Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data

Amos Tanay*, Roded Sharan†, Martin Kupiec‡, and Ron Shamir*§

*School of Computer Science and ‡Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv 69978, Israel; and †International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704

Communicated by Richard M. Karp, International Computer Science Institute, Berkeley, CA, December 25, 2003 (received for review July 1, 2003)

The dissection of complex biological systems is a challenging task, made difficult by the size of the underlying molecular network and the heterogeneous nature of the control mechanisms involved. Novel high-throughput techniques are generating massive data sets on various aspects of such systems. Here, we perform analysis of a highly diverse collection of genomewide data sets, including gene expression, protein interactions, growth phenotype data, and transcription factor binding, to reveal the modular organization of the yeast system. By integrating experimental data of heterogeneous sources and types, we are able to perform analysis on a much broader scope than previous studies. At the core of our methodology is the ability to identify modules, namely, groups of genes with statistically significant correlated behavior across diverse data sources. Numerous biological processes are revealed through these modules, which also obey global hierarchical organization. We use the identified modules to study the yeast transcriptional network and predict the function of >800 uncharacterized genes. Our analysis framework, SAMBA (Statistical-Algorithmic Method for Bicluster Analysis), enables the processing of current and future sources of biological information and is readily extendable to experimental techniques and higher organisms.

Modern experimental techniques in biology collect massive amounts of information on the behavior and interaction of thousands of genes and proteins across diverse conditions (1–7). These techniques are used to interrogate complex biological systems that use highly intricate regulatory mechanisms and control schemes. One cannot fully characterize such complex cellular systems by focusing completely on a single control mechanism, as measured by a single experimental technique. To gain deeper understanding of the systems, it is pertinent to analyze heterogeneous data sources in a truly integrated fashion and shape the analysis results into one body of knowledge.

The challenge of such analysis has become a major bottleneck in expanding our understanding of biology. In this study, we analyzed simultaneously a highly heterogeneous collection of experimental data, spanning many different aspects of biological regulation, including gene expression, protein interactions, phenotypic sensitivity, and transcription factor (TF) binding. The outcome of our analysis is a set of modules, defined as maximal groups of genes that manifest a unique, common behavior across a significant set of the experiments, reflecting a particular function shared by the proteins that encode these genes. As the experimental data we use are of different types and sources, the notion of a module is broad and covers different aspects of organized behavior in molecular networks. We have developed algorithms to uncover statistically significant modules in an unconstrained fashion, without making prior assumption on the organization of the modules in the system. This approach exposes global architectural properties of the molecular network and, at the same time, derives highly specific predictions on gene functions and relations. Previous works have shown modular organization in gene expression (8, 9) and hierarchical modular organization in metabolic pathways and protein networks (10, 11). Here, we provide evidence for hierarchical, modular organization of the global yeast system. We show that small modules can be clustered into supermodules, such that supermodules characterize common behavior of the smaller modules under specific conditions. We show that specific classes of genes (e.g., signaling and transport) form bridges among supermodules, whereas other classes are typically associated with one particular supermodule.

In addition to these broad architectural insights, the extensive collection of identified modules can improve our understanding of specific biological processes. We used TF binding profiles and their correspondence to modules to create a detailed representation of the yeast transcriptional program. We have also automatically generated >800 function predictions for uncharacterized yeast genes and verified some of them experimentally. Our results are accessible in a highly interactive web site (www.cs.tau.ac.il/~rshamir/samba).

Methods

Integrated Modeling of Genomic Data. We model all genomic information as a weighted bipartite graph G (see ref. 12 for basic graph theoretic definitions). Nodes on one side of G represent genes, and nodes on the other side represent properties of genes or proteins encoded by them. An edge with weight w between a property node v and a gene node g represents an assertion that gene g has property v with probability proportional to w. We may define several properties for the same measurement. For example, for a gene expression measurement, we may define four properties representing strong and weak repression and strong and weak induction of expression. For protein interactions we define properties to express the interaction with a given protein. We use the graph to define the notion of a statistically significant module. A module is defined as a set of genes and a set of properties and is interpreted as a subgraph in G. We score a subgraph by calculating the logarithm of the ratio of its probability under two statistical models, one defining the expected high level of dependency in modules and the other specifying the background behavior of our graph. To facilitate efficient computation, we express this probability as a sum of edge weights and transform the problem of finding high-quality modules to the problem of finding heavy subgraphs in a weighted bipartite graph. Additional details are available in *Supporting Text*, which is published as supporting information on the PNAS web site.

