Jennifer A. Chen
Biochem 218 Final Project

Biological network analysis by division into modules, motifs, and themes based on physical and functional interactions between genes

Introduction

The complex system of the cell is a network of interconnections - proteins interact with other proteins or with DNA, and genes can interact functionally with one another. Cellular networks can be broken down into simpler components to facilitate their study. Three different types of analysis – modules, motifs, and network themes – that can be utilized to simplify the intricate web of biological relationships will be discussed. How these analyses are carried out, what type of information they show, what conclusions can be drawn from them, how these methods can vary from study to study, and what kinds of limitations they have will be addressed. The information that can be gained from the study of the individual components and patterns within a network may be extrapolated and provide greater insight about the network as a whole.

Delineating modules for the study of a biological network

Networks can be utilized to describe global interaction data in an organism, with the nodes representing genes and their links, also known as 'edges', describing some form of interaction between them. Since it is extremely difficult to analyze the underlying mechanism of a biological network as a whole, it is advantageous to perform an algorithm for modularization of the network. Modularization is a process which divides a network into smaller units for better understanding and analysis of the entire network. Certain criteria are used to define a module, which is simply a subset of the original biochemical network that tends to have minimal dependency on the rest of the network. Unlike studying the network as a whole, this enables the easier study of individual and somewhat independent modules in order to gain more insight on the entire network.

Popular existing methods for partitioning biological networks include the graph partitioning and community structure detection techniques (1, 2, 9-11). The graph partitioning technique divides a set of tasks among the processors of a parallel computer to minimize the necessary amount of interprocessor communication. The number and size of the groups into which the network is to be split is based upon the number of processors, along with an approximate figure of the number of tasks that each processor can handle. As a result, the best

division of the network is determined without addressing whether a good division even exists. Community structure detection, on the other hand, assumes that the network under study divides naturally into subgroups, and the experimenter's job is to find those groups. This may be considered more ideal than the graph partitioning approach since the number and size of the subgroups are determined by the network itself and not by the experimenter.

Utilization of an algorithm to specify modularization of a signaling pathway

A more novel method of network partitioning is described by Losiana *et al.*, who propose an algorithm for modularization of MAPK and calcium signaling pathways through comparative analysis among different species (7, 8). The algorithm starts with detection of a node *n* having maximum number of relations in the node pool *E* for a given network. The module then grows in size by including immediate neighbors of the starting member in successive steps. Once a module is initialized, the total number of relations (T_n) of every individual member is considered. The neighbors are either included into or excluded from the module depending on the number of their relations being present inside or outside the module. A complexity level *c*, is used to determine whether or not a member is included in the module – a member with less than or equal to a certain number of relations outside the module, termed the complexity level *c*, is included in the module. The complexity level is a term that is specified and varied by the user. For a node in a module, if the number of relations lying inside the module is equal to the total number of relations associated with the node, the member is considered to be *permanent*. If a node in a module has more than *c* relations that lie outside the module, it gets *excluded* from the module. These *extension* and *exclusion* processes continue until there are no new immediate neighboring nodes to be included, or no node is left to be declared permanent. Once a member of declared permanent, it gets removed from the node pool *E* in order to avoid the chance of a single member to be included in more than one module. After successful completion of the creation of a module, the algorithm will search for another starting point and repeat the previous steps to create another module. The process of creating modules will continue until all of the nodes present in the node pool *E* have been exhausted. Figure 1 describes the module construction process.

The proposed modularizing algorithm was applied to calcium signaling pathways and MAPK signaling pathways belonging to various species. Modules were created from both

calcium and MAPK signaling pathways of *H. sapiens* for different values of *c*. The calcium signaling pathway of *H. sapiens* contains 55 nodes, one of which is isolated. These 54 nodes having 59 relations among them is depicted in Figure 2.

Modules were created from the same pathway for complexity level (*c*) of 1, 2, 3, 4, 5, 6, 7 and above. A trend was observed in that the higher the complexity level, the fewer the number of modules were obtained, which is to be expected since the majority of genes tend to have fewer interactions. For example, for $c = 1$, 11 modules were obtained, whereas for $c = 7$, the whole network emerges into a single module. Once biologically significant modules at some value of *c* were obtained, the question of whether to proceed further and continue modularization at higher values of *c* to get more meaningful modules or stop the process was addressed. Modularization for higher values of *c* caused modules to increase by several nodes and relations, making the modules large and complex, which goes against the primary objective of dividing a complex network to simpler units. Therefore, it can be assumed that after a certain level, modularization with increasing *c*-value will yield similar results with that of previous complexity level or the modules will be so much larger that their study and analysis will be difficult. Since approximately biologically significant modules were obtained for $c = 3$, the *c*-value of 3 was fixed to study all signaling pathways in the different species. However, the authors failed to address how the biological significance for a specific *c* value is determined, which leads the reader to believe that it may be determined arbitrarily and may be a source of variation or error. Another limitation of the algorithm is that it is based on the database from which the data is taken. The same algorithm applied to different databases may result in the creation of different modules.

In certain species when only part of the pathway is functional, modularized study is extremely helpful. For instance, one module of the human calcium signaling pathway was consistent among other species studied, while other modules were not present in certain species. Therefore, the other modules can be avoided and the consistent module of human calcium signaling pathway can be compared with that of the other species instead of comparing the whole pathway. When analyzing a very large network, these types of inferences may save time and cost of wet lab experiments by avoiding less important verifications. However, a downside to the limitation of studying a single module for comparative analysis across species is that other modules that may provide additional insight for the entire network are neglected. In addition,

certain genes that are essential to a subnetwork may be excluded from a module because they lack a sufficient number of interactions with the genes within the module. Additionally, the different types of interactions between nodes may be important when deciding which genes to include in a module. In general, the creation of modules within a network is not extremely informative. However, modules facilitate easier analysis and allow for further study of the specific components of the module through the comparison of modules consistent in different species.

