Network Inference from Multiple Data Sources

Introduction

The 21st Century is often said to be the century for biology, just as the previous century has been said to be the century for physics. The main driving force behind the rise of biology is the development of new tools and technologies that allow new knowledge to be uncovered in a much faster and high throughput manner. The discovery of restriction enzymes, which resulted in the 1978 Nobel Prize being awarded to the discoverers, led to the development of recombinant DNA technology. The demand for low cost high throughput sequencing led many laboratories around the world to search for alternatives to the chemical method published by Maxam and Gilbert (1977) and the dideoxy chain termination method published by Sanger *et al.* (1977). One of the most successful high throughput sequencing techniques today is pyrosequencing, which is based on the "sequencing by synthesis" principle (Ronaghi *et al.*, 1996; Ronaghi *et al.*, 1998; Ronaghi, 2001). The whole genome of Nobel laureate James Watson (available online at http://jimwatsonsequence.cshl.edu) has been obtained in 2007 by 454 sequencing, which is a massively-parallel pyrosequencing system currently capable of sequencing roughly 100 megabases of DNA per 7-hour run. DNA microarrays, which first emerged in the 1990s, allow researchers to monitor the expression levels of thousands of genes simultaneously. The combination of rapid, high resolution chromatography systems with fast-scanning mass spectrometers is one of the main technologies that has propelled the field of proteomics forward. Automated microscopy and established image analysis pipelines are also beginning to provide large amounts of spatial information. Last but not least, the generation of large databases and the development of novel computational algorithms have played important roles in helping biologists to understand life.

Types of biological data that can be used

There are many different types of biological data, most of which are publicly and freely available, that can be used to infer networks or functional linkages. The first type of data is sequenced genomes. To date, the genomes of humans, all the major model organisms (for example mouse, *Drosophila melanogaster*, and *C. elegans*), some species closely related to the model organisms, more than 300 microbes, and many others have been sequenced. With such massive amount of data available, we should try to mine as much information as we can out of them. Firstly, we can ask if a group of genes is always coinherited in the same set of bacteria. Secondly, we can ask if a group of genes coevolves. In other words, rather than representing whether the presence of protein A is correlated with the presence of protein B, the coevolution metric assumes that A and B are coinherited and instead represents whether the sequence evolution of A's relatives is correlated with that of B's relatives (Srinivasan *et al.*, unpublished). Thirdly, we can perform phylogenetic profiling. Essentially, we make a multiple sequence alignment of a group of proteins that share some reasonable amount of homology and then calculate the distance matrix, whose entries are measures of the evolutionary distances between each pair of sequences in the multiple alignment. Various methods, namely UPGMA (unweighted pair-group method using arithmetic averages), NJ (neighbor joining), MP (maximum parsimony), and ML (maximum likelihood), can be used to build phylogenetic trees and we can obtain different trees depending on the method used. Fourthly, we can ask if a group

of genes are colocated on the chromosome. In bacteria, proteins of closely related function are often transcribed from a single functional unit known as an operon. Fifthly, we can carry out the Rosetta Stone method, which identifies gene fusion events based solely on sequence comparison (Enright *et al.*, 1999). Some protein pairs with similar functions fuse different domains into one single protein in other species. Proteins that carry out consecutive metabolic steps or are components of the same molecular complexes often end up being expressed as a single polypeptide chain to maximize kinetic or expression efficiency. In such an event, two non-homologous proteins will align large proportions of their sequences to different parts of a third protein, which is referred to as the Rosetta Stone protein (Bowers *et al.*, 2004).

