Published in final edited form as:

Curr Opin Biotechnol. 2008 October ; 19(5): 482–491. doi:10.1016/j.copbio.2008.07.011.

Gene-module level analysis: Identification to Networks and Dynamics

Xuwei Wang, Ertugrul Dalkic, Ming Wu, and Christina Chan

Michigan State University, Chemical Engineering and Materials Science Department, 2527 Engineering building, East Lansing, MI 48823, United States

Summary

Nature exhibits modular design in biological systems. Gene-module level analysis is based on this module concept, aiming to understand biological network design and systems behavior in disease and development by emphasizing on modules of genes rather than individual genes. Module-level analysis has been extensively applied in genome wide level analysis, exploring the organization of biological systems from identifying modules to reconstructing module networks and analyzing module dynamics. Such module-level perspective provides a high level representation of the regulatory scenario and design of biological systems, promising to revolutionize our view of systems biology, genetic engineering as well as disease mechanisms and molecular medicine.

1. Introduction

Since the advent of molecular cell biology, researchers have studied biological phenomenon mainly by analyzing the function(s) of individual genes and proteins, and the change(s) they exhibit in diseased states. This reductionist approach helped discover many of the underlying biological principles[1]. However, researchers subsequently found that the relationship between genotype and phenotype is too complicated to be ascribed to a change in a single gene [2]. Traditional Mendel genetic linkage test cannot explain complex diseases[3] and cross-talk between cell signaling cascades play a significant role[4,5]. These observations, together with the recent availability of ‘omics’ data, have revolutionized the previous view of single gene-phenotype correlation by demonstrating the importance of the inter-relationships between genes. This has intensified the investigation of protein function in the context of complex biological systems, and initiated a novel systematic perspective of biological processes[6].

Modularity is now recognized as a design principle in biological systems[7], and has been observed in protein-protein interaction[8], metabolic[9] and transcriptional regulation networks[10]. Organized modularity in the yeast protein-protein interaction network is evident in the different biological processes captured by each functional sub-network or module[8]. Similarly, metabolic and transcriptional networks organized into topological units of different functions or regulation, respectively, providing striking evidence of modularization of biological systems[9]. The basic notion that underlies the principle of such modularity lies in evolution. Modularity may enhance the adaptability of biological systems to perform new functions[11], and contribute to the systems' robustness to perturbations and their ability to maintain homeostasis[11,12]. For example, network analysis has shown that the modular

structure of regulatory networks enables mutations to be isolated to specific modules without affecting the overall viability of the system[11–13]. Indeed, increasing evidence suggests that a functional gene system is composed of coordinated and interdependent modules[14], and such modularization is shaping current research in biology.

With modularity being ubiquitous in biological systems, applying module-level analysis should aid the study of biological systems at the different levels as well as the system dynamics. Gene module-level analysis emphasizes on groups or modules of genes rather than individual genes [15–17]. Depending on their application, such groups or modules could be cliques of physically interacting proteins, clusters of gene expression profiles, pathways in a signaling, regulatory, or metabolic network, or groups of manually pre-defined and functionally related genes. Module-level analysis has been extensively applied in genome wide level analysis and systems biology[18,19]. For example, clustering has been used frequently with microarray data to identify co-expressed or co-regulated groups or modules of genes[18,19]. Similarly, module-assisted methods have been used to classify genes into coherent functional groups to predict protein functions[18,19]. Characteristics of a module may include a higher number of interior connections relative to interactions[20], similarity of function[20], separability from its surrounding structure[21–23], or conservation of function[24]. With these characteristics, modules qualify as building blocks of biological system, thus one can dissect the system's structure and dynamics by studying these modules in isolation, and then explore the complex function of the whole system through the interactions of the modules. To date, the focus of module-level analysis has been predominantly on identifying individual modules and the qualitative descriptions of the groupings of biological data. Lately, module-level analyses is shifting from a descriptive to a more quantitative analysis of the inter-module relationships, i.e., studying their interplay through module network reconstruction and dynamics, to understand the pathways, mechanisms and network regulations underlying human diseases [25]. This is driven by the need to mine and integrate different large-scale datasets[26], requiring quantitative measures to identify the modules and their interactions or relationships [27]. Gene module analysis produces a high level representation of the regulatory scenario and may help provide insight into biological behaviors by analyzing the dynamic features of the gene groups[23] or summarizing the network in the abstract form of modules[27].

Current gene module-level analysis is focused predominantly on identifying the modules and several methods have been applied, resulting in different approaches of module analysis (Figure 1). In this review we explore the methods and applications of gene module-level analysis from identifying modules, reconstructing module networks and analyzing module dynamics. **Module identification** centers on the internal vertices and connections within the modules, whereas **module network** and **dynamics** focus on the external interactions between the modules. Table 1 summarizes the strategies for module-level analysis discussed in this review.

