Identifying Regulatory Networks
Lessons from Yeast to Humans

With the advent of high throughput techniques, data in molecular biology are accumulating at a staggering rate. Assays such as DNA and tissue microarrays illuminate the molecular mechanisms underlying cellular functions from multiple perspectives. This flood of data bears much promise for novel insights about the workings of living cells and organisms (1).

Despite early enthusiasm and extensive investment, we are still far from achieving this vision. Although the utility of microarrays for diagnosis and molecular characterization of disease has been extensively demonstrated (2–4), and studies are providing increasingly rich detailed signatures of biological and clinical phenotypes, a substantial gap remains between these phenomenological observations and a mechanistic understanding of cellular systems.

Here we focus on attempts to uncover the workings of regulatory (transcriptional) networks from transcription profiles. This requires a shift from describing patterns in the data (e.g., clustering and signatures of differentially expressed genes) toward finding biologically relevant models that explain the observed changes in transcription, thus entailing a methodologic shift in experimental design, analysis methods, and evaluation approaches. As always, a major hurdle is extracting a true biological signal from noisy measurements.

As in many previous cases, such developments are easier to carry out in the context of a unicellular model organism. Indeed, much progress has been made in recent years in elucidating the regulatory networks of bakers' yeast, *Saccharomyces cerevisiae*, using expression profiles, genomic sequences, chromatin immunoprecipitation data, and protein–protein interactions, all collected on a genome-wide scale. These data have been analyzed by a dazzling array of computational methods, with some surprising success stories, as we will show below.

Can these studies scale to mammalian systems? Clearly, mammalian systems are more challenging in several aspects. These include both the increased complexity of the regulatory mechanisms involved and the experimental limitations on manipulating these systems. Nevertheless, as we argue below, by carefully combining experimental design and computational tools we are now poised to directly tackle the challenge of inferring regulatory networks in such settings.

Lessons from Yeast

The common perspective on regulatory networks involves three major components: transcripts, *cis*-regulatory elements, and transcription factors (Figure 1), captured in expression profiles, sequence data, and transcription factor location assays (5). Different approaches to the reconstruction of regulatory networks focus on different combinations of these components and data types.

Virtually all computational works to date in yeast focus on three main expression datasets, which employ strikingly different experimental designs. The *Cell Cycle* dataset (6) provides expression profiles along a time course in synchronized populations of yeast cells; The Rosetta *Compendium* (7) examines a large collection of single gene knockout strains grown in rich medium; and the *Stress* response dataset (8) monitors short time courses following the application of a variety of environmental stresses. Together, these datasets follow the transcriptional programs involved in intrinsic temporal processes, following genetic perturbations, and in response to external stimuli. A major additional resource in yeast is genome-wide location analysis. This method can report on all the *in vivo* targets bound by a transcription factor under a given condition in a single assay. The main dataset (5) includes genome-wide binding information for most known transcription factors in cells grown in rich medium. Finally, in addition to the full genome sequence of *S. cerevisiae*, seven additional genomes from the *Saccharomyces* genus are available to date (9, 10). Although these datasets are of the highest current quality, they all involve significant levels of noise (both biological and experimental) and potentially confounding factors. Most analysis methods attempt to address these concerns both by employing multiple measurements with the same method and by combining different data types.

One common approach attempts to uncover regulatory circuits that associate *cis*-regulatory elements to target transcripts and use them to "explain" the observed expression patterns (11–17). Such methods can identify novel *cis*-regulatory elements, find the targets for a known or novel element, and identify the biological context under which elements modulate target gene expression, alone or in combination. For example, Tavazoie and colleagues (14) identify novel *cis*-regulatory elements that are enriched in clusters of co-expressed genes. In an attempt to uncover more complex combinatorial regulation, Segal and colleagues (17) find combinations of elements that characterize groups of co-regulated genes. Pilpel and colleagues (16) explore an alternative approach to combinatorial regulation, by identifying conditions under which pairs of elements act in synergy. The same general framework can be extended to handle transcription factor location data instead of *cis*-regulatory elements. For example, Bar-Joseph and colleagues (18) find combinations of transcription factors whose co-location can "explain" coherent expression patterns. Finally, by integrating sequence, expression, and location data, Segal and colleagues (19) identify active combinations of transcription factors, their target genes, and the *cis*-regulatory elements that mediate this regulation.