Biclustering and Annotation. The SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) biclustering algorithm searches the genes-properties graph for statistically significant subgraphs. The algorithm uses combinatorial principles to ensure very high

Abbreviations: SAMBA, Statistical-Algorithmic Method for Bicluster Analysis; GO, Gene Ontology; TF, transcription factor.

§To whom correspondence should be addressed. E-mail: rshamir@post.tau.ac.il.

GENETICS

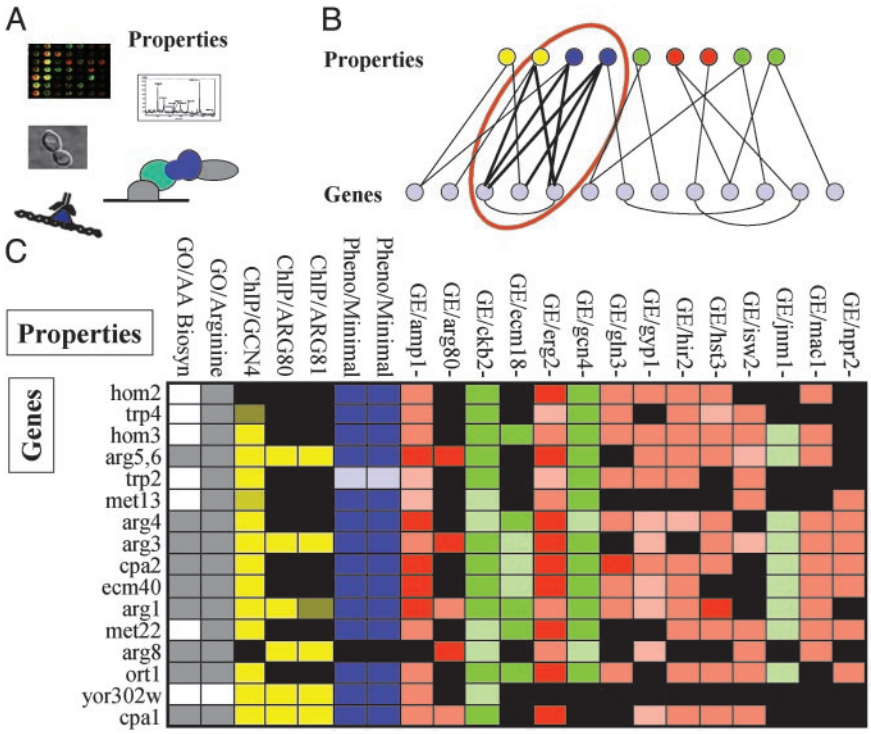

Fig. 1. Integrated analysis of genomic data. Data items (*A*) from diverse sources of biological information are transformed to properties of genes and relations among genes/proteins, together generating a genes-properties bipartite graph (*B*). The graph is represented here schematically as a collection of gene nodes (lower, gray) and property nodes (upper) of different types (yellow, TF binding; blue, knockout phenotypes; green/red, expression profiles; gray, protein interactions). The graph edges represent the (probabilistic) assignment of a property to a gene, with edge weights (not shown here) representing the statistical strength of the assignment. A module (marked by the red oval) corresponds to a set of genes and a set of properties with higher than expected internal degrees. As an example, a real module (*C*) is represented by a matrix of genes by properties. Different types of properties are color-coded differently (using the same color coding as in *B*) with shading to indicate the strength of property assignment (weak–strong binding, low–high phenotype sensitivity, down–up expression regulation). Modules are annotated by testing the enrichment of their genes' GO annotations. The module shown here is strongly enriched with amino acid metabolism genes and more specifically with arginine-related genes. The integrative power of SAMBA is exemplified by the inclusion of *YOR302W* in the module, based on the phenotype and the TF binding properties in the module, even though its expression profile is not sufficiently correlated to the module profile. Indeed, *YOR302W* is *CPA1*'s upstream ORF and is known to function in its translation regulation (16).

efficiency and scalability. Similar modules are filtered so that no reported module may have >20% common gene-property vertex pairs with another (but, still, allowing >20% overlap in their gene sets). We annotated the modules by calculating the hypergeometric functional enrichment score based on the *Saccharomyces* Genome Database Gene Ontology (GO) annotation (13, 14). To search for enriched motifs we used promoters of 600 bp upstream of all yeast ORFs and exhaustively tested all exact gapless motifs of length 5–9 and all motifs of length 6–7 with one gap of length 1–11. For each motif we calculated the hypergeometric *P* value of independence between the set of genes with the motif and the set of module genes, adjusted for multiple testing.