Expression profiling using either spotted DNA microarrays or lithography-printed chips is now a common technique to examine global transcriptional changes under various conditions. Thousands of microarray datasets for experiments done in myriad organisms are publicly available for analysis and they can be downloaded from databases such as KEGG's expression database (http://www.genome.jp/kegg/expression/), the Stanford Microarray Database (Demeter *et al.*, 2007), EBI's ArrayExpress (Parkinson *et al.*, 2007), and NCBI's GEO (Barrett *et al.*, 2007). In particular, yeast is an extremely popular model organism for high throughput studies and its cell cycle and responses to environmental changes have been extensively investigated using microarrays (Spellman et al., 1998; Gasch et al., 2000; Causton et al., 2001). The yeast datasets can be downloaded from http://gasch.genetics.wisc.edu/datasets.html.

Protein-DNA interactions are traditionally of great interest because a few master regulators or key transcription factors can control the expression of many genes to perform important cellular functions such as cell division. Microarrays have been used to determine on a genome scale which parts of the DNA a protein will bind to. This technique, known as ChIP (chromatin immunoprecipitation)-chip, has been used to localize protein binding sites in humans (Boyer *et al.*, 2005; Odom *et al.*, 2006), zebrafish (Wardle et al., 2006), and *Drosophila* (Zeitlinger *et al.*, 2007) among other organisms. The method can also be employed on different cell types, like embryonic stem cells (Boyer *et al.*, 2005) and hepatocytes (Odom *et al.*, 2006). In recent years, due to the advent of high throughput sequencing, a modified method, known as ChIP-seq, has emerged. In essence, high throughput sequencing is performed after the reverse cross-linking step of ChIP instead of hybridizing on a microarray. Since ChIP-seq is fairly new, there are still not that many available experimental data generated by this technique, although it is likely to gain widespread acceptance in the near future.

Intrinsic to the study of protein-DNA interactions is the identification of transcription factor binding sites or the identification of sequence motifs. The general strategy is to look for over-represented short stretches of nucleotides in the promoter regions of a set of genes (McGrath *et al.*, 2007). Typically, a biologist will perform a series of microarray experiments and then look for clusters of genes whose expression profiles are similar (for example, a set of genes whose expression is strongly upregulated during heavy metal stresses or a set of genes whose expression profile shows the same cell cycle-regulated pattern). He or she will then feed the promoter regions of these genes into a motif-finding program, such as BioProspector (Liu *et al.*, 2001), MEME (Bailey and Elkan, 1994), CisModule (Zhou and Wong, 2004), or AlignACE (Roth *et al.*, 1998; Hughes *et al.*, 2000). There is also a subscription-based database, called

TRANSFAC, available that contains data on transcription factors, their experimentally-proven binding sites, and regulated genes.

Even though many biological networks have been built using gene expression data, biological systems ultimately need to be explained in terms of the activity, regulation and modification of proteins. The ubiquitous occurrence of post-transcriptional regulation makes mRNA an imperfect proxy for such information. To obtain global protein levels, Ghaemmaghami *et al.* (2003) have constructed a yeast library where each gene carries a tandem affinity purification (TAP) tag at its native locus. Using a single specific antibody against the tag, they could monitor global protein levels under any desired growth condition.

Another useful type of genome-scale information that can be used for network inference is protein-protein interaction data. Biological functions such as DNA replication or cell division are typically executed by multi-protein complexes. Physical interactions are typically detected by yeast two-hybrid screens (or some variants like yeast three-hybrid screens or bacterial two-hybrid screens) or co-affinity purification experiments followed by high throughput mass spectrometry. Recently, 2 papers that describe similar attempts to find protein complexes in budding yeast have been published. Gavin *et al.* (2006) and Krogan *et al.* (2006) performed genome-wide screens for protein complexes using TAP as well as matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS) and liquid chromatography tandem mass spectrometry (LC-MS). Both papers reported finding about 500 complexes in yeast.

Besides physical interactions, we can also use information about genetic interactions, such as synthetic lethality, synthetic growth defect, dosage lethality, and dosage growth defect, to infer functional linkages. In particular, there is often functional redundancy in the cell for critical pathways, so one has to attempt to make double knockouts in order to detect synthetic lethality. Since deletion strains that cover the whole genome are available for the yeast *Saccharomyces cerevisiae*, a large scale synthetic lethal screen is possible and has been carried out by Tong *et al.* (2001).

