Pathway Analysis of Microarray Data

3

Matteo Pellegrini and Shawn Cokus

Abstract

During the past decade remarkable new techniques for transcriptional profiling have been developed. These include transcriptional profiling using hybridization microarrays as well as methods to sequence transcribed RNAs. No matter which technology is used, these experiments generate data on thousands of genes across multiple conditions and therefore the analysis of these data is often a daunting task. One of the most promising avenues for interpreting large datasets of expression profiles involves pathway-based analysis. Although pathway analysis of expression data is a relatively new field, many important advances have been made over the past few years. Below we outline some the most significant developments in this area of research.

Introduction

Pathways are collections of genes and proteins that perform a well-defined biological task. For instance, proteins that work to successively synthesize metabolites within a cell are grouped into metabolic pathways. Similarly, proteins that are involved in the transduction of a signal from the cell membrane to the nucleus are grouped into signal transduction pathways. These pathways have been established through decades of molecular biology research and are collected in a variety of public pathway repositories (Kanehisa *et al*., 2004; Ashburner *et al*., 2000).

Since the number of known pathways within cells is significantly smaller than the number of genes that is typically profiled, the transformation of data from a gene-centric view to a pathways-centered one represents a dramatic reduction in the number of dimensions. Such a reduction allows a biologist to interpret and understand the data in a manner that is not possible when it is viewed as a collection of individual genes.

Although pathway analysis of expression data is a relatively new field, many important advances have been made over the past few years. Below we outline some the most significant developments in this area of research. These include analyses that attempt to identify the pathways that are overrepresented among significantly perturbed genes in an experiment along with methods that attempt to identify pathways and networks of molecular interactions directly from expression data. Despite the fact that these analyses will undoubtedly continue to evolve rapidly over the next few years, they have already enhanced our ability to understand the biology that underlies complex experiments.

Term enrichment analysis

A typical analysis of microarray expression data generates a long list of genes that are significant according to some criterion. These may be, for example, genes that are differentially regulated in a ratio experiment, or genes that are significant in an analysis of variance (ANOVA) of groups of samples. No matter how the list is generated, it is usually a daunting task to interpret the underlying biology because these lists tend to contain hundreds of genes. In principle, one could search the literature for each gene in the list to attempt to uncover common relationships among them. However such an approach would inevitably require many hours of research without guaranteeing that the search was comprehensive.

Several tools have emerged to automate this type of analysis. These programs rely on a priori classifications of genes into biological function groups. The Gene Ontology (GO) Consortium (Ashburner *et al*., 2000) generates one of the most widely used of these classifications. GO terms are related to each other through a directed acyclic graph (DAG). That is, most terms have both parent and child terms. The parent terms are more general and inclusive than the child terms. For instance, the parent term "ribonuclear protein complex" (GO id 0030529) has a child term "ribosome" (GO id 0005840). The ontology is separated at the highest level into three separate graphs that contain terms for biological processes, cellular components, and molecular functions. To date, GO represents one of the most comprehensive collections of pathway annotations.

An example of an application that uses GO to automatically perform term enrichment analysis is the Expression Analysis Systematic Explorer (EASE) (Hosack *et al*., 2003). This tool measures the overlap between a list of genes with GO biological process categories. The significance of the overlap is calculated using the hypergeometric distribution to estimate the probability of finding at least the observed overlap by chance. As an example, the authors computed the GO terms associated with a gene expression study by Kayo *et al*. (2001) on the influence of aging and caloric restriction to the transcriptional profile of skeletal muscle in rhesus monkey. They find that the terms computed with EASE (mitochondrion and electron transport) matched the terms Kayo *et al*. had found through a manual literature search. However, in contrast to the approximately 200 hours required for the literature search, EASE was able to perform the analysis in a few minutes.

Other applications that perform a similar analysis to EASE include GoMiner (Zeeberg *et al*., 2003), MAPPFinder (Doniger *et al*., 2003), FatiGO (Al-Shahrour *et al*., 2004) and GoSurfer (Zhong *et al*., 2004). These programs differ in the type of gene identifiers that they recognize, the graphical display of the analysis results, the metric that is used to score the enrichment of terms, and the operating system that they work on.