2. Module identification

The first level of module analysis is to identify the gene modules involved in specific biological processes or systems. Currently, the methods to identify gene modules center around three major approaches: (1) *network based approaches* identify highly connected sub-graphs in the biological networks as modules; (2) *expression based approaches* identify groups of co-expressed genes as modules; and (3) *prior pathways based approaches* identify altered pathways as modules.

2.1 Network-based approaches

From a network perspective, modules can be defined as subsets of vertices in the network with high intra-module connectivity (Figure 2). Identification of modules from biological networks focuses predominantly on protein interaction networks.

Methods of finding modules from network topology use mainly hierarchical and graph clustering[18]. The hierarchical clustering approach first classifies pairs of vertices with a weight, and then clusters the vertices based on the weights. Weights may be defined as the number of vertex-independent paths, or the number of edge-independent paths[7] (Figure 2). For example, a high number of vertex-independent paths (i.e., paths between two vertices involving different sets of intermediate vertices) would correspond to a high weight, and have a higher chance of being clustered together. Agglomerative algorithms start with a complete set of vertices and add individual edges based on the weights. Conversely, divisive algorithms start with the whole network and proceed by removing edges between vertices based on the weights. After hierarchical transformation of the network topology, qualitative or quantitative definitions of communities are used to finally collect the modules. For example, a typical quantitative definition requires that each vertex in a community be more connected to vertices within the community and less with those outside[28] (Figure 2). In graph clustering, super-paramagnetic clustering and Monte Carlo optimization have been used to identify highly connected clusters[29]. Alternatively, Bader and Hogue used core clustering coefficients to weight vertices and used these weights to find the dense regions in a large protein interaction network [30].

Additional information, i.e., gene expression, genetic interactions, deletion phenotypes[18], and metabolic information[40] have been incorporated into the network topology to identify modules. An example is to predefine groups of proteins previously known to interact as a “seed”. Members are added to the protein-protein interaction (PPI) network if they satisfy a statistical confidence score. Statistical confidence is derived from distance metrics obtained from multiple sources. Protein pairs with short distances in networks derived from yeast two-hybrid and mass spectrometry data is given a high score, whereas the presence of a long distance lowers the score[39]. Similarly, combining distance metrics from gene expression data (based on high correlation coefficient) with distance metrics from metabolic interaction networks (based on shortest path between two molecules in the network)[40] through hierarchical clustering is another method to identify modules (clusters).

Network based module identification is mainly applied to large scale PPI data[31–35,71]. The identified modules correlate well with protein complexes identified experimentally and tend to contain proteins with similar functions[32,35,71]. Thus, their success in identifying modules encouraged their application in identifying novel modules, pathways, and protein complexes [32,35,71]. For example, a protein of unknown function within the same module as proteins of known function are deemed functionally related. Likewise, these network-based approaches have been applied to identify modules from other biological networks, including metabolic [9], signaling[72] and gene networks[37,38,73]. A disadvantage of protein interaction network analysis is the possibility of identifying modules that may not co-exist *in vivo*[74].

2.2 Expression-based approaches

Network-based approaches identify modules from the topological structure of the biological networks. In contrast, expression-based approaches use gene expression data to infer modules of genes exhibiting similar expression patterns. Clustering methods are applied frequently to identify modules of co-expressed genes. The motivation behind clustering is the assumption that co-expressed genes are coordinately regulated[75], and should reveal gene modules with similar functionality. Traditional clustering methods, including agglomerative and partitional

techniques, are widely used to identify modules of genes co-expressed across experimental conditions[41,76]. Clustering of gene expressions is an active area of research with several recent reviews[41,43]. Commonly used methods include hierarchical and K-means clustering which are easy to implement and highly efficient in discovering expression patterns across time or conditions. Alternatively, bi-clustering method can capture such modules under a *subset* of experimental conditions *and* time[42].

A limitation of clustering methods is their reliance on functional interpretation of the identified modules through a post-hoc analysis, such as, enrichment analysis of the gene categories. To generate gene modules that are more easily interpreted, different sources of information, such as regulator-target gene relationships[77–79] and pre-defined pathways[44], have been integrated with expression data in the clustering procedure. Furthermore, clustering methods do not take into account the temporal dependencies of the expression data, limiting their ability to capture the dynamic behavior of genes[80]. To address this issue, approaches based on dynamic models [45,46] or temporal trends of the gene expression data[47,48] have been developed to identify modules from time-series microarray data.

2.3 Prior pathways based approaches

Pathway based approaches aim to identify pathways, namely, groups of genes, rather than individual genes involved in biological processes. Single gene level analyses typically use univariate statistical tests to detect for genes with significant expression changes. However, these analyses provide a long list of genes that oftentimes are difficult to extract useful information and to put into biological context. In contrast, pathway based approaches detect subtle and coordinated changes in expression of a group of genes in a pathway or with related function[81]. Methods to identify the genes in a pathway include enrichment analysis[49–52] and supervised machine learning approaches[58–62].