There are searchable online databases that contain physical interaction and genetic interaction data for public use. The Biological General Repository for Interaction Datasets (BioGRID) database (http://www.thebiogrid.org) houses and distributes collections of protein and genetic interactions from major model organism. BioGRID currently contains over 198 000 interactions from six different species, as derived from both high-throughput studies and conventional focused studies (for example, Reguly *et al.*, 2006). Other useful databases include the Database of Interacting Proteins (DIP) and the Molecular Interaction Database (MINT).

Gene expression patterns and protein localization images are another important data source for infering networks because 2 genes that are expressed in the same tissue at the same developmental stage or 2 proteins that localize to the same subcellular compartment suggest that the 2 genes or proteins may be functionally linked. Furthermore, with automated microscopy that can take multi-field images over time without human intervention becoming more affordable, such data will likely grow in abundance in the near future. Some research groups have already undertaken large scale attempts to determine gene expression patterns and where most of the

proteins are localized within the cell. Huh *et al.* (2003) built a collection of yeast strains expressing full-length, fluorescently-tagged fusion proteins and they were able to classify these proteins, representing 75% of the yeast proteome, into 22 distinct subcellular localization categories. Haudry *et al.* (2008) built a central public repository called 4D*Xpress* to house all the gene expression patterns obtained by whole mount *in-situ* hybridization for the model organisms zebrafish, medaka, *Drosophila*, and mouse.

Besides those mentioned above, there are still many other types of useful data available for the purpose of network inference. One of these is categorical annotations, like Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and Clusters of Orthologous Groups or (COG). Like other data types, we can use the annotations in 2 ways. The first is to use the annotations as part of the process in building the network. The second is to use the annotations as verification that the built network may potentially be correct. Another type of useful data is protein structures. Proteins that share structural motifs may have the same or similar cellular function and there are databases like the Protein Data Bank (PDB), which contain almost all know protein structures. Last but not least, Belle *et al.* (2006) have determined the half-lives of two-thirds of the yeast proteome. Using a collection of 4200 TAP-tagged strains, they monitored the abundance of each TAP-tagged protein by Western blot analysis as a function of time following inhibition of protein synthesis by cycloheximide.

The need for multiple data sources

The need to use multiple data sources is best illustrated by an example. Beer and Tavazoie (2004) attempted to predict gene expression from sequence. The overall objective is to learn how a given gene will be expressed, under certain experimental conditions, given its 5' upstream DNA sequence. Given a set of microarray data, the authors first used a modified version of K-means clustering to partition the genes into groups that are coexpressed under a set of experimental conditions. Next, they used AlignACE to search for 12bp motifs up to 800bp upstream of each gene. They augmented the 615 putative motifs with 51 known and experimentally documented transcription factor binding sites. They then built a one-layer Bayesian network that mapped these sequence elements together with their properties to the expression patterns. With this network in place, they could theoretically take the sequence of any promoter region, look for the appropriate combination of motif features, and then predict the expression pattern for the corresponding gene under the conditions that the network has been trained.

Lam *et al.* (2008) provide evidence that the method developed by Beer and Tavazoie (2004) will not work. They study the transcription factor Pho4 in yeast, which activate the expression of 2 genes, *pho5* and *pho84*, by binding directly to their promoters. Virtually identical number and type of functional Pho4 binding sites are present in the *pho5* and *pho84* promoters, but yet, the 2 genes have different expression kinetics and thresholds. At intermediate levels of phosphate, the expression of *pho84* is high but the expression of *pho5* is low. Hence, the Bayesian network developed by Beer and Tavazoie (2004) will fail. The reason for the differential expression is that even though both promoters contain the same mixture of high and low affinity binding sites, the positions of the sites relative to nucleosome positions are different. The high affinity sites in the *pho84* promoter are in nucleosome-free regions, unlike those in the

pho5 promoter. Hence, affinity of accessible binding sites determines quantitative expression behavior. In other words, we need information about chromatin structure as well as sequence in order to understand gene expression. The required information on epigenetics covers not just nucleosomal positions, but it also includes patterns of post-translational modifications on histones (Kurdistani *et al.*, 2004), since certain histone marks are well-correlated with gene activation, while others are associated with gene repression.