Regulatory circuits highlight a particular mechanistic aspect of regulation, leveraging our understanding of this aspect (e.g., binding of transcription factors to their cognate elements). However, these methods are constrained to specific data types (e.g., location data and promoter regions) as well as limited in their ability to identify other, indirect, regulatory relations. Complementary approaches focus on interaction networks that describe dependencies between genes, primarily based on their transcriptional profiles (20–27). These studies allow us to detect both direct and indirect targets of regulatory proteins and the conditions under which such regulation occurs. A key component in these methods is the use of statistical tools to identify dependencies between genes. Somewhat surprisingly, these methods can detect both direct transcriptional regulation and indirect regulation, for example, between a protein kinase and its downstream targets. Such methods are particularly suited for experimental systems based on different perturbations, which invoke different but overlapping regulatory programs (Figure 2A).

Inferring interaction networks is based on a strong simplifying assumption that the expression level of the regulator reflects its

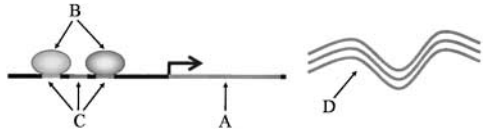

Figure 1. Key components in regulatory networks. The expression of a gene (A) is regulated by transcription factors (B) that bind to specific *cis*-regulatory elements (C) in the gene's promoter, modulating the amount of RNA (D) transcribed from the gene.

activity. Although this is true more often than one might expect, many regulatory relations may be missed. Activity models attempt to distinguish between the activity of a regulator and its expression level (28–30). Such recently developed methods can provide insight into the unmeasured activity profiles of regulators and their quantitative effects on the targets. For example, Liao and colleagues (28) used the regulatory architecture suggested from genome-wide location data to identify the activity levels of 11 cell cycle transcription factors from expression data. Note that such refinements pose harder inference challenges and often require partially known network architecture as a starting point. More refined experiments, such as detailed time courses, may partially alleviate this requirement.

Changes in regulator activity reflect multiple biochemical mechanisms, such as protein degradation, phosphorylation, and relocation that are associated with upstream molecular events. Insights into this relation between transcriptional regulation and signal transduction can be gained by combining activity profiles with prior knowledge about signaling pathways (31) or high-throughput protein–protein interactions maps (32). With the advances in proteomics, these methods will help us elucidate signal transduction mechanisms and resolve cross talk between different pathways.

A crucial aspect relevant to all these approaches is interpretation and evaluation of the inferred networks. Such evaluation allows researchers to gain confidence in the abilities of the methods, understand their limitations, and identify novel testable hypotheses implied by those results. The evaluation of results in yeast is facilitated by the large body of knowledge on regulatory mechanisms, allowing us to contrast inferred results with previously established findings. Note that such comparison is nontrivial, because detailed molecular studies report elaborate details, which may not be captured in full by high-throughput analysis.

The type of evaluation and the potential hypotheses directly depend on the nature of the inferred network. For example, when evaluating a regulatory circuit, we ask if the suggested *cis*-regulatory elements are known to regulate the identified targets. If the element is novel, this suggests a reporter gene experiment, comparing intact and mutated novel elements. In contrast, interaction networks do not describe a specific molecular mechanism and thus their evaluation is more elaborate. For example, we consider one component of the interaction network identified by Segal and colleagues (26) (Figure 2A). Tpk1, a catalytic subunit of the cAMP-dependent protein kinase (PKA), is identified as a regulator of a set of genes involved in energy, osmolarity, and cAMP signaling. These processes are known to be indirectly regulated by this signaling molecule. This finding is further supported by the presence of an STRE motif, known to be bound by transcription factors that are regulated by Tpk1 (Figure 2B).