A number of enrichment analyses are used to identify pathways that distinguish biological phenotypes based upon function. Good reviews exist describing the principles and issues[81, 82] associated with the statistical methodology. In particular, gene set enrichment analysis (GSEA)[49] and similar methods[50–52] have been applied to microarray data to unravel the mechanisms of complex diseases[54–56,83,84]. The gene expression profiles are viewed in terms of functional pathways, and the pathways with a large number of differentially expressed genes are identified as significant among two sets of conditions. Recent enhancements allow handling of more than two groups[53,57]. Nevertheless, GSEA evaluates one pathway at a time and does not capture the pathway–interdependences that contribute to changes in the phenotypes. Furthermore, ontology mapping and GSEA consider all the genes in a pathway equally important. But a recent study demonstrated that by including information on the topology or position of the differentially expressed genes in the pathway, helped to identify pathways that may be relevant to lung cancer, that were otherwise missed[85]. Another concern is the quality of the pathways, since the pathways are usually curated from the literature or other resources, non-relevant genes may be included, or relevant genes may be excluded from the pathways. Several attempts to minimize these misspecifications include, defining signature genes to represent pathway behavior[86], refining pathways to adapt to specific conditions by removing unaltered genes from the dataset[66,87,88], and improving the functional interpretation of gene groups by including additional information (e.g., experimental conditions and biological context) associated with the group[89].

Alternatively, supervised machine learning approaches evaluate multiple pathways simultaneously, and thus could account for pathway interdependence. Various supervised machine learning methods, such as discriminant analysis[60], random forest[62], and partial least square regression[61] have been applied to identify relevant phenotype-specific pathways. The rationale is, if a pathway improves the prediction of the phenotype, its relevance increases.

These machine learning approaches use pathways as input variables, and thus require a quantitative representation of pathway activity. Certain preprocessing methods have been applied to quantify pathway activities by summarizing the gene expressions in the pathways. For example, singular value decomposition-based method[90] uses first principle component to represent pathways, and shrunk centroid-based methods [60,61] use centroids of the gene expressions in the pathways, but information loss is a concern with these preprocessing methods. To address the information loss, an ensemble of classification models may be built in lieu of predictive models based on single-classifier. For instance, a nonparametric pathway-based regression (NPR) method[58,59] uses expression of the genes in a pathway to characterize the activity of the pathway, and the activity levels are regressed to the phenotypes to form a predictive model. Similarly, Pang et al[62] construct a decision tree using the genes in a pathway and combine the decision trees of all the pathways into a “forest” to predict the phenotype. These ensemble approaches provide a nice interpretation of the gene-pathway and pathway-phenotype relationships. Another concern is reduced performance through the introduction of noise or non-relevant information with inclusion of all the genes of the pathways [91]. To address the reduced performance, Edelman et al. formulated an enrichment score to determine the variation in activity of only the differentially expressed genes in a pathway among the individual samples[53].

3. Module networks and dynamics

From a systems biology perspective, modules are building blocks that may be organized into a complex network to achieve cellular regulation[92]. Changes in the modular structure with respect to the inter-module connections are generally associated with alterations in the properties and functions of the cell[22]. Investigation of the inter-module relationships and dynamics enables the prediction and manipulation and eventually our understanding of cellular behavior.

Module networks are constructed predominantly from PPI and gene expression data. Early analyses of module networks relied predominantly on PPI data[30,31]. After identifying the modules in the network, the inter-module interactions are generally analyzed through the role played by the proteins that connect the different modules[31]. For example, in studying the protein interaction network of yeast filamentation, modules, i.e., the MAPK and HOG pathways, were found to be connected by the Ste11 gene. Indeed, support exists in the literature to suggest Ste11 cross-talks between these two pathways[93]. Thus far, these analyses are mostly limited to identifying inter-module connections and lack a quantitative description of module networks.

Using expression data to infer module networks are in its early stages. One such attempt uses an eigengene network approach, where hierarchical clustering of the expression data is used in identifying the modules[65]. The eigengene, captured by the first principle component of the gene expressions in a module, are used in determining the associations between the modules. The associations are characterized by the correlation coefficients between the eigengenes, and modules with high correlations are combined to construct the network[65]. Once the modules are represented by quantitative measures, i.e., the first principle component, available approaches, such as information-theory approaches and probabilistic models[94], can be used to infer the gene module networks.

Attempts to overcome the limitations of single source data have integrated expression with PPI data to construct module networks[27,95]. The incorporation of physical interaction information, such as transcription factor binding and PPI data, facilitate functional interpretation of module-associations. This approach was able to infer co-regulation of functional modules in *Saccharomyces cerevisiae*[64] and crosstalk between modules in

colorectal cancer metastasis[63]. Similarly, incorporating pre-defined gene-sets into the expression analysis can take into account genes that are not differentially expressed[54,64,66].