Global Analysis. The module graph has all modules as nodes and an edge whenever at least one-third of the smaller module's gene set is also present in the other's. The distribution of cluster coefficients in random graphs was computed by using degree-preserving randomization of 1,000 graph instances. Details on the randomization and the layout process producing the module map figure are available in *Supporting Text*.

Mating Experiments. Yeast strains deleted for candidate genes were obtained from a deletion library (Research Genetics, Huntsville, AL). To obtain isogenic a- and α-mating mutants, each strain was crossed to an isogenic WT strain and subjected to sporulation and tetrad analysis. Three mutant spores and a WT control of each mating type were subjected to quantitative mating analysis in all combinations, as described (15).

Results

The SAMBA Framework. The SAMBA program for integrative analysis of genomic information has three components: data representation, analysis, and reporting. To facilitate simultaneous analysis of heterogeneous information we view data items from all sources as properties of genes or proteins encoded by them (Fig. 1). We then use a uniform statistical representation of all data sources and apply highly efficient algorithms (extending ref. 17) to analyze them in a process called biclustering (18). The algorithm aims at discovering sets of genes with statistically significant common properties. We call such sets modules. Our computational framework has a unique combination of features: it requires similarity of the genes in a module only across a subset of the properties, ensures that all modules are statistically significant, and uses all sources of information in one uniform framework. Moreover, it allows overlap among modules, which is essential when analyzing systems with multiple-function genes. Extant analysis techniques (8, 9, 17, 19–21) lack one or more of these characteristics. SAMBA is built to exploit the emerging repositories of very large-scale functional genomics data and is highly efficient and scalable in both memory and speed. The software is available as part of the EXPANDER system (www.cs.tau.ac.il/rshamir/~expander/expander.html).

We applied SAMBA to heterogeneous *Saccharomyces cerevisiae* data. The data included ≈1,000 expression profiles, representing 70 series or sets of conditions from 27 different publications, 110 TF binding location profiles (3), 30 growth profiles (4), 1,031

protein interactions (6), 4,177 complex interactions (5), and 1,175 known interactions from the MIPS (Munich Information Center for Protein Sequences) database (22). SAMBA generated 665 significant ($P < 0.05$) modules with maximal overlap of 20% (see *Methods*). A complete list of modules and complementary information can be found at www.cs.tau.ac.il/~rshamir/expander.

We validated the statistical significance of the modules by performing a randomized control test (see Fig. 4, which is published as supporting information on the PNAS web site). We assigned functions to each module whose gene set showed significant overrepresentation of a particular GO class (13, 14) (see *Methods*).

The Power of Data Integration. The modules we found represent many aspects of metabolism and energy derivation, cell cycle, sporulation, mating, protein biosynthesis, RNA processing, stress response, and more. These categories provide a relatively good coverage of biological processes involved in various kinds of adaptation to the environment. Additional data (primarily better coverage of protein interactions) will be required to further characterize modules defining cellular organization and organelles. The SAMBA modules extend previously identified yeast transcription modules (8), taking advantage of the additional information on TF binding, phenotype, and interactions, and based on improved statistical sensitivity that allows for module detection at finer granularity.