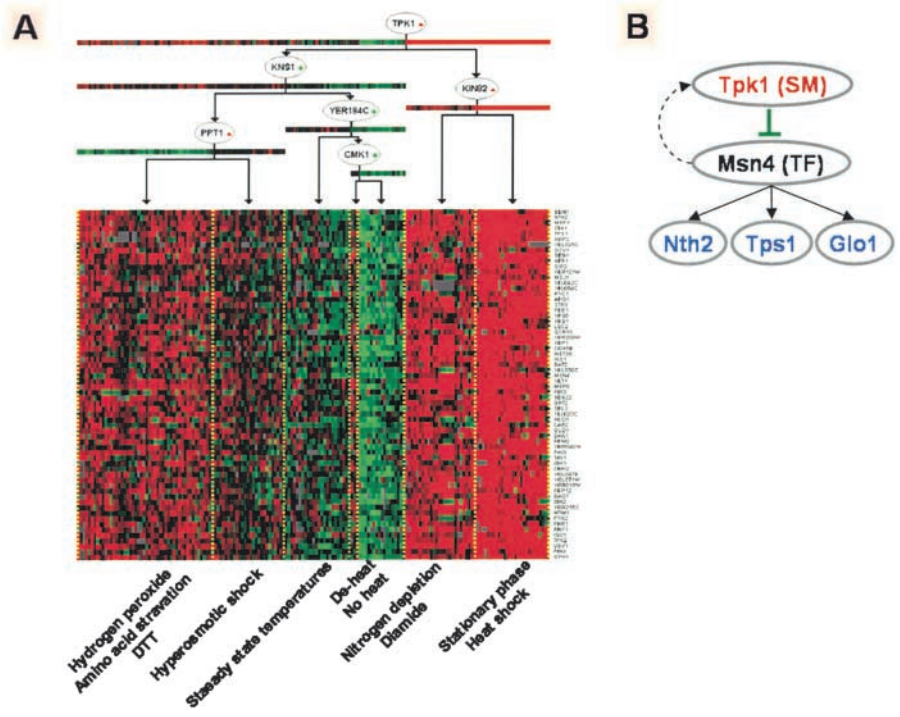

Figure 2. A single module in an inferred interaction network. (*A*) A regulatory module for energy, osmolarity, and cAMP signaling in yeast response to environmental stress. Our method identified a set of genes with coherent expression that function in energy, osmolarity, and cAMP signaling and inferred their shared regulation mechanism. The shared regulation program (*top*) is shown as a tree, where each node represents a regulator (for example, Tpk1) and a query of its qualitative value (for example, *red upward arrow* next to Tpk1 for "is Tpk1 upregulated"?). The expression of the regulators themselves is shown below their respective node. In the gene expression profiles (*bottom:* genes, *rows*; arrays, *columns*) the arrays are arranged according to the regulation tree. For example, the rightmost leaf includes the arrays in which both Tpk1 and Kin82 are upregulated. Contexts that consist primarily of few types of experimental conditions are labeled below. (*B*) The method identified indirect regulation by Tpk1, a signal transduction molecule. Based on previous biological knowledge (*see* Ref. 26), we construct the following scenario. The signaling molecule (Tpk1) inhibits activity of its target transcription factor (e.g., Msn4), by post-translational modification. The transcription factor (Msn4, which binds to the STRE *cis*-regulatory element) induces transcription of the module's genes (*e.g.*, Nth2, Tps1, and Glo1) as well as of the signaling molecule. Thus, the expression level of the signaling molecule changes concordantly with that of its targets and is correctly inferred by the method as the module's regulator. But, because both signaling molecule and the targets are upregulated, the method predicts that Tpk1 activates the module, in contrast to its actual inhibitory role.