There are many different identifiers that are used to denote genes: gene symbols, Entrez gene identifiers, Affymetrix probe identifiers, SwissProt identifiers, etc. Translating from one id type to another is often a necessary step before any analysis is performed since different applications recognize different identifiers. For example, GoSurfer recognizes Affymetrix probe ids while GoMiner recognizes HUGO gene names. A universal identifier translation tool would be extremely useful but is currently not available; therefore one must

manually construct translation files or only use programs that recognize the particular identifiers that one is using.

The output of term enrichment analysis typically comes in two forms: a list of terms that are enriched and a graph of GO terms in which the terms are color-coded according to their statistical significance. If the enriched terms are just reported as a list the complex relationships between them are not apparent and one may not realize that the significant terms are actually related within the ontology. In contrast, if the output is displayed as a network, then one immediately sees how the terms are related to each other, but may not immediately realize which ones are the most enriched. It is therefore ideal to present the output in both formats. An example of the graph output of GoMiner is presented in Fig. 3.1.

Certain programs may be used directly on the web (e.g., FatiGO) while others must be downloaded and installed. Among the latter, some work only on the Microsoft Windows operating system (e.g., GoSurfer) while others are written in Java

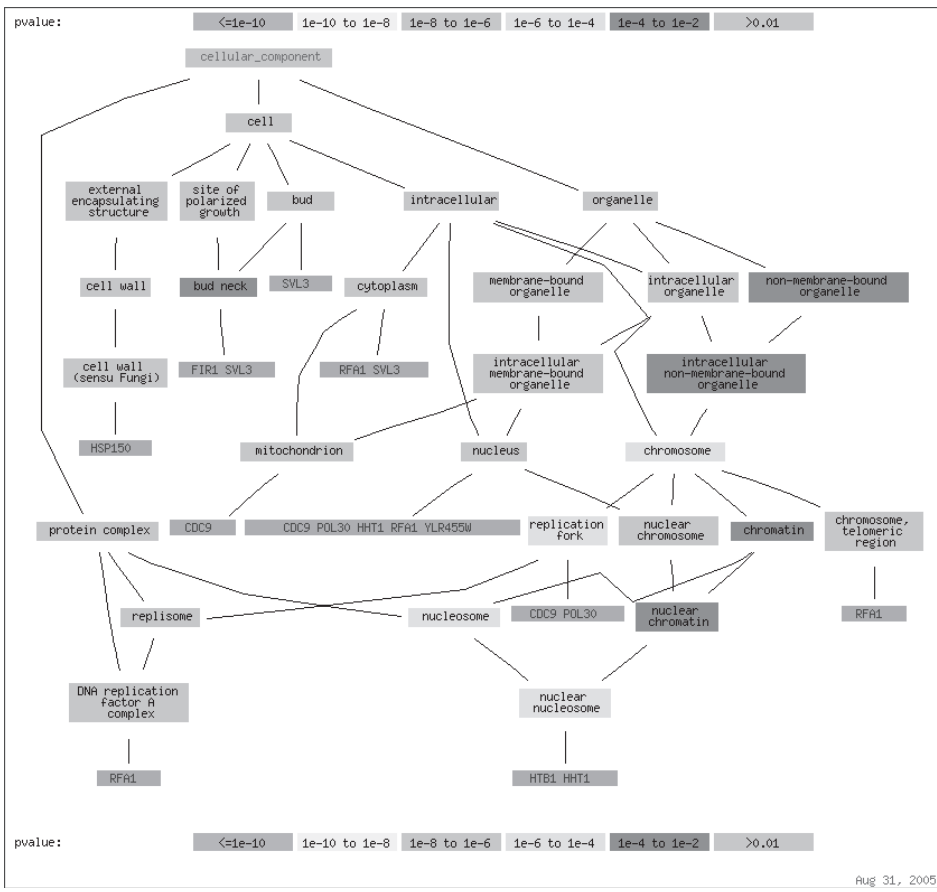

Figure 3.1 The network of enriched GO terms in the cellular component ontology. Blue and cyan indicate that the term is enriched. The figure was generated using the SGD Gene Ontology Term Finder (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder). Genes that are cyclical in synchronized yeast cells were input into the program.

and are therefore platform independent (e.g., GoMiner).

Gene set analysis

In the previous section we discussed using the overlap of significant genes in a microarray analysis with functional groups to identify the groups whose members are overrepresented among these genes. In this case, we only consider genes that are deemed to be significant by some threshold and then ignore the particular numerical values. Although this is often a convenient way to select significant genes, in the process we are effectively converting continuous data (gene log ratios or *p*-values) to binary data (significant or not) and thus losing information. A variety of methods have been developed recently that consider the actual numerical values of genes when attempting to uncover which pathways are of interest.