Conserved network motifs describe protein-protein interaction

A more informative method of analyzing a biological network is the identification of patterns that describe protein interactions. Network motifs have been developed to describe simple patterns of interconnection in networks that occur more frequently than expected when biological networks are compared to randomized networks (6, 12). Motifs can be considered as elementary building blocks of the modules that compose complex networks. In general, a motif describes the physical interaction between genes, which include interactions among transcription factors and their targets, as well as protein-protein interactions.

An algorithm is used to detect network motifs, which are recognized as recurring, significant patterns of interconnections. The network is first scanned for all possible *n*-node subgraphs, or any pattern of interaction between *n* genes, and the number of occurrences of each subgraph is noted. Since each network contains numerous types of *n*-node subgraphs (Fig. 3), the real network is compared to suitably randomized networks and only patterns appearing in the real network at numbers significantly higher than those in the randomized networks are selected (Fig. 4). In order to determine the statistical significance of the patterns found, randomized networks that have the same single-node characteristics as the real network are created. Specifically, each node in the randomized networks has the same number of incoming and outgoing edges as the corresponding node as in the real network. The study of motifs as described by Milo *et al*. determined that the concentration of motifs in subnetworks, pieces of various sizes of the full network, is about the same as that in the full network. Therefore, the study of motifs in individual modules may be both informative and relative to the entire network. However, as shown in the study, it may be more rewarding to analyze larger subnetworks, since the larger the network is, the more significant the motifs tend to become.

Whether motifs in the interaction network provides any useful information to the organization of cellular interactions is oftentimes addressed. Several groups have examined the role of motifs by attempting to describe their biological functions as well as their existence based on evolutionary selective pressure. A well-known indicator of the conservation of specific cellular functions is the evolutionary retention of orthologous proteins that are responsible for performing similar functions. The tendency for motifs, or their specific components, to conserve evolutionarily is therefore indicative of the importance and involvement of motifs in specific biological functions.

To determine the presence of any special evolutionary pressure acting to preserve motifs, Mazurie *et al.* performed a protein comparative analysis between *Saccharomyces cerevisiae* and four hemiascomycetes, which were chosen based on sharing many functional similarities with *S. cerevisiae* yet spanning a broad range of evolutionary distances (5). The transcriptional regulatory network constructed and investigated by Guelzim *et al*. (3) and the protein-protein interaction data in the Database of Interacting Protins (DIP) was utilized to construct a network with 476 nodes, 905 directed transcriptional edges, and 221 undirected protein-protein edges.

Motifs within the different organisms were compared through the quantification of a value, termed the evolutionary fragility, based on the number of the organisms in which a particular motif is found. A small value for the fragility indicated that the genes composing the motif tended to co-occur in the other compared organisms, suggesting an evolutionary pressure to preserve the motif based on functional importance. This may not be the best method to quantify evolutionary conservation, however, because the organisms analyzed were closely related as members of the same class within fungi. The study determined that although there was a higher abundance of genes that have a low evolutionary fragility (suggesting high pressure for co-evolution) in the enriched motifs, there was no statistically significant difference in evolutionary co-existance between enriched motifs and all the interaction patterns found (Fig. 5). However, I do not feel confident about this conclusion and the interpretation of evolutionary fragility- the motifs composed of genes taken at random were not equally distributed when quantified through evolutionary fragility and also had fairly high levels of abundance.

The possible role of motifs in biological functions was also analyzed through the comparison of known biological information (Fig. 6). According to Mazurie *et al.*, the roles of the motifs identified are deemed not evident because the current model of the pathway can be

described without any reference to them. For example, the three proteins Cbf1p, Met4p, and Met28p always act as a complex. However, this information does not emerge when analyzing the topology of the network, since the topology is also compatible with the three proteins acting separately. In general, the motifs were described as insufficient, misleading, and failed to capture the complexity of the interwoven interactions.

The findings of another group, Wuchty *et al*. (13), show that in *Saccharomyces cerevisiae*, proteins organized in motifs are conserved to a substantially higher degree than those that do not form motifs. The conservation of 678 *S. cerevisiae* proteins with an ortholog in each of five higher eukaryotes (*A. thaliana, C. elegans, D. melanogaster, M. musculus* and *H. sapiens*) was studied out of the total 3,183 proteins composing the yeast protein interaction network. This differs from the Mazurie *et al.* study since a substantially larger number of proteins were analyzed and the species compared are more diverse and complex. In addition, the definition of a conserved motif required each of its protein components to have an ortholog in each of the five higher eukaryotes. Substantially different conservation rates for proteins in the different motifs were found (Table 1). A random ortholog distribution was tested to determine the validity of the findings, which showed very low conservation rates compared to that of the natural system. The fraction of the original yeast motifs that is evolutionarily fully conserved, shown in the last column of Table 1, was determined by calculating the ratio between the real and the random conservation rates. The conservation ratio for each motif was greater than one and increased considerably for larger motifs. Larger motifs tended to be conserved as a whole, with each of their components having an ortholog. In general, as the number of nodes in a motif and number of links among its constituents increased, the evolutionary retention of the constituent proteins was more complete. The exceptionally high conservation rates strongly suggest that participation in motifs substantially influences the evolutionary conservation of the specific components.

These findings contrast from those of Mazurie *et al.*, probably because the previous study performed the evolutionary conservation analysis differently. The network motifs were categorized based on the rate of evolutionary conservation but the different types of network motifs were not distinguished. In this study, the natural conservation rate is compared to the random conservation rate for each specific type of network motif. In addition, the rates of random conservation are extremely low, as should be expected for a control.

