data. This network displays the flow and integration of information between modules and can be used to map cellular function.

Contact: szhao@fas.harvard.edu, yong.2.li@gsk.com

INTRODUCTION

Biological functions of the cell are products of complex molecular interactions and thus are difficult to analyze at the single-molecule, or single-gene, level. Functional modules, relatively independent sets of molecular components that have separable biological functions (Hartwell *et al*., 1999), offer a more relevant level of analysis. Furthermore, networks of these modules, where two modules are linked if they functionally influence one another, can map the flow of information in the cell, thus enhancing understanding of the underlying biology (Barabási and Oltavi, 2004, Segal *et al*., 2004).

To date, identification of these functional modules and construction of module networks have generally revolved around the use of three major methods of analysis to infer modules from two major types of data. Expression data have been clustered (Prelić *et al*. 2006; Segal *et al*., 2004), or transformed to a graph and analyzed with topological methods (Chen and Yuan, 2006; Voy *et al*., 2006). Similarly, protein-protein interaction (PPI) data have been transformed into an association matrix and clustered (Rives and Galitski, 2003), or analyzed with topological methods (Pereira-Leal *et al*., 2004; Spirin and Mirny, 2003). Finally, statistical procedures have been used to infer modules from either, or both, of these types of data (Lee *et al*., 2004; Segal *et al*., 2003).

But there are limitations to the identification of modules using these approaches that stem from the types of data on which the approaches are based. First, methods using expression data may overlook some modules. Expression experiments only examine genes that are related through changes in transcription levels, but transcription is not the sole measure of gene activity. Second, modules discovered by methods based on PPI data might not actually exist *in vivo.* A set of genes may form a highly-connected subgraph and thus be identified as a functional module, but if the genes in the set interact under different conditions and these conditions do not coexist *in vivo*, this set is not a true module.

APPROACH

In hopes of overcoming these issues, we propose using human-curated biological pathway data, instead of expression or PPI data, to identify functional modules. Because genes assigned to the same pathway already share a functional relationship, genes that appear together in multiple pathways may have a tight functional relationship. Furthermore, different sets of co-occurring genes are separable, because they co-occur in different sets of pathways, potentially from different biological processes. We see that these sets of co-occurring genes fit exactly the definition of a functional module. In contrast to previous work, ours is a descriptive, rather than an inferential, approach to finding modules.

This approach can potentially address the two problems mentioned previously. First, modules are not likely to be overlooked, because pathway data include genes related not only by

transcription, but also by many other types of interaction, and thus contain modules that would not be found by expression data-based methods. Second, the discovered modules will exist *in vivo,* because genes found together in well-curated pathways were placed there precisely because the curators believed them to function together *in vivo.*

In this paper we evaluate our new approach for finding functional modules. We use frequent pattern mining to extract modules from biological pathway data. Our ultimate goal is to understand how cell functions not through individual modules, but through interactions among many modules. To move toward this goal, we then construct a network of functional modules by linking them together using well-annotated PPI data, which contain information about the direction and the nature of the protein interactions. For example, the PPI data may show that gene A positively influences the phosphorylation of gene B. We link two modules only if the interactions between their genes have clear functional implications. For example, when gene A has a positive effect on the phosphorylation of gene B, it's likely that gene A may influence the activity of gene B. On the other hand, mere binding interactions between A and B would not suffice. We find that our method produces a network of interesting and useful modules. We note that while we were preparing this manuscript, Huang *et al*. published a related approach to finding modules, using the same frequent pattern mining method to find recurrent network patterns in expression data (Huang *et al*., 2007).

MATERIALS AND METHODS

Generating modules

Pathway data

We collected pathway data from the following sources:

- BioCarta (http://www.biocarta.com/)
- Gene Ontology (GO) consortium (http://www.geneontology.org/)
- GeneGO (http://www.genego.com/)
- GenMAPP (http://www.genmapp.org/)
- GlaxoSmithKline internal data
- HumanCyc (http://humancyc.org/)
- Jubilant Biosys (http://jubilantbiosys.com/index.htm)
- Molecular Connections (http://www.molecularconnections.com/home.html)

We first removed all non-leaf nodes from the GO tree, because non-leaf nodes include every gene in the nodes below them in the tree, and this would artificially skew the recurrence count of those genes. We next used only pathways with between 2 and 100 genes, because we believed that genes in large pathways were not as closely related as genes in smaller pathways. This resulted in a total of 5426 pathways containing 8422 unique genes.