In one recently published example, the authors attempted to uncover pathways that are differentially expressed between two patient populations: patients with and without a specific disease (Mootha *et al*., 2003). The approach used was named gene set enrichment analysis (GSEA) and attempted to identify the pathway that contained the most differentially expressed genes between the two populations. The analysis was performed on data collected from healthy and diabetic patients. The authors first ranked genes according to the expression difference between the two groups. They then used a Kolmogorov-Smirnov statistic to determine which set of genes had high-ranking members. They were able to estimate the probability of each observation by comparing the real scores to those of randomly permuted data. The analysis identified the oxidative phosphorylation pathway as the most differentially expressed and demonstrated that the transcription factor PGC1-α, a regulator of this pathway, and mutations in it correlate with diabetes.

Other approaches similar to GSEA have also been developed. For instance, the program GOMapper computes the significance of the expression of a gene set by computing the ratio of the average expression of genes in the set to the average expression of all genes in the array (Smid *et al*., 2004). A similar approach is entitled functional class scoring (FCS), wherein the enrichment of each GO term is calculated by estimating the likelihood of observing the product of probabilities of each individual gene associated with the term (Pavlidis *et al*., 2004). The probabilities associated with each gene are generated from an error model and estimate the likelihood that the gene is perturbed. Monte Carlo simulations estimate the distribution of the products of probabilities to enable the computation of the expectation that a given GO term is enriched for perturbed genes.

The methods described above focus their analysis on datasets in which each gene is assigned a single value. However, these methodologies may be extended to large datasets where gene expression is measured across multiple experiments. In this way, the traditional representation of clustered heat maps of genes versus experiments may be applied to data that measure the activity of pathways across experiments. One example of this type of approach is the map of cancer modules generated by Segal *et al*. (2004). The authors assembled a dataset of 1975 published microarray experiments that span 22 tumor types. They searched for modules that were significantly active within a subset of experiments. Modules are defined as groups of genes that share a common biological function and are derived by combining multiple sources of gene groupings including GO, KEGG

(Kanehisa *et al.*, 2004), and GENMAPP (Doniger *et al.*, 2003).

Among the many conclusions that they could draw from the final module map, they highlight the cell cycle module as active across multiple tumor types, consistent with the observation that all these tumor types involve rapidly dividing cells. Similarly, many tumor types have active osteoblastic module, consistent with the fact that many of these tumors metastasize to bone. In contrast, other modules are specific to tumor types. For example, modules that involve neuronal processes are only repressed in a subset of tumors and are otherwise not active. In general, this type of analysis demonstrates that a module level heat map is significantly more interpretable than a gene-level heat map and therefore this approach represents a useful tool for biologists that are trying to cope with large sets of expression microarrays.

Pathway coherence

In order for GO terms or other pathway groupings to appear activated in the previous analyses, the genes within the pathway must be co-regulated. That is, it is unlikely that a random group of genes will ever show up in a pathway analysis, since the genes are independent of each other and unlikely to be perturbed together. In contrast, if a group of genes acts as a single unit (all perturbed together or unperturbed together) then they are far more likely to appear active. A pathway whose genes are co-regulated may also be called a *coherent pathway*.

It seems reasonable that, if we could determine a priori which pathways are coherent and which are not, it might be advantageous to analyze only coherent pathways. One possible metric to measure coherence was developed by Yang *et al.* (2004), and involves measuring the fraction of gene pairs with a pathway that are significantly co-expressed across a set of experiments. Correlation coefficients between pairs of genes are computed and the probability of observing such a correlation or higher is estimated. One may then compute whether the fraction of statistically significant correlations in a pathway group is greater than in a random group of the same size.

Yang *et al.* performed this experiment with normal and tumor tissue samples. They searched for pathways defined by KEGG that had significant coherence. The found that metabolic pathways and protein complexes are coherent while signal transduction pathways are not. This is not surprising since one expects that both metabolic pathways and protein complexes should contain co-regulated genes, while signal transduction pathways on the other hand are controlled by post transcriptional modifications (e.g., phosphorylation) rather than transcription. A list of the coherent pathways they identified is shown in Table 3.1.