In order to examine if the specific function of the yeast proteins within motifs affects their rate of evolutionary conservation, Wuchty *et al.* assigned each previously identified conserved yeast motif orthologous to humans to the functional class to which its protein components belong (Table 2). Larger motifs seemed to have more functional homogeneity. This seems logical since larger motifs were found to have higher rates of evolutional conservation, and motifs that are evolutionarily conserved would also share the same functional class, since orthologous genes often have similar functions. This is a more generalized analysis than Mazurie *et al.* since motif functions were classified into very broad groups, such as cell cycle, transcription, and protein synthesis, instead of looking to see whether each motif would provide detailed insight into its specific function within its subnetwork. Also, the motifs that were classified into different functions in yeast consisted of the subset of proteins that had an ortholog in humans, and so it would be more likely that a high rate of functional conservation would be observed. Wuchty *et al*. claim to observe substantial functional class-dependent differences in the evolutionary conservation of motifs. For three functional classes (subcellular localization, protein fate, and transcription), each of the 11 studied motifs were considerably overrepresented, whereas few to no functional classes had only one or two characteristic motifs. I would argue that because only three functional classes had only one or two characteristic motifs (transport facilitation, regulation, and cellular transport), this would reflect that the specific function of the yeast proteins within motifs does not affect their rate of evolutionary conservation. It is also important to realize that overrepresented motifs were not found for many important classes, including energy, cellular fate, cellular communication, cellular organization, metabolism, and protein binding.

Motifs may allow greater insight into the larger network in which they are contained. For example, fully connected motifs tend to identify protein complexes. Large numbers of interactions with uncharacterized proteins may indicate functional relation, suggesting that specific motifs could be used to predict the functional role of the unknown protein components. On the other hand, in larger protein complexes, not all proteins have direct interactions with each other, and thus motifs capture only some local, physically interacting components of the entire complex. Transitory macromolecular associations like protein complexes and interactions between a whole protein complex and a target are oftentimes missed and may be represented as individual links between each component and the target. Furthermore, the use of networks to

describe biological interactions may be misleading since it ignores important aspects of detailed biological dynamics, such as localization in both space and time, protein modifications, and the formation of multimeric complexes.

Databases may also incorrectly represent the biological context of gene interactions. The quality of the motifs identified is determined by the quality of results from the databases used. In general, how the motifs were identified in the two studies may influence the types and numbers of motifs that were analyzed. Patterns that are functionally important but not statistically significant may be overlooked, and motifs consisting of higher number of nodes than the ones studied may also be informative but not included. Importantly, the great majority of motifs are embedded in larger networks and are likely to perform a specific functional task depending on the context.

Reconstruction of network pathways into network themes based on direct and indirect gene interactions

Besides the physical interactions involved in signaling between biomolecules, functional interactions should also be taken into consideration (4). For certain types of networks, such as well-characterized signaling pathways, organizational principles can be determined through a set of well-established computational methods in an unbiased fashion, as described for module specification and motif recognition. However, there is a need to develop methods for discovering organizational principles of integrated networks that combine different types of interactions between genes. In addition to conventional protein-protein interaction data, a number of approaches have been developed for identifying co-regulated gene modules. Certain types of interaction information, such as genetic interactions obtained from synthetic lethality screens, do not necessarily indicate direct physical interactions between gene products. Links between two genes can also be derived from computational analysis of datasets that include information about sequence similarity or the level of gene expression under specific conditions.

The study by Zhang *et al.* presents an approach for integrating multiple types of biological interactions (14). The authors simultaneously use five different yeast datasets that include both direct physical interactions (protein complexes and transcriptional regulation) and indirect functional interactions (genetic interactions, gene-expression correlation, and sequence homology). The overall approach used in this work first decomposes the combined interaction

network into multi-color *network motifs*, in which each color corresponds to one type of interaction data, and then assembles these motifs into *network themes* consisting of overlapping motifs. Although network motifs are considered the building blocks of a biological network, they may not necessarily correspond to functional building blocks of the actual networks inside a cell. The network theme describes recurring higher-order interconnection patterns that encompass different network motifs. This reflects a common organizational principle of the same type that can also be tied to specific biological phenomena.

In the study, 12 three-node network themes were identified as enriched, 7 of which described known biological phenomena. For example, multiple motifs containing a transcription factor that regulates two physically interacting proteins can be combined into a single theme corresponding to a protein complex whose component proteins are controlled by the same transcription factor (Fig. 7, a-c). Other types of themes involve genetic interactions. One of these includes the 'alternative subunits' theme, which describes two genes connected to each other by synthetic lethal interactions and to other members of the complex by protein-protein interactions. A second genetic-interaction-based theme, the 'compensatory complex' theme, consists of two protein complexes internally connected by protein-protein interactions that are bridged by a large number of genetic interactions (Fig 7, d-f). This theme indicates a structure in which either of the complexes is needed to perform an essential function, but the complexes can compensate for one another.

An approach that integrates a more diverse set of interaction data, as described in Zhang *et al.*, is advantageous since it may discover network themes that have weaker support from only one type of data such as protein-protein interactions. However, the network theme that is constructed may vary depending on the computational method used and the databases from which the physical and genetic interactions are derived. Also, only a subset of total information about a specific type of interaction may be available. For example, at the time of the Zhang *et al.* study, only 4% of yeast gene pairs had been examined for synthetic genetic interactions. Many potential network themes may be overlooked simply because they fail to achieve statistical significance. Another limitation is that the approach is applied to a static interaction network, whereas in reality only subsets of interactions are active under any particular biological condition. Analysis should therefore be extended to condition-dependent network structures, constructed by combining gene expression and physical interaction data. Overall, the network

theme is an excellent approach that is able to address the limitations of more basic forms of network analysis, specifically through the inclusion of indirect gene interactions and the integration of multiple types of interactions.