The integration of several types of experimental data are often crucial for the detection of modules at fine granularity. Most metabolism modules were identified by combining expression, TF binding, and phenotype data. For example, the arginine module (Fig. 1*C* and Fig. 5, which is published as supporting information on the PNAS web site) associates several known arginine genes (*Arg-1*, -3, -4, -5, -6, and -8) and several other amino acid-related genes. It is supported by diverse expression profiles including many knockout experiments and stress conditions. Mere expression profiles, however, do not suffice to identify that module, and the binding profiles of Arg-80 and Arg-81 are needed to separate it from other amino acid biosynthesis modules. A second, similar example, involving amino acid transport genes is shown in Fig. 6, which is published as supporting information on the PNAS web site. Other biological processes that are not regulated primarily on the transcriptional level are revealed by the combined analysis of growth phenotype and protein interactions. For example, a module involving vesicle transport was detected based on sensitivity to nystatin combined with statistically weaker expression profiles of reaction to methyl methanesulfonate and other agents (Fig. 7, which is published as supporting information on the PNAS web site). The combination of phenotype profiles and expression data, noted before as problematic (4), is in this case essential for the separation of this module from other yeast subsystems. The integration of protein interaction data into the module identification process provides additional information on relations that are not observed in the transcription or phenotype layers. It also allows the interpretation of modules in terms of complexes and cascades. For example, a module related to ubiquitin-dependent protein degradation (Fig. 8, which is published as supporting information on the PNAS web site) is based on the combination of expression data indicating up-regulation in many stress conditions and interactions among various proteins related to ubiquitination. Protein interactions help in separating this module from other stress-responsive genes and also shed light on the cellular mechanisms (complexes, cascades) forming it. Our integrative approach can also help in the analysis of noise-prone protein interactions (e.g., two hybrid screens), by allowing corroboration of interactions by additional functional information.

A Global Map of the Yeast Transcriptional Network. The combined analysis of gene expression and TF binding location was used before to study the transcriptional network of specific processes [e.g., cell cycle (3, 23, 24)]. SAMBA enables the simultaneous analysis of the entire network and the exploration of the relations among TF binding profiles, biological processes, and DNA regulatory motifs in a single map (Fig. 2; for an interactive version see www.cs.tau.ac.il/~rshamir). The transcriptional network map contains as nodes all processes that are significantly overrepresented in at least one module and all of the TFs that are significantly associated with at least one of those modules. We associate a TF with a process whenever there exists a module annotated with that process ($P < 0.01$) that has the TF binding profile as one of its properties. Many of the identified modules contain coexpressed genes. For such modules we can frequently observe one or few common regulating TFs. Indeed, processes that are active during growth in standard conditions are well supported by TF binding profiles (3) that were measured in similar environment. Cell cycle modules, as previously observed (3, 22), are associated with a combination of known TFs acting in a cyclic fashion. Amino acid metabolism modules are associated with combinations of the master regulator Gcn4 and module-specific regulators (Cbf1–Met-4–Met-31 for methionine and sulfur, Arg-80 and Arg-81 for arginine). Respiration modules are regulated by *Hap2-5*, and protein biosynthesis genes are associated with Rap1, Fhl1, and others. Systems that are activated during stress or developmental processes have weaker support of binding profiles (3). For example, sporulation modules are associated with Sum1 but not with other important meiotic regulators (Ndt80, Ume6) (25). Several modules are strongly associated with protein biosynthesis, ribosomal proteins, and RNA processing. Most of the properties defining these modules reflect coordinated expression under stress conditions [e.g., ESR response (26)]. Some of these modules are explained by the binding of Rap1 and Fhl1 (that have similar TF binding profiles) and contain a very high percentage (>90%) of known protein biosynthesis genes. Other modules, enriched in RNA processing, ribosome biogenesis, and stress genes, are not associated with any of the available TF binding profiles. Although these modules contain many Rap1 targets, they have lower percentages of known genes. To analyze these modules in more detail, we screened the promoters of each module's gene set for overrepresented DNA regulatory motifs (see *Methods*). Several statistically significant motifs were detected. The known Rap1/Fhl1 motif (ATCCGTACA) is dominant in the Rap1/Fhl1-bound module, whereas other previously described stress motifs (AAAATTTT, AGGGG, GCGATGAG) (26) are common in modules that are were not associated with a TF based on the binding assays. The transcriptional program inducing the above modules is, thus, apparently based on a combination of TFs, including Rap1, and others that presumably bind the detected active sites. Our analysis indicates in this case the information gaps between current expression and TF binding data.