Dynamic analysis of module networks, thus far, focuses on the changes within the modules and the connections between the modules with time, as well as the dynamic behaviors of the modules in the networks. To study the temporal changes of the module, a common approach is to identify the modules, i.e. the altered pathways, by coupling time-series expression data with pathway information, at each time-point separately[54,66,67,96]. Some studies further analyze the connections between the modules. By qualitatively analyzing the changes in the module relationships across the different stages or time points this approach has provided insight into tumor progression[54,66] and response to acute heat stress[96]. Dynamic models build upon time-series expression data to analyze the dynamic behavior of the modules. Quantitative measures adapted from dynamic system theory, e.g., stability, controllability and steady-state behavior, are applied to analyze the module dynamics, by comparing the difference in these measures between normal and abnormal cells[68]. This approach to exploring module dynamics, however, requires a certain degree of mechanistic knowledge of the modules (i.e., equations that describe the dynamic regulations of the genes), thus limiting their application to well-studied pathways. In cases where such mechanistic knowledge is not available, an alternative approach is to use state-space model (SSM). SSM has been applied to infer modules and dynamic regulation of yeast cell cycle and cancer progression[69,70]. In SSM, latent variables are used to represent unobservable biological entities (i.e., transcription factors or protein complexes), and the temporal expression profiles are mapped to these latent variables to model the dynamics and dependencies of these latent variables. Genes that map to the same latent variables are grouped together, and the dynamics and dependencies of these latent variables are captured by linear dynamic models used to describe the dynamic behaviors of the modules and their interactions.

4. Conclusion

In the past three to four years, large biomolecular interaction networks, constructed from experimental data, have revealed many “emergent” properties of network functionality[8]. Concomitantly new problems have arisen, namely how to model biological processes at different organizational levels[97]. This has provided opportunities as well as challenges in the development of network biology approaches in uncovering the regulations and mechanisms of complex diseases[25]. As discussed above, the property of modularity and module-level analysis play an important role in enabling the exploration of the regulation and design of biological systems at a higher organizational level. Therefore, recent progress in this field of module-level analysis promises to revolutionize our view of systems biology, genetic engineering as well as disease mechanisms and molecular medicine. Nevertheless, challenges remain. For the time being, the quality and coverage of the available data restrict our ability to perform module-level analyses that accurately recapitulate the biological system. To address the quality of the data such as pathway information, an edge ontology has been proposed to standardize pathway information[98]. An edge ontology will help to distinguish different types of interactions by using different edge notations. Since the number of components may vary for the same pathway in different databases, a well-defined edge ontology would provide less ambiguity when comparing pathways across databases as well as enable a more comprehensive representation of the relationships within and among pathways. For coverage, integrating available multi-source data can help improve this situation[99]. Although, the amount and types of data (i.e. continuous vs. discrete, static vs. dynamic) generated are increasing, more computational methodologies are needed to preprocess, analyze, integrate and interpret the various types of data in the pipeline. Finally, many computational methods currently do not take full advantage of the available data or apply strong assumptions that may not be realistic,

for example, many dynamic models assume time-invariant structures making it difficult to capture the time-varying module relationships. As the community strives to address these problems, we envision that ultimately a comprehensive gene module-level view may accurately describe the biological system.

Acknowledgement

C.C is supported in part by the National Institute of Health (1R01GM079688, R21CA126136, R21RR024439), National Science Foundation (BES0425821, DBI 0701709), MUCI Challenge Fund and the MSU Foundation on the Center for Systems Biology. ED is supported in part by the Quantitative Biology and Modeling Initiative.

References

1. Nurse P. Reductionism and explanation in cell biology. *Novartis Found Symp* 1998;213:93–101. [PubMed: 9653717]discussion 102–105
2. Vladutiu GD. Heterozygosity: an expanding role in proteomics. *Mol Genet Metab* 2001;74:51–63. [PubMed: 11592803]
3. Emahazion T, Feuk L, Jobs M, Sawyer SL, Fredman D, St Clair D, Prince JA, Brookes AJ. SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends Genet* 2001;17:407–413. [PubMed: 11418222]
4. Pawson T, Saxton TM. Signaling networks--do all roads lead to the same genes? *Cell* 1999;97:675–678. [PubMed: 10380918]
5. Dumont JE, Dremier S, Pirson I, Maenhaut C. Cross signaling, cell specificity, and physiology. *Am J Physiol Cell Physiol* 2002;283:C2–C28. [PubMed: 12055068]
6. Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* 2003;19:551–560. [PubMed: 14550629]
7. Girvan M, ME N. Community structure in social and biological networks. *Proc Natl Acad Sci U S A* 2002;99:7821–7826. [PubMed: 12060727]
8. Han JD. Understanding biological functions through molecular networks. *Cell Res* 2008;18:224–237. [PubMed: 18227860] A comprehensive review of network biology, from network properties to biological significance and functional and mechanistic insights inferred from analyzing the network. It introduces basic ideas, concepts and approaches in network biology, and reviews the representative strategies used to tackle the different biological problems from a top-down perspective. This paper is a good starting point in learning network biology.
9. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science* 2002;297:1551–1555. [PubMed: 12202830]
10. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. *Nat Genet* 2002;31:370–377. [PubMed: 12134151]
11. Yook SH, Oltvai ZN, Barabasi AL. Functional and topological characterization of protein interaction networks. *Proteomics* 2004;4:928–942. [PubMed: 15048975]
12. Thieffry D, Romero D. The modularity of biological regulatory networks. *Biosystems* 1999;50:49–59. [PubMed: 10235650]
13. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature* 2000;407:651–654. [PubMed: 11034217]
14. Wagner GP, Pavlicev M, Cheverud JM. The road to modularity. *Nat Rev Genet* 2007;8:921–931. [PubMed: 18007649] Nature induced modularity into molecular networks to enhance their adaptability to new functions. This review summarizes the evidence for modularity, introduces recent models of modular evolution, and discusses the origin of modularity.
15. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007;3:140. [PubMed: 17940530]
16. Von Dassow G, Odell GM. Design and constraints of the *Drosophila* segment polarity module: robust spatial patterning emerges from intertwined cell state switches. *J Exp Zool* 2002;294:179–215. [PubMed: 12362429]