Extracting modules

We used frequent pattern mining to extract recurrent sets of genes to be used as candidate for functional modules. An introduction to frequent pattern mining can be found in Yahia *et al*. (Yahia *et al*., 2006), but briefly, a transaction is a set of items, and a frequent itemset is a set of items that occurs in more than a user-specified number of transactions. The support of a frequent itemset is the number of transactions containing the itemset. A frequent closed itemset is a frequent itemset for which there is no superset that has the same support. Treating pathways as transactions and genes as items, we used CLOSET+ (Wang *et al*., 2003), an algorithm available from Illimine (http://illumine.cs.uiuc.edu/), to generate a comprehensive list of frequent closed itemsets with a minimum support of five. We then removed all modules that came only from one pathway data source. This list formed our list of functional modules candidates.

To be more confident that the genes in each candidate actually functioned together, we checked whether the genes co-occurred more often than they would have by mere chance. To do so, we first generated a randomized set of pathways. Using our original data, we determined the frequency of occurrence of every gene (the number of pathways containing that gene). We then replaced each gene in each pathway of our original dataset with another gene randomly selected from a set of genes with similar frequencies of occurrence. Each of these sets was constructed to contain an adequate number of genes. The probability of selecting a particular gene from a set was weighted by the frequency of that gene relative to the frequencies of the other genes in the set. We simulated 1000 of these randomized pathway datasets. For each frequent closed itemset

in our functional module candidate list, we determined the support of that itemset in each of the randomized datasets. We then selected only those module candidates whose supports were higher than 95% of their supports under the 1000 randomized pathway datasets.

In order to select high-quality significant modules, we ranked the candidates according to a composite score function that combined the sizes of the pathways containing the module, the support, the number of data sources, and the size of the module:

$$Score = \frac{1}{Support}\left(\sum_{i=1}^{Support}\frac{1}{PathwaySize_i}\right) Sources \cdot \log_2 Support \cdot \log_2 ModuleSize$$

Our score function was adapted from Okada *et al*. (Okada *et al*., 2007), and penalizes modules supported by larger pathways, since genes in larger pathways are less closely related. We have also tried ranking by support level or number of data sources (data not shown), but the score function produced modules with better size distribution (about one-third of final modules have at least 3 genes) and higher support for each size groups (Supplementary Table 1). Next, moving from highest to lowest rank, we counted the number of genes in each successive module that had already been included in previously selected modules. If the number of redundant genes made up more than 25% of the genes in the module, the module was not selected. The cutoff of 25% was chosen because this would allow for at most one out of 4 genes in a module to overlap with pre-

vious modules, which we thought was a reasonable allowance. In this way we produced a final module list containing high-quality, and relatively mutually independent, modules.

Generating the module network

PPI data

We collected PPI data from the following sources:

- GeneGO (http://www.genego.com/)
- Ingenuity (http://www.ingenuity.com/)
- Jubilant Biosys (http://jubilantbiosys.com/index.htm)
- Molecular Connections (http://www.molecularconnections.com/home.html)
- PRIME (http://prime.ontology.ims.u-tokyo.ac.jp:8081/)
- Transfac (http://www.gene-regulation.com/)
- TRRD (http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/trrdintro.html)

We limited our dataset to functionally relevant interaction types supported by 2 or more publications to obtain 22099 total interactions averaging 3.82 publications each. The included interaction types were transcription, phosphorylation, activation, cleavage, transport, dephosphorylation, degradation, and modulation. The data also included the direction of the interaction and the effect of the source gene on the target gene, which could be either positive, negative, or unknown.

Constructing the network

All final module pairs were examined for interactions present in our PPI dataset. Module pairs where at least one gene in one module interacted with at least one gene from the other module were connected by an edge.

RESULTS

Modules

CLOSET+ produced 258982 module candidates in under 21 seconds on a 1.5 GHz Pentium Centrino machine with 512 MB of RAM, and after removing all modules containing only one gene, and all modules supported by only one pathway data source, we were left with 227961 modules. Our randomization procedure next produced 197542 significant modules, and our ranking and selection process reduced the number to 623 functional modules.