There are two general reasons why we should use multiple data sources to infer networks and functional linkages, if possible. The first is that the coverage and reliability of a single data source are inherently limited as one data type illuminates only limited aspects of the underlining biological mechanisms (Linghu *et al.*, 2008). This explains why the method by Beer and Tavazoie (2004) will fail to predict the expression of *pho5* and *pho84*. The second reason is that there are many functional linkages that are not very obvious if one is to look at a single data source. Instead, such linkages will only be revealed by moderate support from multiple evidences (Srinivasan *et al.*, unpublished). There are also some false linkages that appear to be strong in one data type (which might perhaps be very noisy) but will fall apart when other data are included.

Ways of integrating the data

In the last decade, multiple heterogeneous data sources with different noise level and different coverage have often been utilized and integrated using primarily machine learning procedures, such as Bayesian methods, neural network, and decision tree. The reported integration results support the intuitive expectation that such integrated functional linkage networks can be more reliable than networks based only on a single data source (Linghu *et al.*, 2008). Here, I will describe some of the ways various data sources are used to infer networks.

Segal *et al.* (2003) use two types of data, namely gene expression data and categorical annotations. In the paper, they describe the use of a method named module networks to construct different regulatory programs for various clusters of genes. Module networks are fundamentally Bayesian networks, where dependency structure between the observed variables (genes in this case) is learnt. However, module networks rely on the assumption that many variables have similar behavior and therefore partition the variables into modules, so that the variables in each module share the same parents in the network and the same conditional probability distribution, i.e. similar variables are modeled by and constrained to the same parameters. The clustering is useful especially when the number of variables is large, as in the case of gene expression data.

For each cluster of genes, a multi-layer Bayesian network is constructed. The features used rely on the sequence annotations and are a pre-determined set of 466 candidate regulators comprising transcription factors and signal transduction molecules. Each feature or regulator is allowed to have three discretized states, namely up-regulated, no change, or down-regulated.

The learning procedure is iterative and is based on the Expectation Maximization (EM) algorithm. After taking as input a gene expression data set and a set of regulators and clustering of the data as initialization, the procedure alternates between two steps. In the M-step, it searches through the space of all possible models to find the most likely model, i.e. the model that

maximizes the likelihood (Bayesian) score, for each intermediate cluster. In the E-step, the procedure finds the model (regulatory program) that best explains the expression vector of each gene and then re-assigns the gene, if necessary, to the cluster having that model. Hence, the size and members of each cluster (module) may change with each iteration due to the re-assignments. By repeatedly considering combinations of possible regulator-target relationships, the procedure learns the best (locally optimal) network that can explain the underlying data.

There are several shortcomings in the method developed by Segal et al. (2003). The procedure is constrained such that it does not assign a regulator gene to a module in which it is also a regulatory input. This is done as a gene can predict its own expression relatively easily. However, there cannot be auto-regulatory feedback loops in any of the learned models, which is physiologically unrealistic as many transcription factors are known to regulate their own expression.

A key assumption of the paper is that the transcription level of regulators serves as a good proxy for their actual activity level. In some cases, this assumption is true, but unfortunately, many transcription factors are regulated by post-transcriptional mechanisms. These mechanisms include translational control (for example by miRNAs), nuclear import and export, phosphorylation, proteolytic degradation, or binding with small ligands. Ideally, we need to correlate the nuclear concentration of a regulatory protein in its active state with a regulated gene's mRNA transcript abundance, but this information is difficult to obtain in general.