17. von Dassow G, Meir E, Munro EM, Odell GM. The segment polarity network is a robust developmental module. *Nature* 2000;406:188–192. [PubMed: 10910359]
18. Sharan R, Ulitsky IRS, R S. Network-based prediction of protein function. *Mol Syst Biol* 2007;3:88. [PubMed: 17353930] A review for the field of protein-function annotation. Besides providing an overview of classical, network-based, and module-assisted methods, the authors discuss recent developments in this area and compare the performance of different methods in terms of accuracy and robustness. The concept of the functional module is clearly-defined and frequently used in protein function annotation
19. Wong DJ, Chang HY. Learning more from microarrays: insights from modules and networks. *J Invest Dermatol* 2005;125:175–182. [PubMed: 16098025]
20. Alon U. Biological networks: the tinkerer as an engineer. *Science* 2003;301:1866–1867. [PubMed: 14512615]
21. Lipson H, Pollack JB, Suh NP. On the origin of modular variation. *Evolution Int J Org Evolution* 2002;56:1549–1556.
22. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* 1999;402:C47–C52. [PubMed: 10591225]
23. Nordling TE, Hiroi N, Funahashi A, Kitano H. Deduction of intracellular sub-systems from a topological description of the network. *Mol Biosyst* 2007;3:523–529. [PubMed: 17639126]
24. Ettema T, van der Oost J, Huynen M. Modularity in the gain and loss of genes: applications for function prediction. *Trends Genet* 2001;17:485–487. [PubMed: 11525815]
25. Ideker T, Sharan R. Protein networks in disease. *Genome Res* 2008;18:644–652. [PubMed: 18381899] A comprehensive review of the applications of protein networks to disease and medicine.
26. Curtis RK, Oresic M, Vidal-Puig A. Pathways to the analysis of microarray data. *Trends Biotechnol* 2005;23:429–435. [PubMed: 15950303]
27. Xue H, Xian B, Dong D, Xia K, Zhu S, Zhang Z, Hou L, Zhang Q, Zhang Y, Han JD. A modular network model of aging. *Mol Syst Biol* 2007;3:147. [PubMed: 18059442] Combines gene expression data with PPI networks to reveal a modularized network view of the aging process, where the stochastic effects of aging are linked to the dynamics of the modular network. To the best of our knowledge, it is the first attempt at modeling a complex biological process as the result of inter-module communications in a modular network. It may also be the first attempt at studying the modular dynamics on a large time-scale, within the time range of human life.
28. Radicchi F, Castellano C, Cecconi F, Loreto V, D P. Defining and identifying communities in networks. *Proc Natl Acad Sci U S A* 2004;101:2658–2663. [PubMed: 14981240]
29. Spirin V, LA M. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* 2003;100:12123–12128. [PubMed: 14517352]
30. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003;4:2. [PubMed: 12525261]
31. Rives AW, Galitski T. Modular organization of cellular networks. *Proc Natl Acad Sci U S A* 2003;100:1128–1133. [PubMed: 12538875]
32. Luo F, Yang Y, Chen C, Chang R, Zhou J, Scheuermann R. Modular organization of protein interaction networks. *Bioinformatics* 2007;23:207–214. [PubMed: 17092991]
33. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* 2003;21:697–700. [PubMed: 12740586]
34. Brun C, Herrmann C, Guénoche A. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics* 2004;5:95. [PubMed: 15251039]
35. Chen J, Yuan B. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* 2006;22:2283–2290. [PubMed: 16837529]
36. Zhang S, Ning X, Zhang XS. Identification of functional modules in a PPI network by clique percolation clustering. *Comput Biol Chem* 2006;30:445–451. [PubMed: 17098476]
37. Aggarwal A, Guo DL, Hoshida Y, Yuen ST, Chu KM, So S, Boussioutas A, Chen X, Bowtell D, Aburatani H, et al. Topological and functional discovery in a gene coexpression meta-network of gastric cancer. *Cancer Res* 2006;66:232–241. [PubMed: 16397236]