Finally, the method correctly predicts that Tpk1 is actively regulating these targets in heat shock, high osmolarity, stationary phase, and nitrogen depletion. Other novel hypotheses of this form can be tested, for example, by examining expression profiles in yeast knockout strains lacking the regulator grown under the predicted conditions. Differential expression of the predicted targets under these conditions can validate the method's predictions. Indeed, the predictions for three regulators—a transcription factor, a protein kinase, and a protein phosphatase—were experimentally validated (26).

Challenges and Prospects in Mammalian Systems

Despite the large availability of microarray studies in mammalian systems, little work has been done so far on inferring regulatory networks. We attempt to estimate the prospects of scaling this work to mammals, bringing into account the complexities involved.

Although yeast is undoubtedly an important eukaryotic model, mammalian molecular networks pose many additional challenges. For example, the genome is significantly less compact, with longer promoters and intergenic regions. Promoters and regulatory sequences are notoriously difficult to find, both experimentally and computationally, as they are dispersed and remote. Furthermore, many of the responses involve complex combinatorial regulation by several transcription factors. Higher order chromatin organization plays a key and mostly uncharacterized role in gene regulation by multiple mechanisms. Moreover, the signal transduction networks modulating transcriptional responses to exogenous and endogenous stimuli are significantly larger and more elaborate in mammalian systems. Many of these additional elements are still not measured by high throughput methods and have confounding effects on existing measurements.

This molecular complexity is partially due to multicellularity, where different configurations of the molecular networks participate in differentiating and maintaining a wide range of cellular phenotypes. Multicellularity also affects the available data. Most tissues are composed of a heterogenous population of cells, from different lineages and at various differentiation states that are organized in elaborate spatial and temporal patterns. This heterogeneity is a potentially confounding factor in many whole tissue studies. Changes in the composition of cell populations may lead to differential expression patterns that do not reflect a concomitant change in cellular state. Furthermore, when such changes in cellular states do occur, they are often obscured by the aggregate measurement of several superimposed responses. For example, lung tissue includes epithelial cells, fibroblasts, smooth muscle cells, endothelial cells, neutrophils, and lymphocytes, and its cellular content can change rapidly in different physiologic conditions.

A more fundamental problem of complex regulatory networks in multicellular organisms is the multifaceted and context specific nature of biological responses. For instance, the transforming growth factor (TGF)-β response is markedly different in distinct cell types or transient states, although they share many of the building blocks of the underlying mechanism. Moreover, some cell types, such as fibroblasts, can transdifferentiate in response to stimuli and change their subsequent behavior. These cellular responses are subsequently integrated into a tissue- or organ-wide coherent response, involving intercellular communication mechanisms. Thus, to achieve meaningful results, a well-defined cellular state is often required. This, however, poses significant experimental challenges, in particular for *in vivo* studies. Even under the most controlled experimental conditions, this inherent context specificity will pose difficulties for interpretation. An inconsistency between new predictions and previous findings can almost always be excused as a different variant of the response (e.g., a specific protein can be both pro- and antiapoptotic depending on context). Finally, integrating the inferred networks across multiple systems and generalizing their findings has to be done carefully, so as not to confound distinct responses.

These challenges seem to put into question the utility of the methods we discussed above in mammalian context, perhaps explaining the scarcity of such studies. Nevertheless, a few studies and preliminary results report some promise. Several researchers have identified regulatory circuits in expression data from synchronized HeLa cells (17, 33, 34), finding both known cell cycle regulatory elements and targets and suggesting novel ones. In our preliminary work we found interaction networks using a dataset measuring the response to various cytokines of three primary cell types (normal human bronchial epithelial cells, normal human lung fibroblasts, and bronchial smooth muscle cells) in culture (N. Kaminski, unpublished data). For example, our method identified the correct role of Smad3 and JunB in the TGF-β response in fibroblasts. This included a module, which had both Smad3 and JunB as key regulators. The genes in this module contain many known TGF-β targets (such as cTGF) and the decision tree below SMAD3 and JunB involves experiments that include early (6 h) and late (24 h) response to TGF-β in lung fibroblasts.

Are these merely anecdotal findings or is there a prospect for systematic achievement?