The method presented can fail to pick up regulator-target relationships for a few other reasons. Firstly, regulation may occur only in some specific conditions where microarray data are not yet available. Secondly, there may be redundant pathways. If several regulators regulate the same target through parallel pathways, then the method identifies only one representative of the group. Thirdly, some regulatory relationships may occur between a regulator and only a few genes and thus cannot be generalized to an entire module.

The module networks approach can be improved in one important way. As implemented by Segal *et al.* (2003), it currently uses only annotations (list of regulators) as its features. There is so much available biological information that has not been utilized. Instead, it should use richer features that include more data sources. For example, the authors can include information about sequence motifs. Besides the presence or absence of certain combinations of motifs, their features can include position of a motif from translational start (ATG), motif orientation, as well as order and spacing between particular motifs. The authors can also include information about the chromatin in the features, for example post-translational modifications of histones that indicate an active or a repressed gene state.

Nguyen and D'haeseleer (2006) use 3 data types, namely gene expression data, categorical annotations (list of regulators), and sequence motifs. In their paper, the authors presented a deterministic mathematical strategy based on matrix algebra, called motif expression decomposition (MED) method, for explaining expression data under various conditions at the gene-by-gene level. Specifically, the method aims to minimize the following objective function

$$\|E - MA\|, \quad (1)$$

where E is a m genes by n conditions expression matrix (the experimental data), M is a m genes by k motifs matrix of condition-independent motif strengths (effect of each motif on gene expression), and A is a k regulators by n conditions matrix of condition-dependent regulator activities. Essentially, we have a matrix decomposition problem, where we want to break down the observed experimental data into 2 separate matrices M and A.

The general scheme is as follows. The matrix M is first initialized. The entry Mij is constrained to zero if motif jth does not exist in the promoter of gene ith. The non-zero elements of M are set either as the weighted sum of the number of motif instances if motifs are represented in a position-specific weight matrix (PWM) (motifs that are closer to the consensus sequence are weighted more heavily) or by simply counting the number of times a motif occurs in a promoter if no PWM is available. Then, the algorithm iterates between finding the matrices A and M by trying to minimize the objective function. So, given E and M, the algorithm uses least squares to find the matrix A. Then given E and A, the algorithm minimizes over M, and so on.

Nguyen and D'haeseleer (2006) note the importance of motif constraints or specific promoter context: motif geometry (location and orientation), motif exact sequence (i.e. its similarity to the consensus sequence), multiplicity, and co-occurrence with other motifs. By focusing primarily on motif location and orientation, the authors conclude that motifs do not always have the same level of influence on gene expression simply owing to their presence in the gene promoter. They also do not necessarily exert the largest influence on expression when they are near the start codon (ATG). The authors further conclude that there are four main types of motifs, namely short-range type (within 150bp from the start codon), mid-range type (150-300bp), long-range type (300-450bp), and orientation-dependent type.

There is one main area where the method can be improved. The authors should have extended their analysis to account for nonlinear motif-motif interactions. Such nonlinearities are commonplace in all organisms from simple bacteria promoters to complicated *Drosophila* enhancers. For example, many transcription factors bind as dimers or different transcription factors interact with each other to enhance their binding to the promoter, resulting in cooperativity. Naturally, the linear objective function (1) will have to be modified.

The previous 2 papers (Segal *et al.*, 2003; Nguyen and D'haeseleer, 2006) describe transcriptional networks. A certain set of transcription factors and signaling molecules (given by annotations) regulate a group of genes whose expression data under various conditions are known. The latter paper also includes motif information and tries to explain the actions of the regulators on the genes through the motifs of variable strengths.