Around two-thirds of our final modules were two-gene modules (see Supplementary Table 1). This is not surprising because it is more likely for fewer numbers of genes to be found together, so smaller modules tend to have larger supports and thus are more likely to be chosen by our ranking and selection system. However, there were also quite a few three-, four- and five-gene modules, indicating that our procedure is not limited to finding only small modules. Our largest module contained 22 genes and corresponded to an NADH dehydrogenase complex, which had a support of five and came from three pathway data sources.

Table 1 displays a few examples of the modules that we discovered, along with broad functional annotations. Of interest are the modules MAPK1|MAPK3 and MAP2K1|MAP2K2. It

is not surprising that these modules, so important to cell signaling, were included in our list with very large support levels and supported by nearly every data source we used. Note that the kinases (MAPK) and the kinase-kinases (MAP2K) are found in separate modules, indicating the strong resolving power of our approach. Finally, sets of genes that have been well-characterized to function together are also found in our list, such as the BCL2L1|BID|CASP|CYCS module involved in apoptosis, and the ADCY1|GNAS1|GNB1|GNGT1 module involved in G-protein signaling.

Network

Our final network contained 532 nodes and 6611 edges, involving 1289 of the 8422 unique genes from our original pathway dataset and covering 4410 of the 5426 original pathways (in other words, at least 4410 pathways contained at least one gene of the 1289 genes in the network). The module IL1B|TNF, with a support of 247 coming from 5 pathway sources, was the highest-degree node in our network, connected to 236 other modules. MAPK1|MAPK3 was the second-most connected module, with 182 connections, and IFNG|IL10|IL4 (support of 80, from 5 pathway sources) rounded out the top three network hubs with a degree of 178. These three modules are involved in many aspects in inflammation, cell signaling, and immune response, respectively, and thus it makes sense that they are so well-connected.

Because our PPI data included information about the direction of the interaction as well as the type of effect (positive or negative), different edges in our network had different charac-

ters. Analysis of these edges may help us understand how two modules functionally influence each other and generate testable hypotheses for follow-up experiments. Figure 1a illustrates some examples of different edge types. Some modules can be linked by positive feedback loops, as in i). REN is an activator of AGT, so the edge seen in ii) encapsulates a negative feedback loop. Finally, since CDKN1A and CDKN1B are both inhibitors of CDK4, the edge in iii) is actually a unidirectional edge, with an overall inhibitory effect on the cell cycle control module.

The value of a module network, however, lies in its ability to depict the patterns of interconnectedness of not just two, but many, modules. Figure 1b gives one example of such a pattern. DNA replication is initiated after entry into S phase and the MCM2|MCM4|MCM6|MCM7 module plays an essential role during replication initiation. Two modules, namely CCNE1|CDK2|E2F1|RB1|TFDP1 and E2F1|E2F2|E2F3|E2F5, which are well known to promote entry into S phase, are shown here to activate the MCM module. On the other hand, DNA replication must not occur after entry into mitosis, and so the CDC2|CDC25A|CDC25B|CDC25C module, a master regulator of mitosis, acts to inhibit the MCM module.

CONCLUSION AND DISCUSSION

We have presented a new approach to extracting functional modules and constructing a network of these modules. Our method is based on human-curated pathway and PPI data and has advantages over the current expression data- and PPI data-based methods: we can capture modules based on functional relationships other than coexpression, and we avoid modules that cannot

actually exist *in vivo*. We have found many modules that make biological sense (Table 1) and we have demonstrated the usefulness of a network representation of the modules (Figures 1a and 1b).

We believe that the modules we found are likely to be functionally relevant because the pathway data that we used are human-curated, and the curators can use their expertise to group into a pathway genes that would not be grouped by computational methods. One potential limitation to this approach is that the curators can make mistakes, so we do not rely merely on one source for pathway data (we used eight sources in this report).

Our module selection procedure itself, however, needs improvement, for it is very difficult to operationalize the selection of "good" modules. While the consensus among biologists is that modules are real, most don't agree on what actually constitutes a module (Wolf and Arkin, 2003). Modules must have "separable functions," but the difficulty is that function is context-specific. There have been efforts to deal with this issue (Huang *et al*., 2007) but no definite rules have emerged. We may benefit from some computational techniques meant to deal with uncertainty, such as fuzzy frequent pattern mining techniques (Hüllermeier, 2005) or itemset summarization techniques (Yan *et al*., 2005).