Conclusion

Through the breakdown of the complex cellular network into smaller components, the physical and functional relationships among genes, proteins, and other macromolecules can be studied. The organization of networks into modules, motifs, and themes was discussed with respect to the types of information they can provide in order to gain more insight about the network as a whole. Modularization is a process that divides a network into self-sufficient units which facilitates analysis of its specific components. The identification of similar modules in different species may allow the function of similar proteins within the module to be extrapolated. Network motifs are overrepresented patterns that describe protein interactions. Whether motifs play a role in specific biological function and contribute to evolutionary selective pressure of the proteins within the motif is up to debate. Network themes, which describe higher-order interconnection patterns that encompass different types of network motifs, are a more comprehensive and integrative analysis of the different types of physical and functional interactions that occur between genes in a network. Overall, these different methods of reducing a complex network into smaller components can provide information that can be tested experimentally in order to gain greater insight about how the network is able to function as a whole.

FIGURES

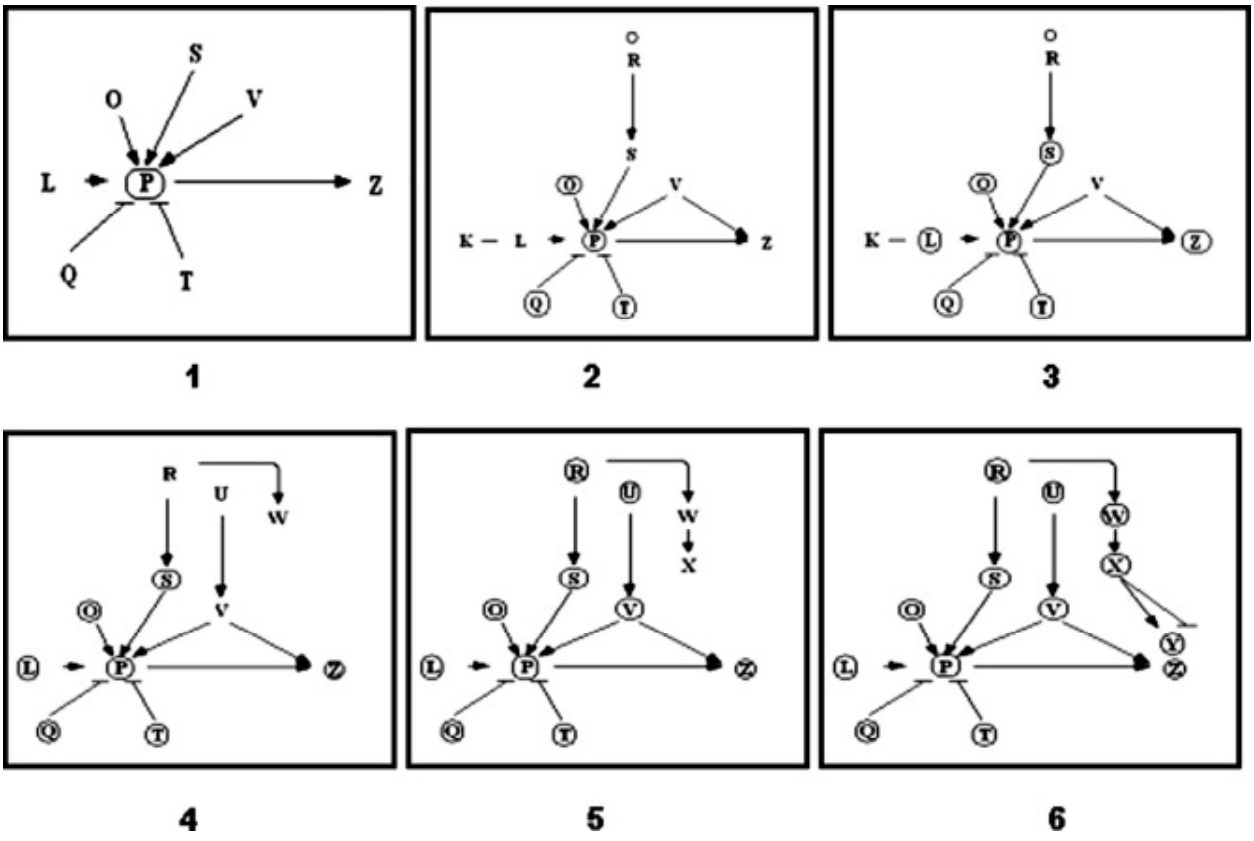


Figure 1. Stages in construction of a module. This figure gives stepwise construction of a module for c = 2. After each extension, nodes having all their relations inside the module are declared permanent. Nodes having more than two out-relations are excluded from the expanding module, and the rest are taken as possible members under consideration and their immediate neighbors are included during the next phase of extension. (7, 8)

Figure 2. KGML layout for calcium signaling pathway of H. sapiens. (7, 8)