Lee *et al.* (2004) sought to construct an extensive gene network by considering functional, rather than physical, associations, realizing that each experiment, whether genetic, biochemical, or computational, adds evidence linking pairs of genes, with associated error rates and degree of coverage. The types of data they use include gene expression data, protein-protein interaction data from DIP, protein-protein interaction data from co-immunoprecipitaion experiments followed by mass spectrometry, protein-protein interaction data from yeast two-hybrid assays, genetic interaction data from synthetic lethal screens, phylogenetic profiling,

Rosetta Stone method, literature mining data, and categorical annotations. In their framework, gene-gene linkages are probabilistic summaries representing functional coupling between genes.

The authors develop a unified scoring scheme for testing the many heterogeneous datasets, even when the datasets are accompanied by their own intrinsic scoring schemes. This re-scoring by a single criterion allows the authors to measure directly the relative merit of each dataset and then to integrate the datasets with weights that reflect the merit. Each experiment is evaluated for its ability to reconstruct known gene pathways and systems by measuring the likelihood that pairs of genes are functionally linked conditioned on the evidence, calculated as a log likelihood score (LLS):

$$LLS = \ln \frac{P(L|E)/\neg P(L|E)}{P(L)/\neg P(L)}, \quad (2)$$

where the ratio $P(L)/\neg P(L)$ is estimated by counting the number of gene pairs with any shared functional annotation and those without any shared functional annotation among all possible gene pairs chosen from the set of annotated yeast genes. The ratio $P(L|E)/\neg P(L|E)$ is estimated by counting the number of gene pairs that share or do not share functional annotation and that are also supported by the given evidence. The formula can therefore be interpreted as the log likelihood of the linkage conditioned on the given evidence and corrected for background expectations of linkages.

Various approaches for integrating information in order to more accurately define physical or functional interactions between proteins have been explored by various research groups. These approaches range from simple intersection or union to more sophisticated approaches, like the Bayesian method. However, the relative independence of the various datasets can be difficult to estimate in the Bayesian framework. Lee *et al.* (2004) have found empirically that a heuristic modification to the strict Bayesian approach performs well for integrating diverse functional linkage datasets. They rank order the log likelihood scores and then calculate the weighted sum (WS) scoring the functional linkage between a pair of genes as:

$$WS = \sum_{i=1}^{n} \frac{LLS_i}{D^{(i-1)}}, \quad (3)$$

where *LLS* represents the log likelihood score for the gene linkage from a single dataset, *D* is a free parameter roughly representing the relative degree of dependence between the various data sets, and *i* is the rank index in order of descending magnitude of the *n* log likelihood scores for the given gene pair. The free parameter *D* ranges from 1 to ∞ and is chosen to optimize overall performance on the functional benchmark. When $D = 1$, WS represents the simple sum of all log likelihood scores. *D* exhibits an optimal value of 1 in the case that all datasets are completely independent. As the optimal value of *D* increases, WS approaches the single maximum value of the set of log likelihood scores, indicating that the various datasets are strongly redundant (i.e. no new evidence is offered by the additional datasets over what is provided by the first set). Intermediate values of *D* represent exponentially diminishing belief in the additional evidence.

Conclusions

Today, with the sheer abundance of biological data, much of which are obtained through high throughput technologies like microarrays, the key issue is how to make sense of all these

data to gain novel biological insights. In the 1990s, when systems biology was still very much in its infancy, most biologists worked with just a single data source. For example, they would perform some microarray experiments to look at global expression levels in certain conditions and perform clustering analysis to find sets of genes that might be coregulated. Such an approach is still in use today and is actually sufficient in many instances. However, we can learn much more when we combine multiple data sources and try to infer networks and functional linkages.