However, not all the genes within a metabolic pathway are co-regulated. Ihmels *et al.* investigated in great detail which components of metabolic pathways are coherent (Ihmels *et al.*, 2004). For example, they found that of the 46 genes assigned to the glycolysis pathway in KEGG, only 24 were correlated in their expression patterns across one thousand diverse experiments. These 24 genes are linearly arranged along the central part of the pathway. They find that in general the central components of metabolic pathways are the most coherent part of the pathway, and that such a central component represents a set of linear reactions.

Ihmels *et al.* also extended their analysis to isozyme pairs contained within metabolic pathways (i.e., pairs of genes with similar sequences that perform slightly dif-

Table 3.1 Coherent pathways

Fructose and mannose metabolism
Sterol biosynthesis
Urea cycle and metabolism of amino acids
Pyrimidine metabolism
Arginine and proline metabolism
Glycoprotein degradation
Ubiquinone biosynthesis
Inositol phosphate metabolism
Sphingoglycolipid metabolism
Nicotinate and nicotinamide metabolism
Apoptosis
Starch and sucrose metabolism
Valine, leucine and isoleucine degradation
Lysine biosynthesis
Propanoate metabolism
Butanoate metabolism
Protein export
Photosynthesis
Aminoacyl-tRNA biosynthesis
Oxidative phosphorylation
ATP synthesis
Ribosome
Proteasome

ferent functions). They find that most of the pairs were separately co-regulated with alternative sub pathways. In other words, KEGG pathways may often be broken up into distinct sub pathways that utilize different members of isozyme pairs. They also identify genes that are co-expressed with the sub-pathway and are therefore likely components of the pathway. Often such genes code for transporters of the metabolites utilized in the pathway or transcription factors that regulate the pathway.

In summary, it is not only possible to define which pathways are coherent, but also to refine these pathways so that they become more coherent. This involves identifying the most coherent core of the pathway and then extending these cores with additional genes that were not initially associated in the pathway but are co-expressed with the core.

Reconstruction of networks using expression data

In the preceding sections, we have discussed techniques for using pre-existing pathway information to interpret microarray expression data. An alternative approach attempts to reconstruct pathways directly from the data. In other words, the previous sections were aimed at supervised analysis whereas here we discuss unsupervised approaches.

One of the first approaches developed to analyze microarray expression data was the clustering program of Eisen *et al*. (1998). They computed pairwise "distances" between genes and clustered genes based on these distances. The approach proved to be remarkably successful in facilitating the interpretation of data. Clusters typically contain genes that function within related pathways or biological processes. It was therefore possible assign functions to previously uncharacterized genes based on the functions of the genes it clusters with.

The reason that pairwise clustering approaches facilitate our interpretation of expression data is that genes with correlated expression tend to function within the same biological process. However, the converse is not often true. That is, genes that are known to function together are not always correlated. In fact, in the majority of cases genes that function together

are not significantly correlated. This behavior is consistent with the observation of the previous section that only a minority of pathways are coherent, and that only a subset of a typical coherent pathways is in fact truly coherent.

To overcome this difficulty, several methods have been developed to search for functional associations between genes that are not correlated in their expression patterns. One such approach has been to consider higher-order relationships between genes beyond the pairwise ones used in the original clustering methods. For instance, Zhou *et al*. have developed what they term *second-order analysis* (Zhou *et al*., 2005). Rather than simply calculating pairwise correlations between genes within datasets, the authors compute the correlation between the correlations of two pairs of genes across multiple datasets. In other words, it is possible to first compute the correlation between genes *A* and *B* across datasets *X* and *Y*: c*AB*(*X*), c*AB*(*Y*). It is then possible to identify a second pair of genes, C and D, whose correlations in datasets *X* and *Y* are correlated with c*AB*(*X*) and c*AB*(*Y*). The two pairs of genes may have statistically significant second-order correlations even though the pairwise correlations between *A* and C or *B* and *D* are not significant. Thus, it is possible to find relationships between genes that are not captured by pairwise correlations.

To compute second-order relationships Zhou *et al*. looked at 618 yeast expression arrays that comprised 39 datasets. The analysis revealed 5,142 pairs of genes with significant correlations across some of these datasets and 178,799 statistically significant quadruplets. They observed that 83% of these quadruplets were functionally homogeneous by measuring how often they shared GO terms, implying that the genes participate within the same pathway. In contrast, only 53% of the pairwise relationships were functionally homogeneous. Statistically significant quadruplets seem to group genes into pathways more effectively then pairs. Clustering second-order profiles allows the authors to assign genes to functions more effectively than clustering using Eisen's original approach.