Global Modular Organization in Yeast. We next turn to explore the global organization of the yeast system as revealed by the association of different modules into one functional network. To this end we constructed and analyzed two graphs. The gene graph (data not shown) contains as nodes all yeast genes, with an edge between two genes whenever they are both contained in some module. The module graph (Fig. 3) contains as nodes all of the modules, with an edge between two modules whenever their gene set intersection is sufficiently large (see *Methods*). Because the gene graph is induced by gene modules (cliques in the graph), it is expected to have a modular structure. The module graph, on the other hand, could not be preassumed to exhibit modularity. To analyze the topology of the two graphs, we computed their clustering coefficients (11). The cluster coefficient of a node is the fraction of the pairs of its

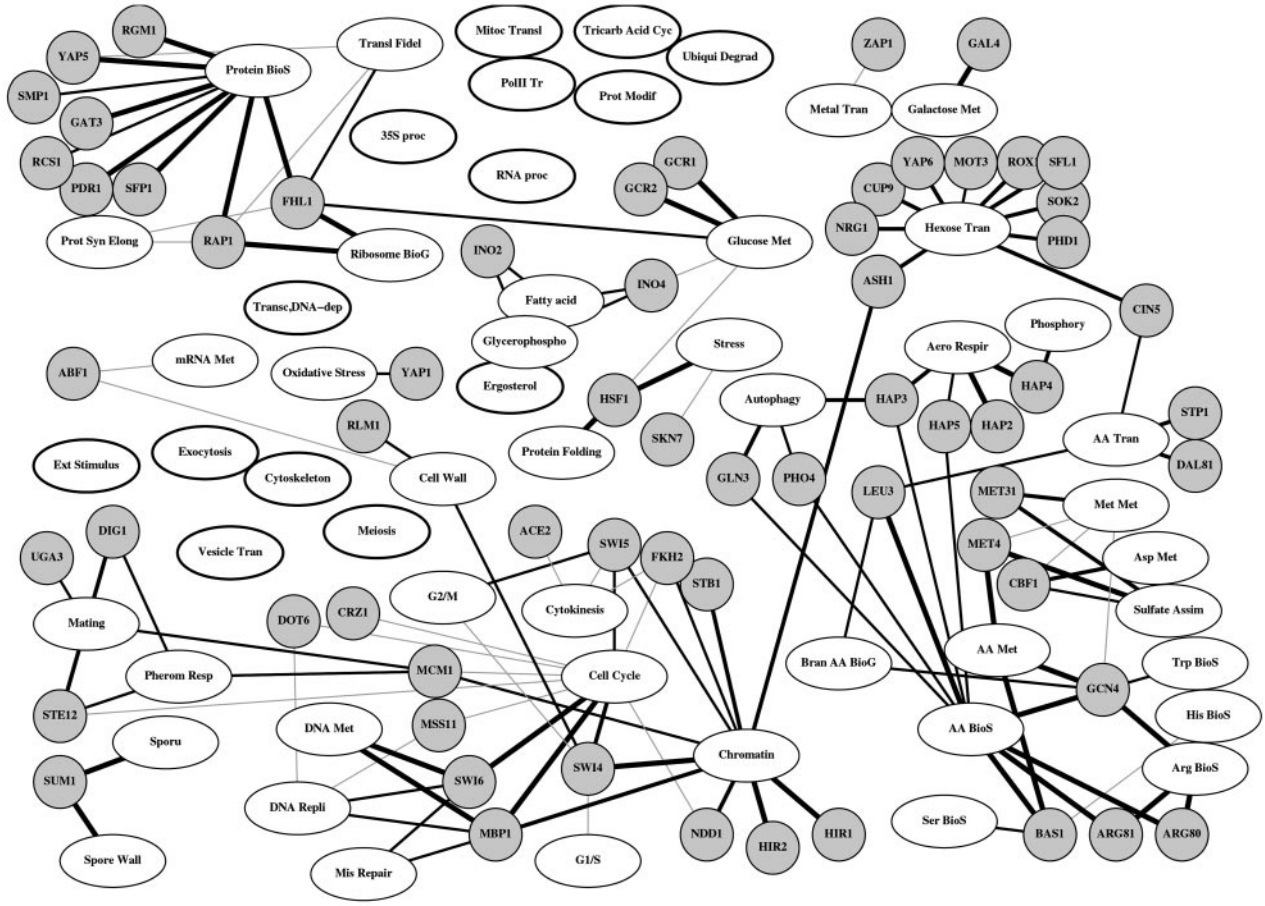

Fig. 2. Functional modules and their TFs in the yeast system. Modules with significant functional enrichment for a particular process ($P < 0.01$) are grouped and plotted as an oval with the process name. TFs with binding profiles associated with any of these modules are marked as gray circles and connected to the associated process. Modules may be enriched in more than one process and thus contribute to several regions in the map. The thickness of the connecting lines is inversely proportional to the *P* value of the functional enrichment in the associated module. The map was automatically generated by SAMBA using no prior biological knowledge. Met, metabolism; Tran, transport. An interactive version of this figure is available at www.cs.tau.ac.il/~rshamir/samba.