Even though the last decade has seen many attempts to integrate heterogeneous data to predict gene function or build networks, there is still room for improvement. To the best of my knowledge, there are some datasets that have never been used as input to the training or determination process. One of them is protein structures. I believe the reason is that there are currently still too few solved crystal structures or solution structures available. Structural biology is still very much a time-consuming field and high resolution structures take time to be deciphered. Nevertheless, with large consortiums being given grants to solve the structures of well-conserved hypothetical proteins, the situation might change in the future as more structural information becomes publicly available. Another type of data that has never been used is images. The problem here is not quantity but instead lies in the need to convert an image into something that can be easily integrated into the training or determination process in a high throughput manner. We, as humans, can look at a set of images and easily say two proteins show the same cell cycle localization pattern in a bacterial cell or two genes are expressed in the notochord during E10.5 of mouse embryonic development. But to a computer, identifying such patterns in images is a difficult task. From identifying the boundary of an object (which may vary in size from image to image as the bacterial cell grows or the embryo develops) to demarcating the separate compartments within an object to making a decision whether there is fluorescence or staining in certain regions are all non-trivial. Hence, we need to improve or perhaps even develop new pattern classification algorithms in order to be able to utilize the numerous available images effectively.

Besides the need to better utilize all the types of data, there are two other areas that I feel can be improved. Firstly, all the current methods or procedures to infer gene function, predict functional linkages, or build networks are extremely time-consuming. Running WUBLAST to align the sequences of approximately 920000 microbial proteins versus each other (Srinivasan *et al.*, unpublished) or building module networks, an iterative process (Segal *et al.*, 2003), requires a lot of computational time and memory. The problem is going to get worse as more and more genomes are sequenced, more motifs are characterized (which will result in a much larger feature set for a Bayesian network), and a lot more of the other data types are being generated. Hence, there is an urgent need for computer scientists and mathematicians to device novel strategies or algorithms that might involve parallelization and task distribution to multiple processor units. Secondly, I feel that there is a lack of good interactive visualization tools that biologists can use to effectively examine constructed networks. The only one I know of is networks.stanford.edu, which is based on the work by Srinivasan *et al.* (unpublished). A biologist studying some particular system or cellular process should also have some say into what sort of data goes into building the network. Naturally, one does not always need to use every available dataset and the data sources to use is very much context dependent. For example, if we are investigating embryonic stem cells, there is in general no need to include ChIP-chip data from differentiated lymphocytes (unless we are looking at epigenetic memory during somatic cell nuclear transfer).

Hence, the ideal web interface, to me at least, is an input form asking me to choose the datasets, another form asking me to choose the algorithm or method to use to build the network, and the final result with interactive nodes and links so that I can zoom in and look at subgraphs if necessary.

References

1. Bailey T.L. and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*: 28-36.

2. Barrett T., D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. 2007. NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res.* **35**: D760-D765.

3. Beer and Tavazoie. 2004. Predicting gene expression from sequence. *Cell* **117**: 185-198.

4. Belle A., A. Tanay, L. Bitincka, R. Shamir, and E.K. O'Shea. 2006. Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci.* **103**: 13004-13009.

5. Bowers P.M., M. Pellegrini, M.J. Thompson, J. Fierro, T.O. Yeates, and D. Eisenberg. 2004. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**: R35.

6. Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., Melton, D.A., Gifford, D.K., Jaenisch, R. and Young, R.A. 2005. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. Cell 122: 947-956.

7. Causton H.C., B. Ren, S.S. Koh, C.T. Harbison, E. Kanin, E.G. Jennings, T.I. Lee, H.L. True, E.S. Lander, and R.A. Young. 2001. Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell.* **12**: 323-337.

8. Demeter, J., C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J.C. Matese, M. Nitzberg, F. Wymore, Z.K. Zachariah, *et al.* 2007. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* **35**: D766-D770.

9. Enright A.J., I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**: 86-90.

10. Gasch A.P., P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.* **11**: 4241-4257.

11. Gavin A.C., P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. Dümpelfeld, *et al.* 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**: 631-636.

12. Ghaemmaghami S., W.-K. Huh, K. Bower, R.W. Howson, A. Belle, N. Dephoure, E.K. O'Shea, and J.S. Weissman. 2003. Global analysis of protein expression in yeast. *Nature* **425**: 737-741.