Finally, Zhou *et al*. also apply this approach to transcription factor modules. These are the sets of genes controlled by specific transcription factors. They show that applying second-order analysis allows them to infer regulation motifs in which two transcription factors are being controlled by a third or where one transcription factor is controlling another. They demonstrate that these types of relationships exist between cell cycle transcription factors. For example, the SWI4 and NDD1 modules are correlated in second-order analysis even though none of the genes in one module are correlated with any of the genes in the other. The second-order relationship implies that SWI4 is controlling the transcription of NDD1, and hence the genes regulated by these factors are related in a second-order fashion (Fig. 3.2).

An approach related to second-order analysis and developed by Li *et al*. is named *liquid association* (Li *et al*., 2004; Li, 2002). The idea underlying this method is that uncorrelated genes may in fact appear correlated when their relationships are conditioned on the state of a third gene. For example, two genes *A* and *B* may appear uncorrelated over a large dataset. However, the pair might appear positively correlated when the values of a third gene C are high and negatively correlated when the values of C are low. This relationship between *A*, *B*, and C may arise if C is somehow controlling the expression of both *A* and *B*.

To illustrate the utility of liquid association the authors looked at oncogene

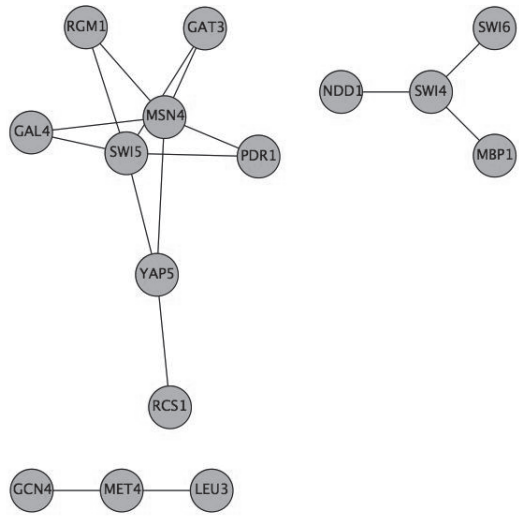

Figure 3.2 Network of second-order interactions between transcription factor modules. Each transcription factor is known to control a group of genes based on transcription factor binding data (Harbison *et al*., 2004). These gene modules are related by second-order analysis (i.e., the genes from one module are not correlated to the genes in another module, but the correlations between the genes in the two modules are correlated across conditions). These relationships imply that the transcription factors are either interacting or controlling each other, or being controlled by a third factor.

P53. It is known from the literature that *P53* interacts with *TP53INP1* (which encodes a P53-inducible nuclear protein) and *TPBP1* (which codes for P53-binding protein 1). However, these three genes show very low correlation across expression datasets. The authors therefore searched for a fourth gene that possibly interacted with these three and generated a high liquid association score. Their top candidate was *SMARC4*, a gene that encodes a protein that is known to interact with *P53*. The correlations between the three initial genes were therefore significant when conditioned on the expression of *SMARC4*. This may be due to *SMARC4*'s participation in the SWI/SNP transcription factor complex that is necessary for the activation of P53-mediated transcription.

Both second-order analysis and liquid association attempt to identify relationships between small numbers of genes. However, methods have also been developed to reconstruct large networks of gene associations. One such approach utilizes the formalism of Bayesian networks to infer gene networks (Friedman, 2003; Friedman *et al*., 2000). Bayesian networks model the probability of any state of the system based on the conditional probability distribution of each gene with respect to "parent" genes:

$$P\left(X_1,\ldots,X_n\right)=\prod_i P\left(X_i\middle|U_i\right) \qquad (3.1)$$

where gene expression values are denoted by X_i and corresponding parent genes U_i. The relationships between genes are to form a DAG that must be inferred from the data. Inferring the DAG is often computationally expensive. Furthermore, many DAGs provide solutions of roughly the same quality so it is customary to construct an "average network" from all the nearly

optimal inferred DAGs. An example network reconstructed using this technique is shown in Fig. 3.3.