neighbors that have edges between them (11). High average clustering coefficient is an indication of modularity. For example, a tree graph has value zero, and a complete graph has value 1. As expected, the average cluster coefficient of the gene graph is a high 0.473. Interestingly, the module graph also has a very high average clustering coefficient of 0.49 (compared to mean coefficient 0.0398 and SD 0.008 on random graphs). Indeed, modules are themselves organized into supermodules (Fig. 3). The overall organization is thus hierarchical: Genes are grouped into modules that are clustered into supermodules. Note that genes can participate in more than one module, and modules can be part of more than one supermodule. Given this architecture, it is important to characterize the genes that connect different supermodules and tie together different processes. To this end, we recalculated the average cluster coefficient in the gene graph over sets of genes annotated with each GO entry. A class with low average coefficient contains genes that are more likely to bridge different supermodules. Indeed, the classes with lowest coefficients are related to signaling (e.g., G protein-coupled receptor with value 0.27 and mitogen-activated protein kinase with 0.29) and transport (e.g., iron transporter with value 0.21 and phospholipid transport with 0.25). Closer examination of genes with low cluster coefficient may help in identifying genes that have multiple functions and improve our understanding of the way in which different biological processes are organized together.

Functional Annotation. We used SAMBA to derive functional annotation of uncharacterized yeast genes. Uncharacterized genes in modules showing high enrichment ($P < 0.01$ and >40% of the annotated genes) for one biological process are likely to participate in the same process. We tested the specificity of this approach by performing a five-way cross validation: we repeatedly applied SAMBA to data sets in which one-fifth of the known gene annotations were hidden and tested the specificity of predicting the function of these genes. Overall we obtained 40–100% specificity for a variety of classes including mating (GO:0007322, 65%), amino acid metabolism (GO:0006520, 40%), sporulation (GO:0030435, 55%), glucose metabolism (GO:0006006, 100%), lipid metabolism (GO:0006629, 92%), and more (Fig. 9, which is published as supporting information on the PNAS web site). Average specificity ranged between 58% and 78%, depending on the strictness threshold used for annotation (see *Methods*). In many cases, the classification errors result from ambiguous annotation terms or too general categories and may represent missing information rather than misclassification. For example, stress response and cell cycle are very general categories that intersect many other processes. Stress-annotated genes are often also related to carbohydrate metabolism and transport, so our classification for such genes may reflect an additional function and not an error. In total, our scheme generated putative functional annotations for 874