38. Ruan XG, Wang JL, Li JG. A network partition algorithm for mining gene functional modules of colon cancer from DNA microarray data. *Genomics Proteomics Bioinformatics* 2006;4:245–252. [PubMed: 17531800]
39. Bader J. Greedily building protein networks with confidence. *Bioinformatics* 2003;19:1869–1874. [PubMed: 14555618]
40. Hanisch D, Zien A, Zimmer R, Lengauer T. Co-clustering of biological networks and gene expression data. *Bioinformatics* 2002;18:S145–S154. [PubMed: 12169542]
41. Kerr G, Ruskin HJ, Crane M, Doolan P. Techniques for clustering gene expression data. *Comput Biol Med* 2008;38:283–293. [PubMed: 18061589]
42. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* 2004;1:24–45. [PubMed: 17048406]
43. Miller DJ, Wang Y, Kesidis G. Emergent unsupervised clustering paradigms with potential application to bioinformatics. *Front Biosci* 2008;13:677–690. [PubMed: 17981579]
44. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. *Nat Genet* 2004;36:1090–1098. [PubMed: 15448693]
45. Ma P, Castillo-Davis CI, Zhong WX, Liu JS. A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* 2006;34:1261–1269. [PubMed: 16510852]
46. Zeng YJ, Garcia-Frias J. A novel HMM-based clustering algorithm for the analysis of gene expression time-course data. *Computational Statistics & Data Analysis* 2006;50:2472–2494.
47. Ernst J, Nau GJ, Bar-Joseph Z. Clustering short time series gene expression data. *Bioinformatics* 2005;21:1159–1168. [PubMed: 15961453]
48. Sacchi L, Bellazzi R, Larizza C, Magni P, Curk T, Petrovic U, Zupan B. TA-clustering: Cluster analysis of gene expression profiles through Temporal Abstractions. *International Journal of Medical Informatics* 2005;74:505–517. [PubMed: 15941669]
49. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545–15550. [PubMed: 16199517]
50. Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 2005;6:144. [PubMed: 15941488]
51. Klebanov L, Glazko G, Salzman P, Yakovlev A, Xiao Y. A multivariate extension of the gene set enrichment analysis. *J Bioinform Comput Biol* 2007;5:1139–1153. [PubMed: 17933015]
52. Jiang Z, Gentleman R. Extensions to gene set enrichment. *Bioinformatics* 2007;23:306–313. [PubMed: 17127676]
53. Edelman E, Porrello A, Guinney J, Balakumaran B, Bild A, Febbo PG, Mukherjee S. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 2006;22:e108–e116. [PubMed: 16873460]
Presents a method to quantify the activity of individual pathways in the samples. This method may be very useful for inferring networks of pathways. Once the pathways are quantified, computational methods could be used to infer the network.
54. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, et al. Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 2007;39:41–51. [PubMed: 17173048] An impressive amount of gene sets (around 14,000) were collected, and thus provides a useful resource for gene-set analysis.
55. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, Girtman K, Mathew S, Ma J, Pounds SB, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* 2007;446:758–764. [PubMed: 17344859]
56. Li Z, Srivastava S, Yang X, Mittal S, Norton P, Resau J, Haab B, Chan C. A hierarchical approach employing metabolic and gene expression profiles to identify the pathways that confer cytotoxicity in HepG2 cells. *BMC Syst Biol* 2007;1:21. [PubMed: 17498300]
57. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* 2005;102:13544–13549. [PubMed: 16174746]