Another approach that has been recently implemented to reverse engineer entire networks is called ARACNE, which stands for Algorithm for the Reconstruction of accurate cellular networks (Basso *et al.*, 2005). This approach uses mutual information to identify pairs of genes that are likely co-regulated. It then applies a filtering step to eliminate pairwise relationships that are likely to be indirect. This filtering step is performed using the "data processing inequality" from data transmission theory. The authors claim that the resulting network is enriched for direct interactions. They also compare this approach to Bayesian networks and demonstrate that in certain cases it yields superior results.

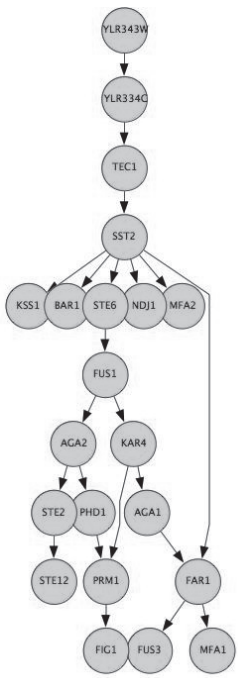

Figure 3.3 Yeast mating Bayesian network constructed by Friedman *et al*. (2003) from a yeast expression dataset of 300 mutant strains. Many of the genes are involved in mating. For instance, we see the mating pheromone a-factor (Mfa1 and Mfa2), along with genes involved in cell fusion (Fus1 and Fus3) and a protease that allows cells to recover from alpha-factor-induced cell cycle arrest by degrading alpha factor (Bar1).

ARACNE was recently applied to the reconstruction of networks in human B-cells. The analysis was performed on 336 B-cell expression arrays that represented a wide collection of normal, transformed and experimentally manipulated cells. The resulting network includes 129,000 interactions and is therefore difficult to analyze on a global scale. To validate their approach the authors focused their attention on a sub network centered around oncogene *MYC*. This network includes 2,063 genes, 56 of which were directly connected to *MYC*. Among the genes that are directly interacting, about half are already known *MYC* targets. They tested 12 of the remaining genes using chromatin immunoprecipitation (ChIP), a technique that allows one to experimentally determine where *MYC* is binding on the genome. They discovered that 11 of these were also *MYC* targets. Therefore, the approach seems to reliably predict which genes are *MYC* targets although its coverage of known *MYC* targets remains sparse.

Integrated pathway analysis of expression data and transcription factor binding data

The techniques for pathway analysis of expression data discussed so far have utilized pre-existing pathway information to interpret the data or have attempted to reconstruct networks from expression data. A third class of techniques is now emerging that integrates multiple genome-scale data types. In particular, several approaches have been published recently that combine

expression data with transcription factor binding data.

ChIP allows one to experimentally determine where transcription factors are binding on the genome. This sort of data has been systematically collected for most of the known transcription factors in *Saccharomyces cerevisiae* (Harbison *et al.*, 2004) and has also been collected less systematically in many other organisms. Over the past few years, a number of techniques have emerged that attempt to interpret expression data in terms of transcription factor binding data and vice versa.

For example, one of the questions that can be addressed by integrating expression and binding data is in which phases of the cell cycle a transcription factor is active. Alter *et al.* (2004) provided an answer to this question by describing transcription factor binding data in terms of expression data. This technique is inspired by the expression deconvolution technique, in which expression data are represented as a linear combination of basis states (Lu *et al.*, 2003). In Alter's work, the basis states are the components of cell cycle data obtained using singular value decomposition and correspond to the different phases of the cell cycle: G1, S, G2, and M. This analysis allows her to demonstrate that cell cycle transcription factors such as SWI4 and SWI6 are active in the G1 phase whereas the origin replication complex components (such as ORC1) are active in the S phase. In Fig. 3.4, we show that the analysis may be reversed and expression data may be interpreted in terms of transcription factor binding basis states to gauge the activity of each transcription factor within a specific experiment.

Luscombe *et al.* performed a more global analysis of transcription factor activity in yeast (Luscombe *et al.*, 2004). They set out to characterize the transcriptional network across multiple conditions: cell cycle, sporulation, diauxic shift, DNA damage, and stress response. In each condition they reconstructed a network by identifying transcription factors that were expressed and genes that were differentially expressed and created a link between the two when the binding data suggested the factor bound the gene. They then performed extensive statistical analyses on these networks to identify changes in their properties.