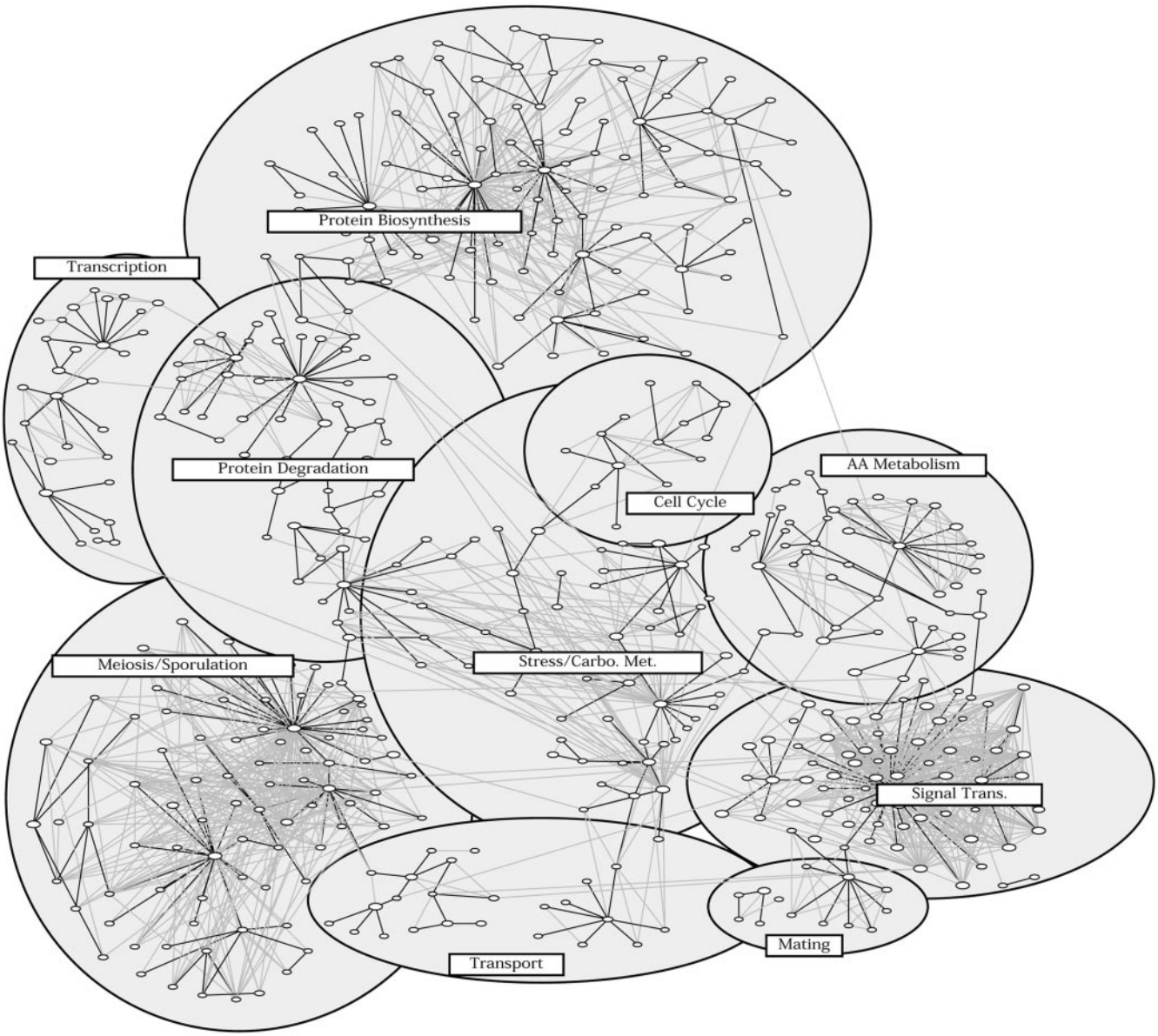

Fig. 3. Hierarchical organization of the yeast molecular network. The module graph was generated by connecting two modules (small ovals) if more than one-third of the genes in one (the smaller) are present in the other. We used our module annotations to manually classify regions in the graph (shown as shaded large ovals). The graph reflects a hierarchical organization that arranges modules in clusters. Some of the clusters (e.g., protein biosynthesis) are organized in more than two hierarchical levels: large modules are composed of several smaller modules, giving a star-like topology.

uncharacterized yeast genes. The complete list is available at www.cs.tau.ac.il/~rshamir/samba.

As an additional test of our annotation accuracy, we analyzed in the laboratory five yeast strains deleted for ORFs predicted by SAMBA to be involved in mating. Quantitative mating experiments showed that four of the strains (*YDR429*c, *YBR223*c/*TDP1*, *YNL106*c/*INP52*, and *YOL106*w) exhibited reduced mating ability, compared to the WT, confirming the involvement of these genes in the mating process. A fifth ORF, *YFL027*c/*GYP8*, exhibited mating ability indistinguishable from that of the WT control.

Discussion

Our integrative analysis of diverse yeast genomewide data provides a detailed view of both the global and local organization of the yeast system. The organization of the molecular network into functional modules is shown here to span many types of regulation mechanisms and to follow organizing principles that govern the nature of interactions among different modules. The module graph we have analyzed is found to associate modules into supermodules in a hierarchical fashion. A similar structure was observed previously in metabolic and protein networks. The topological properties of the molecular network can facilitate better understanding of the way biological systems are organized and can focus our interest on specific genes that are shown to have a role in several regions of the network. By joining basic modules into larger ones in a hierarchical way, we may vastly reduce the complexity of molecular network reconstruction. One of the current challenges of functional genomics is to enable biologists using established assays, as well as those developing novel experimental technologies, to analyze their new data, while taking into consideration the large public sources of various