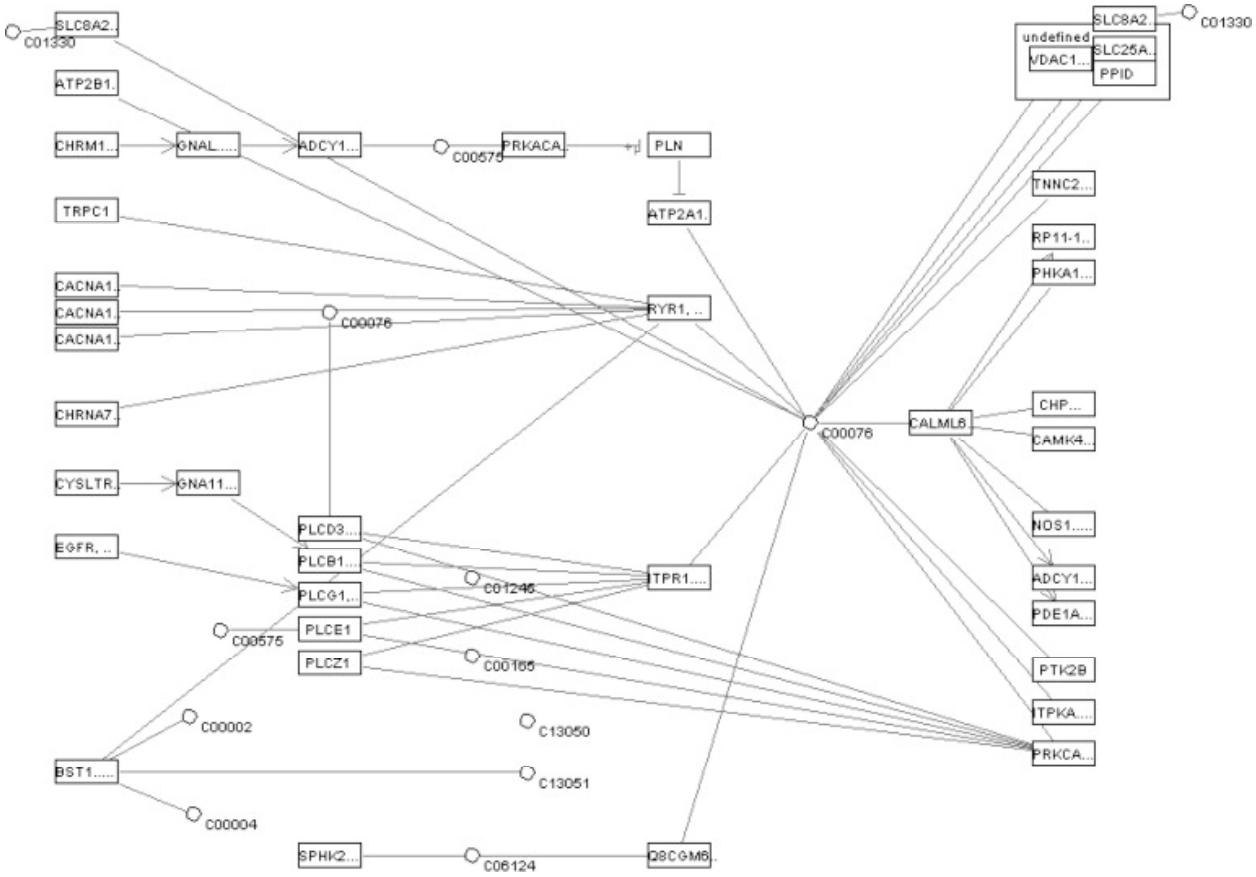


Figure 3. All 13 types of three-node connected subgraphs. (6)

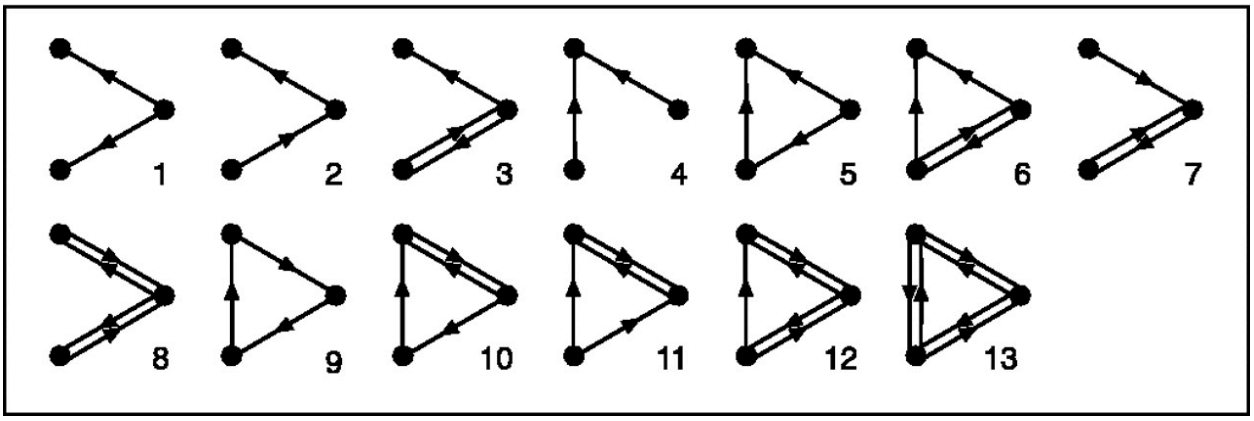


Figure 4. Schematic view of network motif detection. Network motifs are patterns that recur much more frequently (A) in the real network than (B) in an ensemble of randomized networks. Each node in the randomized networks has the same number of incoming and outgoing edges as does the corresponding node in the real network. Red dashed lines indicate edges that participate in a specific motif, depicted in the bottom left corner of (A), which occurs five times in the real network. (6)