These analyses lead to the classification of the experiments into two broad groups: endogenous processes (cell cycle and sporulation) and exogenous states (diauxic shift, DNA damage, and stress response). The former are complex multistage processes. These have low out degrees (the number of target genes for a given transcription factor), large average pathlengths (the number of links connecting two proteins), and high clustering (the level of transcription factor inter-regulation). In contrast, exogenous states produce rapid, large-scale responses and this is best accomplished with high out degree, small pathlength, and low clustering. In exogenous states, a few transcription factors drive a large number of genes without much "cross-talk."

Conclusions

We have discussed a variety of recent techniques that have been developed to analyze expression data from a pathway perspective. These techniques either leverage existing pathway information or attempt to deduce pathways from the expression data themselves. We have also illustrated how complementary data, such as transcription factor binding, may be used to enhance our understanding of the expression data.

A common perception among biologists is that the interpretation of expression data is one of the primary bottlenecks

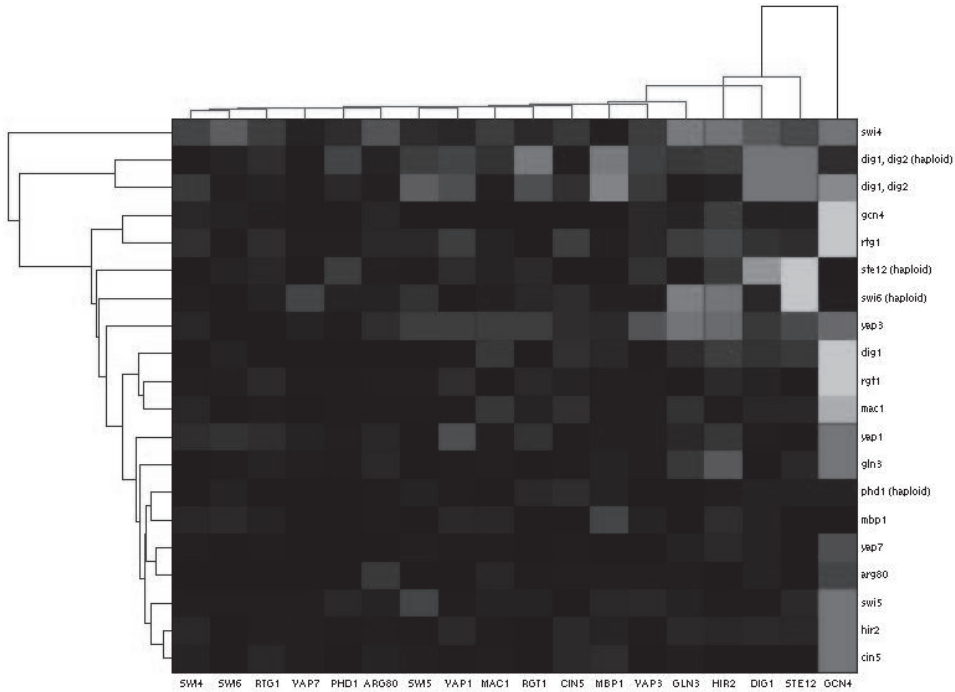

Figure 3.4 A heat map of the activity of transcription factors in yeast deletion experiments. In this example, basis states are defined as the genes bound by a specific transcription factor (the columns). Red indicates that the genes bound by the transcription factor are over expressed in the deletion experiments (the rows), while green indicates that the genes are under expressed. This example indicates that expression deconvolution may be used as a proxy for measuring the activity of transcription factors.

in the path to scientific discoveries. The pathway analyses we have described are attempts to remove, or at least ameliorate, this bottleneck. They allow scientists to look underneath expression data and interpret what biological phenomena are driving the observed expression patterns. As these techniques mature and become more accessible to the average biologist, expression profiling should become an even more powerful tool than it is already.

Finally, it is important to note that pathway analysis approaches are evolving in parallel with genomic data collection techniques. The availability of new data allows scientists to understand expression data in a deeper manner as we saw in the case of the integration of expression and binding profiles. Since we are merely at the beginning of the technological development of these new profiling techniques, it is reasonable to assume that over the next few years a variety of new pathway analyses approaches will be developed that utilize new types of data. As this occurs, our goal of using expression profiling to transparently interpret the inner workings of the cell will become more of a reality.

Future developments and trends

The work we have described above provides a static picture of expression data. That it, it allows one to assess which pathways and processes are active in a specific experiment, but not how these change

with time. In the future one may imagine more sophisticated models of expression data that provide such a dynamical view of expression. The advantage of such descriptions would be that one could generate predictions of how the expression of genes would change if experimental conditions are altered.