58. Wei Z, Li H. Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics* 2007;8:265–284. [PubMed: 16772399]
59. Luan Y, Li H. Group additive regression models for genomic data analysis. *Biostatistics* 2008;9:100–113. [PubMed: 17513311]
60. Tai F, Pan W. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics* 2007;23:3170–3177. [PubMed: 17933851]
61. Tai F, Pan W. Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics* 2007;23:1775–1782. [PubMed: 17483507]
62. Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H. Pathway analysis using random forests classification and regression. *Bioinformatics* 2006;22:2028–2036. [PubMed: 16809386]
63. Li Y, Agarwal P, Rajagopalan D. A Global Pathway Crosstalk Network. *Bioinformatics*. 2008
64. Petti AA, Church GM. A network of transcriptionally coordinated functional modules in *Saccharomyces cerevisiae*. *Genome Res* 2005;15:1298–1306. [PubMed: 16109970]
65. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 2007;1:54. [PubMed: 18031580] Proposed a proof of concept approach to construct module network from gene expression data, and a general framework to identify modules and infer module networks.
66. Edelman EJ, Guinney J, Chi JT, Febbo PG, Mukherjee S. Modeling Cancer Progression via Pathway Dependencies. *PLoS Comput Biol* 2008;4:e28. [PubMed: 18282083] Develops a data-driven pipeline to construct pathway networks from condition-specific gene expression data. The pipeline includes identifying curated pathways relevant to the phenotypes, refining the curated pathways based on the expression data, and finally constructing networks of the refined pathways.
67. Li Z, Srivastava S, Findlan R, Chan C. Using dynamic gene module map analysis to identify targets that modulate free fatty acid induced cytotoxicity. *Biotechnol Prog* 2008;24:29–37. [PubMed: 18052188]
68. Xiong H, Choe Y. Dynamical pathway analysis. *BMC Syst Biol* 2008;2:9. [PubMed: 18221557] Characterizes the dynamic behavior of pathways by applying several metrics from control theory. Pathway states are evaluated in a holistic and quantitative manner.
69. Hirose O, Yoshida R, Imoto S, Yamaguchi R, Higuchi T, Charnock-Jones DS, Print C, Miyano S. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics* 2008;24:932–942. [PubMed: 18292116] Provides a state-space model to infer module networks from short time-series microarray data. Suggests a method to improve the parameter estimation from limited time points.
70. Inoue LY, Neira M, Nelson C, Gleave M, Etzioni R. Cluster-based network model for time-course gene expression data. *Biostatistics* 2007;8:507–525. [PubMed: 16980695]
71. Zhang S, Ning X, Zhang X. Identification of functional modules in a PPI network by clique percolation clustering. *Comput Biol Chem* 2006;30:445–451. [PubMed: 17098476]
72. Baudot A, Angelelli J-B, Guenoche A, Jacq B, Brun C. Defining a Modular Signalling Network from the Fly Interactome. *BMC Syst Biol* 2008;2:45. [PubMed: 18489752]
73. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 2007;8:22. [PubMed: 17250769]
74. Zhao SD, Li Y. Extracting functional modules from biological pathways. *Nature Precedings*. 2007><http://dx.doi.org/10.1038/npre.2007.1457.1031><
75. Boutros PC, Okey AB. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. *Brief Bioinform* 2005;6:331–343. [PubMed: 16420732]
76. Zhao Y, Karypis G. Data clustering in life sciences. *Mol Biotechnol* 2005;31:55–80. [PubMed: 16118415]
77. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 2003;34:166–176. [PubMed: 12740579]
78. Ihmels J, Bergmann S, Barkai N. Defining transcription modules using large-scale gene expression data. *Bioinformatics* 2004;20:1993–2003. [PubMed: 15044247]

79. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, et al. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 2003;21:1337–1342. [PubMed: 14555958]
80. Bar-Joseph Z. Analyzing time series gene expression data. *Bioinformatics* 2004;20:2493–2503. [PubMed: 15130923]
81. Nam D, Kim SY. Gene-set approach for expression pattern analysis. *Brief Bioinform*. 2008
82. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007;23:980–987. [PubMed: 17303618]
83. Baur JA, Pearson KJ, Price NL, Jamieson HA, Lerin C, Kalra A, Prabhu VV, Allard JS, Lopez-Lluch G, Lewis K, et al. Resveratrol improves health and survival of mice on a high-calorie diet. *Nature* 2006;444:337–342. [PubMed: 17086191]
84. Lagouge M, Arghmann C, Gerhart-Hines Z, Meziane H, Lerin C, Daussin F, Messadeq N, Milne J, Lambert P, Elliott P, et al. Resveratrol improves mitochondrial function and protects against metabolic disease by activating SIRT1 and PGC-1 α . *Cell* 2006;127:1109–1122. [PubMed: 17112576]
85. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R. A systems biology approach for pathway level analysis. *Genome Res* 2007;17:1537–1545. [PubMed: 17785539] Develops a novel method to identify altered pathways by integrating the topological information of pathways, such as the positions and interactions of the genes in the pathways. This method takes into account the importance of the genes in the pathways based on the topological information, unlike previous methods which consider the genes in a pathway equally important.
86. Panteris E, Swift S, Payne A, Liu X. Mining pathway signatures from microarray data and relevant biological knowledge. *J Biomed Inform* 2007;40:698–706. [PubMed: 17395545]
87. Novak BA, Jain AN. Pathway recognition and augmentation by computational analysis of microarray expression data. *Bioinformatics* 2006;22:233–241. [PubMed: 16278238]
88. Gat-Viks I, Shamir R. Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Res* 2007;17:358–367. [PubMed: 17267811]
89. Hummel M, Meister R, Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics* 2008;24:78–85. [PubMed: 18024976]
90. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 2005;6:225. [PubMed: 16156896]
91. Tsai CA, Chen CH, Lee TC, Ho IC, Yang UC, Chen JJ. Gene selection for sample classifications in microarray experiments. *DNA and Cell Biology* 2004;23:607–614. [PubMed: 15585118]
92. Yanai I, Baugh LR, Smith JJ, Roehrig C, Shen-Orr SS, Claggett JM, Hill AA, Slonim DK, Hunter CP. Pairing of competitive and topologically distinct regulatory modules enhances patterned gene expression. *Mol Syst Biol* 2008;4:163. [PubMed: 18277379] Provides experimental evidence that demonstrates molecular mechanisms attributed to coordination of the gene modules.
93. Cullen PJ, Schultz J, Horecka J, Stevenson BJ, Jigami Y, Sprague GF Jr. Defects in protein glycosylation cause SHO1-dependent activation of a STE12 signaling pathway in yeast. *Genetics* 2000;155:1005–1018. [PubMed: 10880465]
94. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. *Mol Syst Biol* 2007;3:78. [PubMed: 17299415]
95. Ulitsky I, Shamir R. Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction networks. *Mol Syst Biol* 2007;3:104. [PubMed: 17437029]
96. Szustakowski JD, Kosinski PA, Marrese CA, Lee JH, Elliman SJ, Nirmala N, Kemp DM. Dynamic resolution of functionally related gene sets in response to acute heat stress. *BMC Mol Biol* 2007;8:46. [PubMed: 17550601]
97. Schlitt T, Brazma A. Modelling gene networks at different organisational levels. *FEBS Lett* 2005;579:1859–1866. [PubMed: 15763564]
98. Lu LJ, Sboner A, Huang YJ, Lu HX, Gianoulis TA, Yip KY, Kim PM, Montelione GT, Gerstein MB. Comparing classical pathways and modern networks: towards the development of an edge ontology. *Trends Biochem Sci* 2007;32:320–331. [PubMed: 17583513]
99. Kutalik Z, Beckmann JS, Bergmann S. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat Biotechnol* 2008;26:531–539. [PubMed: 18464786]

100. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–2504. [PubMed: 14597658]

Figure 1. Schematic representation of gene-module level analysis, from module identification to module network and dynamics

The first level of analysis is module identification. The modules may be identified using different methods, including network-based, expression-based and pathway-based approaches. Based on these approaches, different types of modules may be obtained (as indicated in the box with dash line), including condense subnetworks, clusters of co-expressed genes and altered pathways. After identifying the modules, higher level of analysis including module network reconstruction and module dynamic analysis, may be performed on the modules. Module networks aim to represent the relationships between modules in network form; such networks have been constructed based on correlation between the expressions of the genes in the modules or physical interactions between the genes in the modules. Module dynamics aim to analyze and predict the temporal behavior of modules; dynamic models, such as state-space models, may be used for this purpose.

Figure 2. A representative network with 7 vertices ($v_1, v_2, \dots, v_7$) and 9 edges ($e_1, e_2, \dots, e_9$)

A path is a sequence of vertices and edges that start from one vertex and end at another vertex and passes through each vertex only once. An example of a path between vertices v_4 and v_1 is a sequence of v_4, e_3, e_2, v_1 . However, this is not the only possible path between v_4 and v_1 . Among the different possible paths between two vertices, vertex-independent paths do not share any vertices except the first and the last vertices, whereas edge-independent paths do not share any edges. For example, another path between v_4 and v_1 is a sequence of v_4, e_4, v_1 . Excluding the starting and ending vertices, this path does not share any edges or vertices with the previous path (v_4, e_3, e_2, v_1), therefore these two paths are both vertex and edge-independent. Note, all paths between v_4 and v_5 are both edge and vertex-dependent, because all paths have to pass through v_1 and e_1 . Two different modules can be identified in this figure, with the first module consisting of v_1, v_2, v_3, v_4 and the second module consisting of v_5, v_6, v_7 . The first module contains 5 edges and the second module contains 3 edges, whereas there is only one edge between the modules. The figure was generated by using Cytoscape[100].

Table 1

A summary of module-level analysis strategies.

Level of Modular-analysis	Input data	Methods	Output	Application examples
Module Identification	Network-based approaches	Hierarchical Clustering [7,28] Graph clustering[29] Clustering Coefficient[30] Seed expansion[39] Incorporating expression data [40] Traditional clustering [41] (hierarchical clustering, K-means) Bi-clustering[42] Model-based clustering[45, 46] Template-based matching [47, 48] Enrichment analysis[49–53] Supervised machine learning: Nonparametric regression [58, 59] Discriminant analysis[60] Partial least square regression [61] Decision tree/random forests [62]	Condense sub-network	Characterize modularity of biological systems [31,32]; Predict protein functions [33,34]; Identify functional groups of genes[35–38]; Identify functional groups of genes [41,43]; Identifying pathways activated under specific conditions [44];
	Expression-based approaches	Static gene Expression Time-series expressions	Clusters of co-expressed genes	
	Pathway-based approaches	Gene expressions in pathways	Differentially expressed pathways Phenotype-relevant pathways	Identify disease-related gene sets[54]; and genetic alterations[55]; Discovering novel pathways[56,57];
	Module network construction	Physical interactions (i.e. protein-protein) Expressions of genes in modules Time-series expression data	Networks of modules Difference between static module networks Dynamic models of module behaviors	Identifying pathway crosstalk [63]; Discovering coordinated transcriptional modules [64]; Identifying targets for modulating cell response [67]; Identifying pathways altered in disease progression [54,66]; Describing and prediction module dynamics[68];