Despite these limitations, the modules extracted by our approach can aid biologists in making sense of the large amounts of information often produced by high-throughput experiments. Mapping this information onto functional modules rather than large pathways makes

high-throughput experiments easier to understand. Furthermore, our representation of the modules in a functional network can help biologists trace the transfer and integration of information and interaction between and among modules, and can lead to experimentally verifiable hypotheses. Cell biology is a complex phenomenon in which the interactions of modules give rise to a whole that is greater than the sum of its parts. Mapping and understanding those interactions is a step closer toward a complete understanding of the cell itself.

ACKNOWLEDGEMENTS

We thank Liwen Liu for providing the pathway datasets, and Dilip Rajagopalan, Pankaj Agarwal, Michael Lutz, and David Searls for their encouragement, support, and comments on the manuscript.

Conflict of Interest: None declared.

REFERENCES

Barabási, A.-L. and Oltvai, Z.N. (2004) Network Biology: Understanding the cell's functional organization. *Nat*. *Rev*. *Genet*., **5**, 101-113.

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

Lee, I. *et al*. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555-1558.

Hartwell, L.H. *et al*. (1999) From molecular to modular cell biology. *Nature*, **402**, C47-C52.

Huang, Y. *et al*. (2007) Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics*, **23**, i222-i229.

Hüllermeier, E. (2005) Fuzzy methods in machine learning and data mining: status and prospects. *Fuzzy Sets and Systems*, **156**, 387-406.

Okada, Y. *et al*. (2007) A biclustering method for gene expression module discovery using a closed itemset enumeration algorithm. *IPSJ Trans*. *on Bioinformatics*, **48**, 39-48.

Pereira-Leal, J.B. *et al*. (2004) Detection of functional modules from protein interaction networks. *Nucleic Acids Res*., **54**, 49-57.

Prelić, A. *et al*. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122-1129.

Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA*, **100**, 1128-1133.

Segal, E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*., **34**, 166-176.

Segal, E. *et al*. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet*., **36**, 1090-1098.

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

Voy, B.H. *et al*. (2006) Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PloS Comput. Biol*., **2**, 757-768.

Wang, J. *et al*. (2003) CLOSET+: Searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the 9th ACM SIGKDD International Conference On Knowledge Discovery and Data Mining*.

Wolf, D.M. and Arkin, A.P. (2003) Motifs, modules and games in bacteria. *Curr. Opin. Microbiol*., **6**, 125-134.

Yahia, S.B. *et al*. (2006). Frequent closed itemset based algorithms: a thorough structural and analytical survey. *SIGKDD Explorations*, **8**, 93-104.

Yan, X. *et al*. (2005) Summarizing itemset patterns: a profile-based approach. In *Proceedings of*

the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

FIGURES

Table 1. Examples of discovered modules

Genes (support, sources)	*Involved in*
BCL2L1, BID, CASP9, CYCS *(16, 5)*	Apoptosis
MAPK1, MAPK3 *(899, 7)*	Cell signaling
MAP2K1, MAP2K2 *(213, 6)*	Cell signaling
MMP2, MMP9, TIMP1, TIMP2 *(17, 5)*	Maintenance of ECM
MCM2, MCM4, MCM6, MCM7 *(8, 5)*	DNA replication
ADCY1, GNAS1, GNB1, GNGT1 *(19, 3)*	G-protein signaling
C1S, C2, C3, C4B, C5, C6, C7, C9 *(6, 4)*	Blood coagulation
MYH1, MYH4, MYH6, MYH7, MYH8, MYL1, MYL3, MYL9 *(5, 3)*	Cytoskeleton
CCND1, CDK2, CDK4, CDKN1A, CDKN1B *(26, 5)*	G1/S phase transition

Selected examples of functional modules. The support, the number of data sources, and a broad functional annotation are given for each module.