Detailed dynamical models of biological systems to date have described only small systems that include a few dozen genes and have therefore not been useful in interpreting expression arrays in a general fashion. One exception is the work of Holter *et al*. that generates a simple dynamical model of time series expression data (Holter *et al*., 2001). Here the authors attempt to derive a model that predicts the state of the system at time t based on its state at an earlier time point:

$$Y(t + \Delta t) = M \cdot y(t) \qquad (3.2)$$

where $Y(t)$ are the expression levels of all genes in an array at time t and M is a time translation matrix. However, it is not usually possible to solve this equation since the number of time points is typically much smaller than the number of genes. The authors therefore resort to modeling only the primary modes of the time series that are apparent from singular value decomposition. When they apply this approach to model yeast cell cycle data they demonstrate that the first two modes do a very good job at reproducing the system and that a particular 2×2 time translation matrix M reliably captures the behavior of the system.

Although these results seem promising, this model does not allow one to reliably predict how the system will change in response to different experimental conditions (e.g., mutations or environmental stresses) and therefore the model is still primarily descriptive. Nonetheless, it suggests that future approaches that possibly build upon these types of approaches may in the next few years bring us closer to the realization of truly predictive models. We might then find ourselves in a situation where expression arrays are used to verify models.

References

Al-Shahrour, F., Diaz-Uriarte, R., and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics *20*, 578–580.

Alter, O., and Golub, G.H. (2004) Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. Proc. Natl. Acad. Sci. USA *101*, 16577–16582.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B-cells. Nat. Genet. *37*, 382–390.

Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C., and Conklin, B.R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. *4*, R7.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601–620.

Friedman, N. (2003). Probabilistic models for identifying regulation networks. Bioinformatics I, Suppl. 2: II57.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T.,

Lander, E.S., Gifford, D.K., Fraenkel, E., and Young, R.A. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99–104.

Holter, N.S., Maritan, A., Cieplak, M., Fedoroff, N.V., and Banavar, J.R. (2001). Dynamic modeling of gene expression data. Proc. Natl. Acad. Sci. USA *98*, 1693–168.

Hosack, D.A., Dennis, G. Jr, Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol. *4*, R70.

Ihmels, J., Levy, R., Barkai, N. (2004) Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. Nat. Biotechnol. *22*, 86–92.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M.(2004). The KEGG resources for deciphering the genome. Nucleic Acids Res. *32*, D277–D280.

Kayo, T., Allison, D.B., Weindruch, R., and Prolla, T.A. (2001). Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. Proc. Natl. Acad. Sci. USA 98, 5093–5098.

Lamb, J., Ramaswamy, S., Ford, H.L., Contreras, B., Martinez, R.V., Kittrell, F.S., Zahnow, C.A., Patterson, N., Golub, T.R., and Ewen, M.E. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. Cell *114*, 323–334.

Li, K.C. (2002) Genome-wide coexpression dynamics: theory and application. Proc. Natl. Acad. Sci. USA *99*, 16875–16880.

Li, K.C., Liu, C.T., Sun, W., Yuan, S., and Yu, T. (2004). A system for enhancing genome-wide coexpression dynamics study. Proc. Natl. Acad. Sci. USA *101*, 15561–15566.

Lu, P., Nakorchevskiy, A., and Marcotte, E.M. (2003). Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. Proc. Natl. Acad. Sci. USA 100, 10370–10375.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature. 431, 308–212.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., and Groop, L.C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

Pavlidis, P., Qin, J., Arango, V., Mann, J.J., Sibille, E. (2004). Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. Neurochem. Res. *29*, 1213–1222.

Segal, E., Friedman, N., Koller, D., Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nat. Genet. *36*, 1090–1098.

Smid, M., Dorssers, L.C. (2004) GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. Bioinformatics 20, 2618–25.

Yang, H.H., Hu, Y., Buetow, K.H., Lee, M.P. (2004). A computational approach to measuring coherence of gene expression in pathways. Genomics *84*, 211–217.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C., Weinstein, J.N. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 4, R28.

Zhong, S., Storch, K.F., Lipan, O., Kao, M.C., Weitz, C.J., and Wong, W.H. (2004). GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontology. Appl. Bioinformatics *3*, 261–264.

Zhou, X.J., Kao, M.C., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O.M., Finch, C.E., Morgan, T.E., and Wong, W.H. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. Nat. Biotechnol. *23*, 238–243.