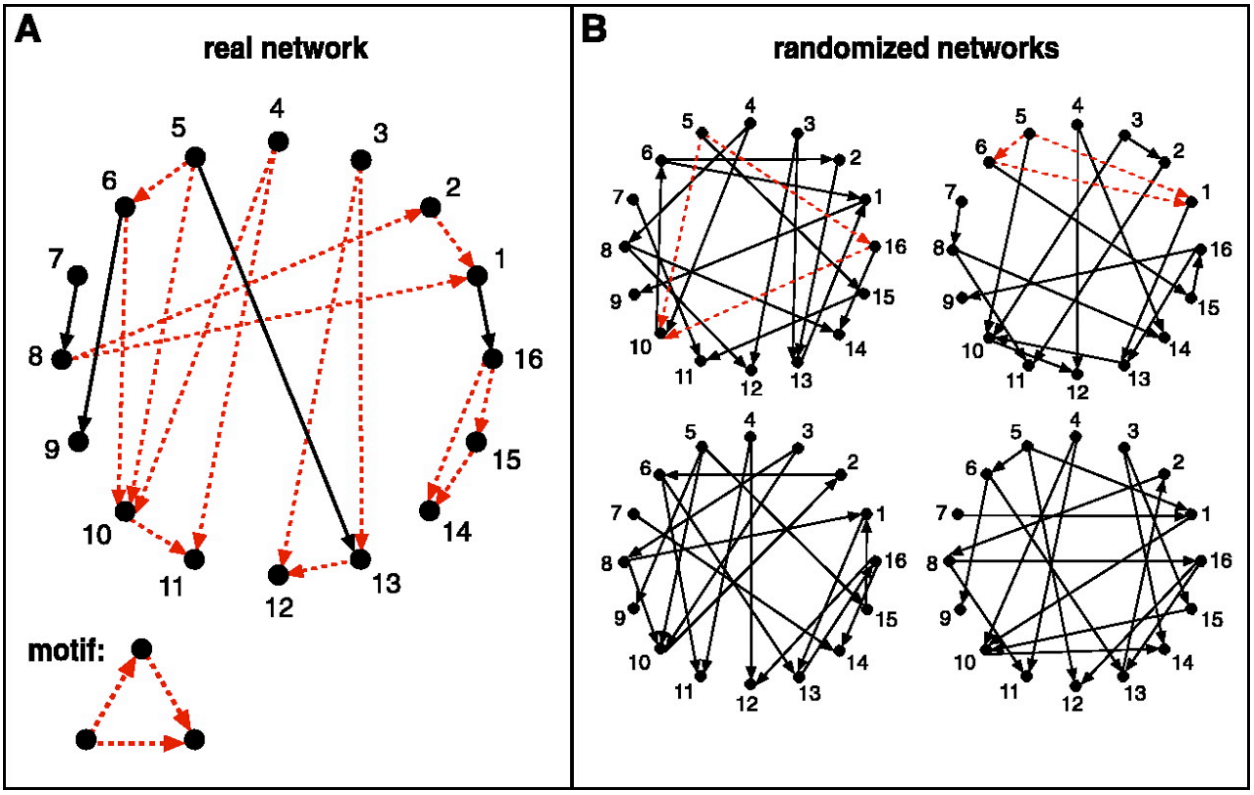


Figure 5. Categorization of evolutionary fragility among different interaction patterns. Histograms of the evolutionary fragility of interaction patterns belonging to the following three classes are shown: instances of network motifs (red); generic patterns of interacting genes, irrespective of their abundance (black); patterns composed of genes taken at random (white). The five possible values (in increasing value 0 to 4) of the evolutionary fragility are reported on the x-axis. A small fragility value indicates that all the genes composing the interaction patterns tend to co-occur in the other genomes compared and point to evolutionary pressure acting to preserve the interaction pattern. (5)

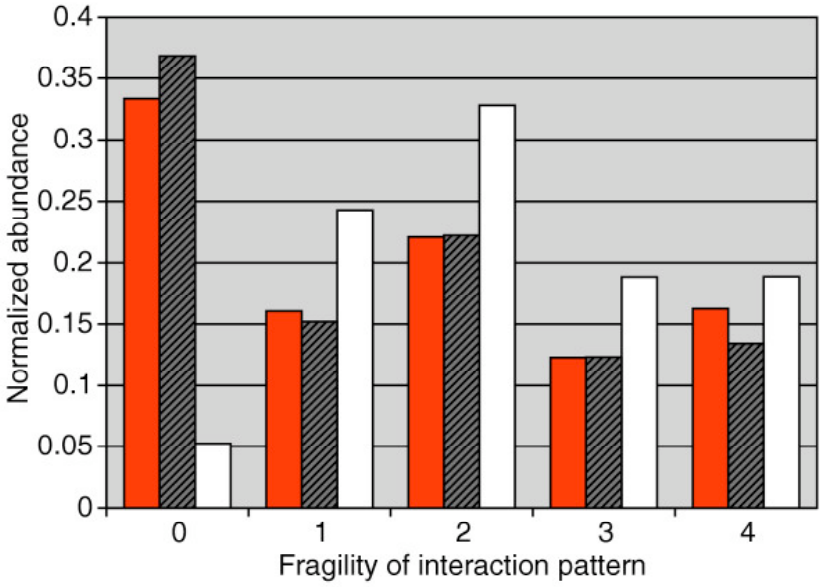


Figure 6. Comparison of motifs to known biological pathways.
(a) Methionine (MET)
(b) nitrogen catabolite repression (NCR)
(c) pseudohypal growth/mating (HYPE)
(d) regulation of early meiotic genes (CCYCLE)
(e) pleiotropic drug resistance (PDR)
Subnetworks with the identified motifs within the pathway drawn from the interaction databases are shown on the left. A schematic representation of the regulation mechanisms for the same pathways, based on the present experimental knowledge, is shown on the right. Full lines represent transcriptional regulation, dashed lines non-transcriptional regulation, and wavy lines transformations and syntheses. Arrowheads, positive regulation; lines ending in a terminal bar, negative regulation. (5)