genomewide measurements. Here, we report on the development and application of key computational steps toward the fulfillment of this vision. The SAMBA platform and methodology enable a unified, high-level representation of heterogeneous biological information and provide a means for the analysis of the biological system under study in light of very large functional genomics databases. The framework provides the statistical robustness and computational efficiency that is required for large-scale studies and is readily extendable to future experimental techniques. Our study of budding yeast data exemplifies the power of integrative analysis and shows that the merger of heterogeneous data has a synergistic effect. We derive global views of the yeast transcriptional network and assign functions to a large number of uncharacterized yeast genes. Integrating more information into our framework is straightforward and can be done by nonexperts. Thus, our methodology enables researchers to use as much of the existing public information as possible, by adding new (possibly private) data to a vast database of multisource information, and analyzing the new data in the context of all available information. Using SAMBA, large repositories of functional genomics data can be used with maximum effect to enable the characterization of complex organisms and heterogeneous biological processes.

We thank Irit Gat-Viks, Rani Elkon, Aviv Regev, and Dana Pe'er for comments on the manuscript and the anonymous referees for many helpful comments and suggestions. M.K. was supported by grants from the Israel Science Foundation and the Binational Science Foundation. R. Shamir was supported by a pilot grant from the McDonnell Foundation and a grant from the Ministry of Science and Technology, Israel. R. Sharan was supported by a Fulbright grant.

1. Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al*. (2000) *Science* **290**, 2306–2309.
2. Iyer, V., Horak, C., Scafe, C., Botstein, D., Snyder, M. & Brown, P. (2001) *Nature* **409,** 533–538.
3. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002) *Science* **298,** 799–804.
4. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., *et al.* (2002) *Nature* **418,** 387–391.
5. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., *et al*. (2002) *Nature* **415,** 180–183.
6. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al*. (2000) *Nature* **403,** 623–627.
7. Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., *et al*. (2001) *Science* **294,** 2364–2368.
8. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. & Barkai, N. (2002) *Nat. Genet.* **31,** 370–377.
9. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. & Friedman, N. (2003) *Nat. Genet.* **34,** 166–176.
10. Rives, A. W. & Galitski, T. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 1128–1133.
11. Ravasz, E., Somera, A., Mongru, A., Oltvai, Z. & Barabasi, A. (2002) *Science* **297,** 1551–1555.
12. Even, S. (1979) *Graph Algorithms* (Computer Science Press, Potomac, MD).
13. Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J. E., *et al*. (2004) *Nucleic Acids Res.* **32,** D311–D314.
14. Gene Ontology Consortium (2004) *Nucleic Acids Res.* **32,** D258–D261.
15. Clark, K. & Sprague, G. (1989) *Mol. Cell. Biol.* **9,** 2682–2694.
16. Wang, Z., Gaba, A. & Sachs, M. (1999) *J. Biol. Chem.* **274,** 37565–37574.
17. Tanay, A., Sharan, R. & Shamir, R. (2002) *Bioinformatics* **18,** S136–S144.
18. Cheng, Y. & Church, G. (2000) in *Proceedings for the Eighth International Conference on Intelligent Systems for Molecular Biology*, eds. Bourne, P. & Gribskov, M. (AAAI Press, Menlo Park, CA), pp. 93–103.
19. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
20. Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A. & Holstege, F. C. (2002) *Mol. Cell* **9,** 1133–1143.
21. Troyanskaya, O., Dolinski, K., Owen, A., Altman, R. & Botstein, D. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 8348–8353.
22. Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., *et al*. (2004) *Nucleic Acids Res.* **32,** D41–D44.
23. Simon, I., Barnett, J., Hannett, N., Harbison, C., Rinaldi, N., Volkert, T., Wyrick, J., Zeitlinger, J., Gifford, D., Jaakkola, T. & Young, R. (2001) *Cell* **106,** 697–708.
24. Segal, E., Barash, Y., Simon, I., Friedman, N. & Koller, D. (2002) in *Proceedings of the Sixth International Conference on Computational Molecular Biology*, ed. Lengauer, T. (Assoc. for Computing Machinery, New York) pp. 263–272.
25. Primig, M., Williams, R., Winzeler, E., Tevzadze, G., Conway, A., Hwang, S., Davis, R. & Esposito, R. (2000) *Nat. Genet.* **26,** 415–423.
26. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000) *Mol. Biol. Cell* **11,** 4241–4257.