Much effort in recent years has been devoted to amassing samples of diseased tissue from human subjects, especially tumor tissues. Although such studies suggest many candidate genes for diagnosis and therapy, they are not well suited for inferring detailed networks underlying disease for several reasons. These include confounding genetic and environmental factors (unavoidable in human subjects), heterogeneous cell populations, and the lack of a time course of disease progression. More importantly, they usually comprise of many "similar" samples and hence lack the detailed perturbations necessary to follow regulatory events. Thus, *in vivo* studies to infer networks should also be carefully designed to minimize confounding genetic and environmental factors (35), to incorporate multiple perturbations (36), and to monitor the progression of disease. Finally, various molecular and microdissection techniques (37) may be employed to reduce the heterogeneity of examined samples.

Interestingly, all these early successes were achieved by analyzing data collected from homogenous cell populations in culture under well-defined conditions. Indeed, many of the aforementioned pitfalls can be avoided by careful design of experiments. Although our ultimate goal is to elucidate the workings of molecular networks in the context of the whole organism, we believe that studies on well-defined cell populations in controlled conditions are a crucial first step. Depending on the specific biological system, these can be done *in vitro* in cell lines or primary cultured cells, or *in vivo* (e.g., in hemtopoietic cells [36]). Such studies can elucidate the "basic" building blocks of cellular reponses. With these in hand, we will be better poised to interpret results from composite responses in heterogenous cell populations *in vivo*.

Although the availability of data from carefully designed studies is crucial for successfully identifying regulatory networks, so is the development of more sophisticated computational methods. These should address the increased complexity of mammalian systems. For example, most current methods for detecting *cis*-regulatory elements are better suited for compact, less complex genomes with short intergenic regions. Naively

applying such methods to complex mammalian genomes has had limited success, but extending such methods to look for clusters of elements has shown promising results (34, 38, 39). Clearly, better modeling of promoter organization and employing comparative genomics approaches (34, 40) is needed.

Another computational challenge is dealing with whole tissue samples. These will always be required, because much insight can only be gained from studying the interactions between cells in their native environment. We need methods that build on expression patterns in pure cell populations and information from other sources (e.g., *in situ* hybridization) to deconvolve individual responses from whole tissue measurements (*see*, for example, Ref. 37). More generally, we emphasize the need for "transferable" models that will allow us to combine insights from multiple experimental systems addressing related phenomena. For example, we would like to combine similar pathologic conditions (e.g., fibrosis or cancer) in different tissues or similar biological conditions in different organisms (e.g., a mouse animal model and human clinical samples).

Finally, to understand the commonalities and differences between various settings, we need to directly characterize the underlying biological context that determines the mode of response of different regulatory components. This context might involve differentiation patterns, concentration levels of different molecules, external perturbations, and genetic state. Although context is typically invoked to explain variation in responses, we still need a rigorous approach to define and identify biological contexts.

Conclusion

We briefly reviewed the current research of computational methods for inferring regulatory networks from high throughput data, in particular gene expression profiles, and discussed encouraging success stories in unicellular eukaryotes. Although mammalian systems raise several new challenges, we believe that they are not insurmountable. In short, a combination of realistic goals and expectations, thoughtful experimental design using the full range of cell culture, animal models and human samples, and novel computational methods will allow us to successfully infer relevant regulatory networks.