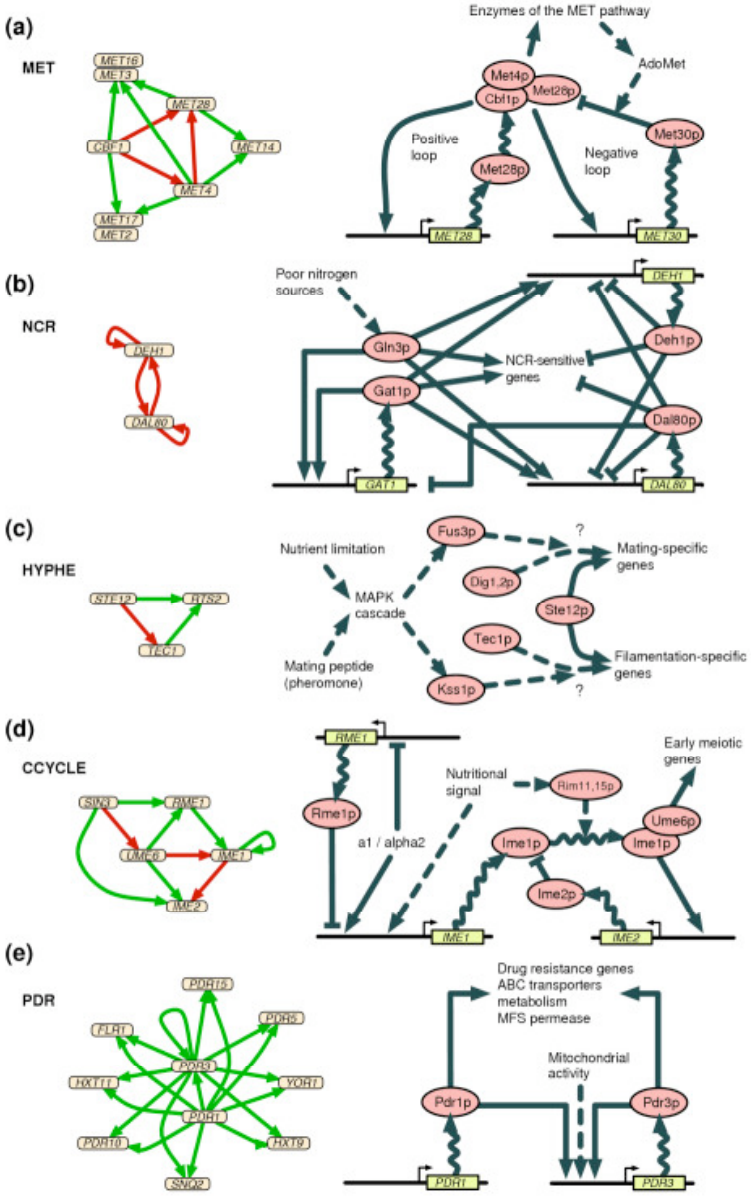
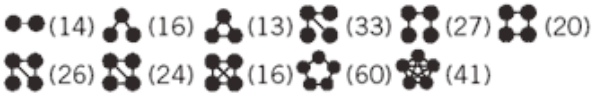
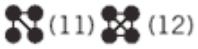
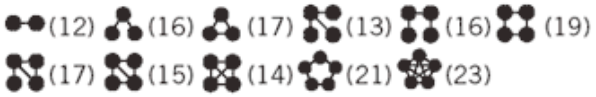


Table 1. Evolutionary conservation of motif constituents. (13)

#	Motifs	Number of yeast motifs	Natural conservation rate	Random conservation rate	Conservation ratio
1		9,266	13.67%	4.63%	2.94
2		167,304	4.99%	0.81%	6.15
3		3,846	20.51%	1.01%	20.28
4		3,649,591	0.73%	0.12%	5.87
5		1,763,891	2.64%	0.18%	14.67
6		9,646	6.71%	0.17%	40.44
7		164,075	7.67%	0.17%	45.56
8		12,423	18.68%	0.12%	157.89
9		2,339	32.53%	0.08%	422.78
10		25,749	14.77%	0.05%	279.71
11		1,433	47.24%	0.02%	2,256.67

The third column gives the number of motifs of a given type found in the yeast protein interaction network of 3,183 proteins. 678 proteins were identified that have an ortholog in each of the five higher eukaryotes studied. In this subset, all motifs for which each component belong to were identified. The natural conservation rate indicates the fraction of the original yeast motifs that is evolutionarily fully conserved. The conservation ratio is determined by comparing the natural conservation rate to the random conservation rate for each type of motif.

Table 2. Overrepresentation of human orthologous motifs in various functional classes of yeast proteins. (13)

Functional class	Overrepresented motifs
Transport facilitation	 (10)
Subcellular localization	 (21) (21) (26) (15) (27) (23) (29) (20) (63) (45)
Regulation	 (10)
Protein fate	 (14) (16) (13) (33) (27) (20) (26) (24) (16) (60) (41)
Cell cycle	 (11) (14) (13) (11) (14)
Cellular transport	 (11) (12)
Transcription	 (12) (16) (17) (13) (16) (19) (17) (15) (14) (21) (23)
Protein synthesis	 (12) (11) (17) (11) (24)

The number of each characteristic motif belonging to a specific functional class was determined. All motifs that are overrepresented by a factor of at least ten compared with a random configuration is listed, with the specific Z value shown next to the motifs.