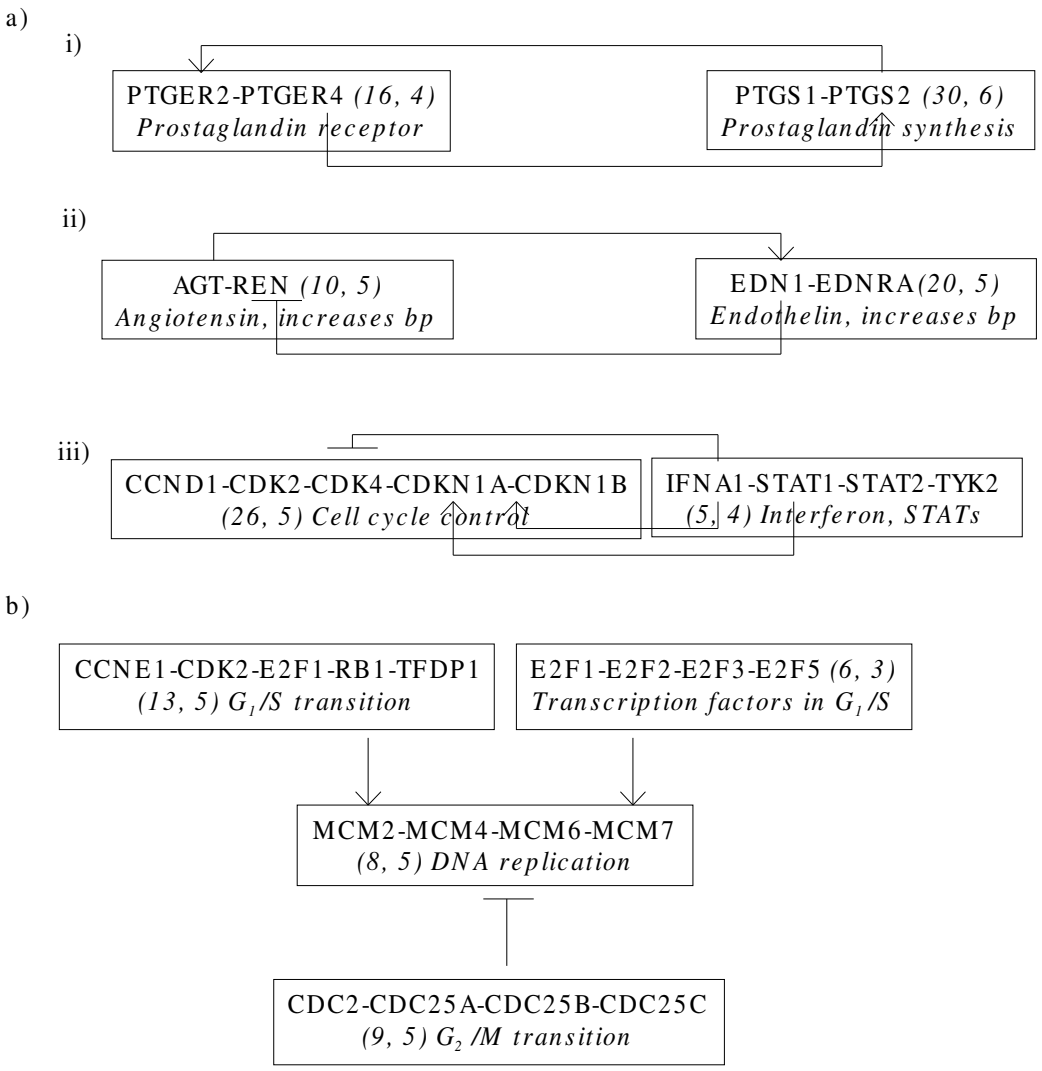

Figure 1. A deeper look into the module network.

a) Example of the different types of edges that can exist between modules. Lines represent interactions between genes. Lines ending in arrows signify activation, and lines ending in bars signify inhibition. **b)** Example of information integration between and among modules. Lines represent interactions between modules. Support level and number of data sources are shown in parentheses after each module followed by a functional annotation of that module.

Supplementary Table 1. Frequency of module sizes, with supports

Size	Frequency	Mean Support	Median Support
2	454	15.13	7
3	76	11.36	7
4	64	11.09	7
5	13	16.31	8
6	6	8.83	7.5
7	2	25	25
8	3	5.67	6
9	1	6	6
12	1	6	6
13	2	6.5	6.5
22	1	5	5

Characteristics of the final list of functional modules. Module size is equivalent to the number of genes in the module. The number of modules of any given size is shown, as well as the average and median supports for modules of that size.