References

1. Lander, E. S. 1999. Array of hope. *Nat. Genet.* 21:3–4.
2. Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503–511.
3. van de Vijver, M. J., Y. D. He, L. J. van't Veer, et al. 2002. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347: 1999–2009.
4. Beer, D. G., S. L. Kardia, C. C. Huang, et al. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8:816–824.
5. Lee, T. I., N. J. Rinaldi, F. Robert, et al. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298:799–804.
6. Spellman, P. T., G. Sherlock, M. Q. Zhang, et al. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9:3273–3297.
7. Hughes, T. R., M. J. Marton, A. R. Jones, et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102:109–126.
8. Gasch, A. P., P. T. Spellman, C. M. Kao, et al. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11:4241–4257.
9. Cliften, P., P. Sudarsanam, A. Desikan, et al. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301:71–76.
10. Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–254.
11. Brazma, A., I. Jonassen, J. Vilo, and E. Ukkonen. 1998. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.* 8:1202–1215.
12. Barash, Y., and N. Friedman. 2002. Context-specific Bayesian clustering for gene expression data. *J. Comput. Biol.* 9:169–191.
13. Hughes, J. D., P. W. Estep, S. Tavazoie, and G. M. Church. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.* 296: 1205–1214.
14. Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22:281–285.
15. Bussemaker, H. J., H. Li, and E. D. Siggia. 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* 27:167–171.
16. Pilpel, Y., P. Sudarsanam, and G. M. Church. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29: 153–159.
17. Segal, E., R. Yelensky, and D. Koller. 2003. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* 19:I273–I282.
18. Bar-Joseph, Z., G. K. Gerber, T. I. Lee, et al. 2003. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* 21:1337–1342.
19. Segal, E., Y. Barash, I. Simon, N. Friedman, and D. Koller. 2002. From Promoter Sequence to Expression: A Probabilistic Framework. *In* Proc. of the 6th Annual International Conference on Computational Molecular Biology. ACM Press, New York.
20. Friedman, N., M. Linial, I. Nachman, and D. Pe'er. 2000. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7:601–620.
21. Hartemink, A. J., D. K. Gifford, T. S. Jaakkola, and R. A. Young. 2002. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* 7:437–449.
22. Hartemink, A. J., D. K. Gifford, T. S. Jaakkola, and R. A. Young. 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* 6:422–433.
23. Kim, S. Y., S. Imoto, and S. Miyano. 2003. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief. Bioinform.* 4:228–235.
24. Pe'er, D., A. Regev, G. Elidan, and N. Friedman. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17:S215–S224.
25. Pe'er, D., A. Regev, and A. Tanay. 2002. Minreg: inferring an active regulator set. *Bioinformatics* 18:S258–S267.
26. Segal, E., M. Shapira, A. Regev, et al. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34:166–176.
27. Tamada, Y., S. Kim, H. Bannai, et al. 2003. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics* 19:II227–II236.
28. Liao, J. C., R. Boscolo, Y. L. Yang, et al. 2003. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA* 100:15522–15527.
29. Perrin, B. E., L. Ralaivola, A. Mazurie, et al. 2003. Gene networks inference using dynamic Bayesian networks. *Bioinformatics* 19:II138–II148.
30. Ronen, M., R. Rosenberg, B. I. Shraiman, and U. Alon. 2002. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA* 99:10555–10560.
31. Wang, W., J. M. Cherry, D. Botstein, and H. Li. 2002. A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* 99:16893–16898.
32. Ideker, T., O. Ozier, B. Schwikowski, and A. F. Siegel. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18:S233–S240.
33. Elkon, R., C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh. 2003. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.* 13:773–780.
34. Sharan, R., I. Ovcharenko, A. Ben-Hur, and R. M. Karp. 2003. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* 19:I283–I291.
35. Tarnavski, O., J. R. McMullen, M. Schinke, et al. 2004. Mouse cardiac surgery: comprehensive techniques for the generation of mouse models of human diseases and their application for genomic studies. *Physiol. Genomics* 16: 349–360.
36. Gilman, A. G., M. I. Simon, H. R. Bourne, et al. 2002. Overview of the Alliance for Cellular Signaling. *Nature* 420:703–706.
37. Stuart, R. O., W. Wachsman, C. C. Berry, et al. 2004. *In silico* dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proc. Natl. Acad. Sci. USA* 101:615–620.
38. Frith, M. C., U. Hansen, and Z. Weng. 2001. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* 17:878–889.
39. Sinha, S., E. Van Nimwegen, and E. D. Siggia. 2003. A probabilistic method to detect regulatory modules. *Bioinformatics* 19:I292–I301.
40. Kent, W. J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100:11484–11489.