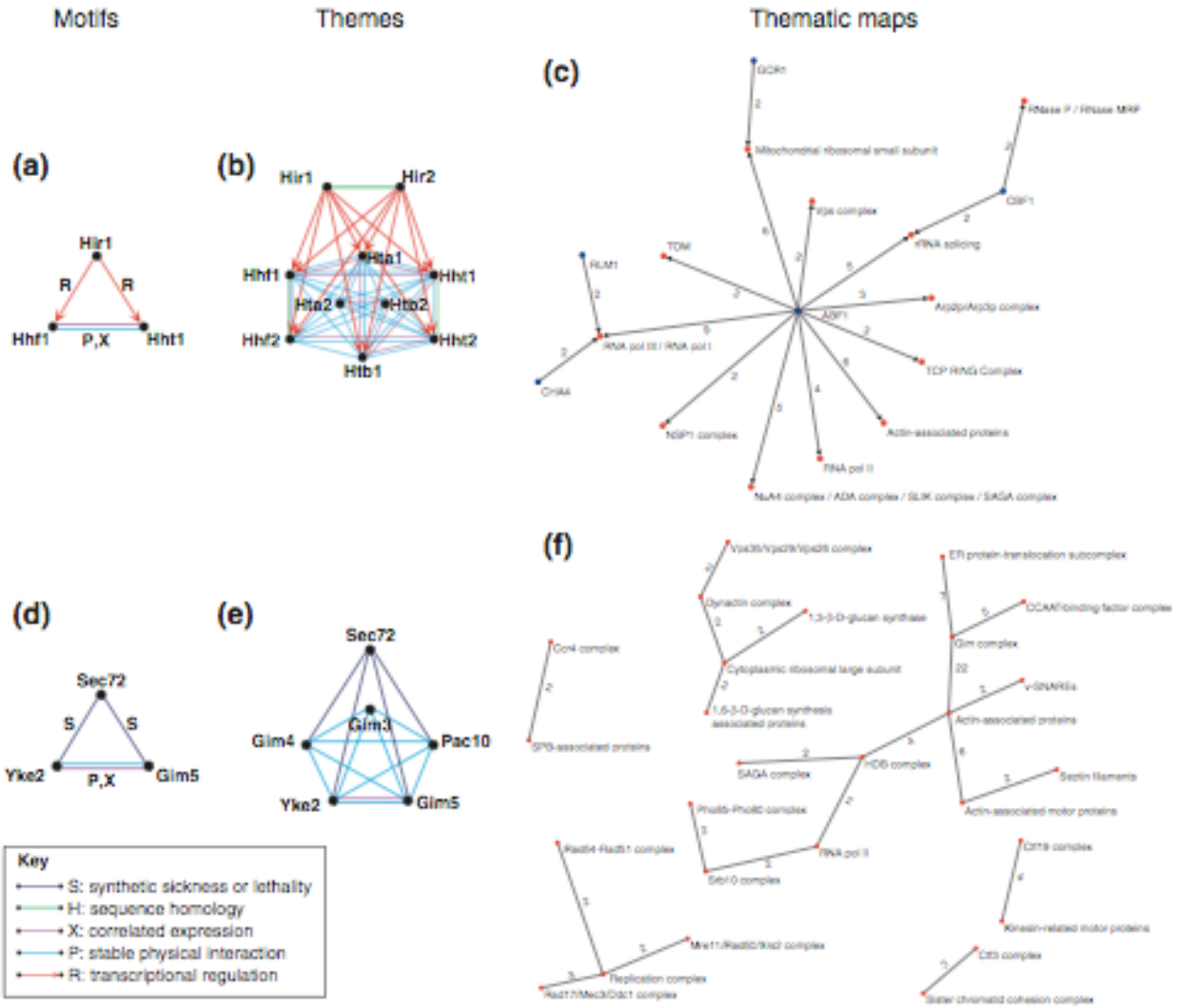


Figure 7. Examples of motifs, themes, and thematic maps in biological processes. (a-c) A theme capturing the co-regulation of members of a protein complex by a pair of transcription factors. (d-f) The 'compensatory complex' theme, in which one complex can compensate for the other in function. (4, 14)

REFERENCES

1. **Elsner, U.** 1997. Graph partitioning: a survey. Tech. Rep **97**.
2. **Fjallstrom, P.** 1998. Algorithms for graph partitioning: a survey. Tech. Rep **98**.
3. **Guelzim, N., S. Bottani, P. Bourgine, and F. Kepes.** 2002. Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet **31:**60-3.
4. **Herrgard, M. J., and B. O. Palsson.** 2005. Untangling the web of functional and physical interactions in yeast. J Biol **4:**5.
5. **Mazurie, A., S. Bottani, and M. Vergassola.** 2005. An evolutionary and functional assessment of regulatory network motifs. Genome Biol **6:**R35.
6. **Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon.** 2002. Network motifs: simple building blocks of complex networks. Science **298:**824-7.
7. **Nayak, L., and R. K. De.** 2007. An algorithm for modularization of MAPK and calcium signaling pathways: comparative analysis among different species. J Biomed Inform **40:**726-49.
8. **Nayak, L., and R. K. De.** 2007. Modularized study of human calcium signalling pathway. J Biosci **32:**1009-17.
9. **Newman, M. E.** 2006. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E Stat Nonlin Soft Matter Phys **74:**036104.
10. **Newman, M. E.** 2006. Modularity and community structure in networks. Proc Natl Acad Sci U S A **103:**8577-82.
11. **Newman, M. E., and M. Girvan.** 2004. Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys **69:**026113.
12. **Shen-Orr, S. S., R. Milo, S. Mangan, and U. Alon.** 2002. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet **31:**64-8.
13. **Wuchty, S., Z. N. Oltvai, and A. L. Barabasi.** 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet **35:**176-9.
14. **Zhang, L. V., O. D. King, S. L. Wong, D. S. Goldberg, A. H. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F. P. Roth.** 2005. Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J Biol **4:**6.