13. Haudry Y., H. Berube, I. Letunic, P.D. Weeber, J. Gagneur, C. Girardot, M. Kapushesky, D. Arendt, P. Bork, A. Brazma, *et al.* 2008. 4DXpress: a database for cross-species expression pattern comparisons. *Nucleic Acids Res.* **36**: D847-D853.

14. 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.

15. Huh W.-K., J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, and E.K. O'Shea. 2003. Global analysis of protein localization in budding yeast. *Nature* **425**: 686-691.

16. Krogan N.J., G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, *et al.* 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**: 637-643.

17. Kurdistani S.K., S. Tavazoie, and M. Grunstein. 2004. Mapping global histone acetylation patterns to gene expression. *Cell* **117**: 721-733.

18. Lam F.H., D.J. Steger, and E.K. O'Shea. 2008. Chromatin decouples promoter threshold from dynamic range. *Nature* (in press).

19. Lee I., S.V. Date, A.T. Adai, and E.M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science* **306**: 1555-1558.

20. Linghu B., E.S. Snitkin, D.T. Holloway, A.M. Gustafson, Y. Xia, and C. Delisi. 2008. High-precision high-coverage functional inference from integrated data sources. *BMC Bioinformatics* **9**: 119

21. Liu X., D.L. Brutlag, and J.S. Liu. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*: 127-138.

22. Maxam A.M. and W. Gilbert. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* **74**: 560-564.

23. McGrath P.T. H. Lee, L. Zhang, A.A. Iniesta, A.K. Hottes, M.H. Tan, N.J. Hillson, P. Hu, L. Shapiro, and H.H. McAdams. 2007. High-throughput identification of transcription start sites, conserved promoter motifs, and predicted regulons. *Nat. Biotechnol.* **25**: 584-592.

24. Nguyen D.H. and P. D'haeseleer. 2006. Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Syst. Biol.* **2**: 2006.0012

25. Odom D.T., R.D. Dowell, E.S. Jacobsen, L. Nekludova, P.A. Rolfe, T.W. Danford, D.K. Gifford, E. Fraenkel, G.I. Bell, and R.A. Young. 2006. Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Syst. Biol.* **2**: 2006.0017.

26. Parkinson H., M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, *et al.* 2007. ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **35**: D747-D750.

27. Reguly T., A. Breitkreutz, L. Boucher, B.J. Breitkreutz, G.C. Hon, C.L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, *et al.* 2006. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5**: 11.

28. Ronaghi M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Res.* **11**: 3-11.

29. Ronaghi M., S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**: 84-89.

30. Ronaghi M., M. Uhlén, and P. Nyrén. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281**: 363-365.

31. Roth F.R., J.D. Hughes, P.E. Estep, and G.M. Church. 1998. Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**: 939-945.

32. Sanger F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:** 5463-5467.

33. Segal E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**: 166-176.

34. Spellman P.T., G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**: 3273-3297.

35. Srinivasan B.S., A.F. Novak, J.A. Flannick, S. Batzoglou, and H.H. McAdams. Network integration speeds protein discovery in 305 microbes (unpublished).

36. Wardle F.C., D.T. Odom, G.W. Bell, B. Yuan, T.W. Danford, E.L. Wiellette, E. Herbolsheimer, H.L. Sive, R.A. Young, and J.C. Smith. 2006. Zebrafish promoter microarrays identify actively transcribed embryonic genes. *Genome Biol.* **7**: R71.

37. Tong A.H., M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader, N. Pagé, M. Robinson, S. Raghibizadeh, C.W. Hogue, H. Bussey, et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**: 2364-2368.

38. Zeitlinger J., R.P. Zinzen, A. Stark, M. Kellis, H. Zhang, R.A. Young, and M. Levine. 2007. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes & Dev.* **21**: 385-390.

39. Zhou, Q. and W.H. Wong. 2004. CisModule: de novo discovery of *cis*-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci.* **101**: 12114-12119.