9 Predicting Protein Function and Networks on a Genomewide Scale

Edward M. Marcotte

WHAT IS PROTEIN FUNCTION?

Perhaps the most significant finding from the more than 80 genomes that have been sequenced as of 2002 has been the extent of our ignorance about the constituents of cells. In virtually every genome sequenced, the majority of genes have never been studied directly. In spite of this, for about half of the genes at least one near or distant relative has been studied, so we glean our knowledge from the activities of these relatives. Until recently such methods for extending information to proteins with similar sequences or structures (homology-based methods) have been the only form of inference about protein function.

Homology-based annotation, with algorithms such as BLAST (Altschul et al., 1997; `http://www.ncbi.nlm.nih.gov/BLAST`), has been wildly successful in extending knowledge from the small set of experimentally characterized proteins to the tens of thousands of proteins found in genome sequencing projects. However, these methods perform as one might expect: they provide information only for proteins with very closely related functions. They reveal little about proteins that work together but typically have unrelated sequences or structures. Thus, the homology-based methods cannot be used to reconstruct metabolic or signaling pathways or other protein interaction networks. That such a bias exists shows that there are different aspects to protein function; methods that reveal one aspect do not necessarily reveal others.

The two most important aspects of protein function, defined in figure 9.1, will be referred to as the *molecular function* and the *cellular function* of proteins. The homology-based methods tend to find only the molec-

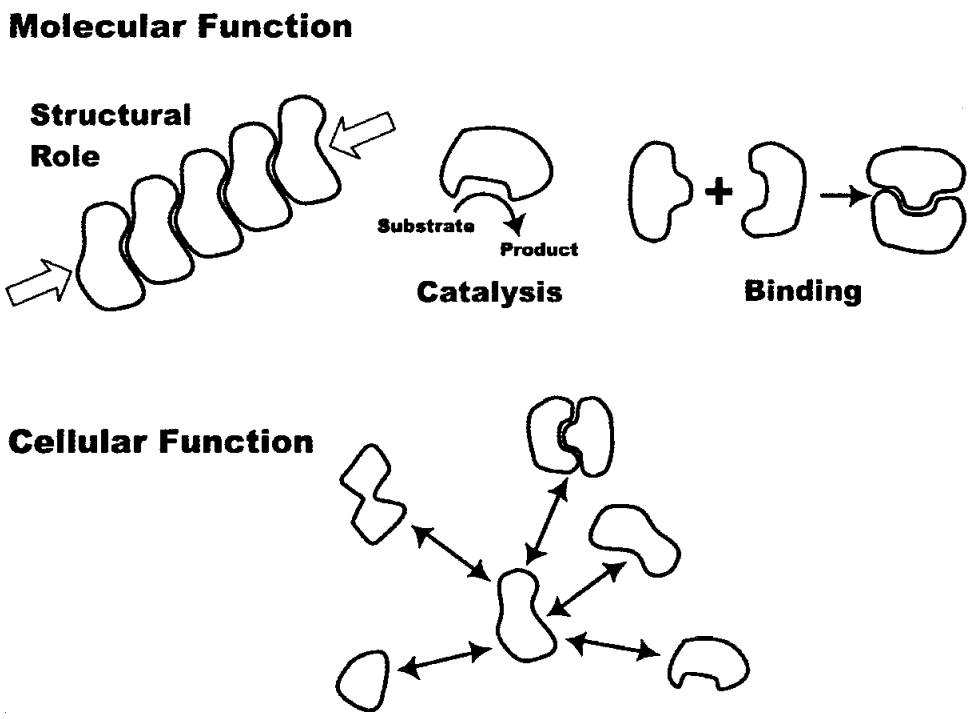


Figure 9.1 Two important components of protein function are the molecular (biochemical) function and the cellular (contextual) function (e.g., see Kim, 2000; Eisenberg et al., 2000). The molecular function of a protein is essentially the traditional view. It is the specific action that the protein engages in, such as binding, activation, inhibition, catalysis, fulfilling a structural role, etc. The cellular function is the system of interactions that the protein participates in, the context within which it operates. Other aspects of function include the intracellular location of proteins and the times and conditions under which proteins are expressed.

ular functions of proteins, but tell little about the context in which proteins operate. In fact, the context is crucial: proteins virtually never function alone in cells, but often interact with many partners. It has been estimated that an average protein will physically interact with 2–10 partners (Marcotte et al., 1999a). It can also be estimated that a protein will functionally interact—that is, participate in the same pathway—with even more proteins, perhaps two to three times the number of physical interactions. This interconnectedness is an important feature of the cellular organization and regulation of proteins. For this reason, protein networks are the subject of widespread study.

A new class of computational methods has been developed that finds the cellular function of proteins. This type of method is not based on comparisons of sequence or structure, but instead analyzes other

attributes associated with genes. Broadly speaking, these nonhomology methods draw inferences about relationships between genes by analyzing the context in which the genes are found. This chapter will present an overview of these methods, along with a discussion of their applications for finding protein function, reconstructing cellular pathways, revealing new metabolic systems, and even revealing physical properties of proteins, such as their locations in cells.

GENOMES CONTAIN CONSIDERABLE INFORMATION ABOUT PROTEIN FUNCTION

It is easy to think only of the coding potential of genes, since that seems most immediately important for producing a protein. However, genes have many different properties besides their coding potential, and information about the relationships between genes is often encoded in these other properties. Important contextual properties of genes include their position and order on the chromosome, the flanking control regions, the distribution of homologues in other species, the occurrence of fusions between genes, and so on. Table 9.1 summarizes many such genomic sources of functional data and lists data derived from measurements of protein and mRNA expression patterns.

Just as homology-based methods analyze conserved sequences or structures to find proteins with related molecular function, so nonhomology methods analyze conserved contextual properties to find proteins with related cellular function. At the heart of nonhomology methods is the fact that proteins working together in the cell have shared constraints—they must be encoded by the same genome, they often are coregulated, they occasionally are fused into a single gene, they must at some point be coexpressed, and so on. Nonhomology methods exploit these constraints to identify proteins working together.

DISCOVERING PROTEIN FUNCTION FROM GENOMIC DATA

Finding Function from Domain Fusions

One of the most straightforward nonhomology methods needs large numbers of protein sequences but does not require complete genomes. It has been known for years that proteins encoded as separate genes in one organism often are found in another organism fused into a single

Table 9.1 Analysis of "contextual" information associated with genes

	Contextual Information	Applications
Information in genomes about the relationships between genes		
Information derived from a single genome	Intergenic distance	Operon reconstruction
	Intragenomic conservation of regulatory sequences	Operon and regulon reconstruction
Information derived from comparisons of multiple genomes	Distribution of sequence homologues among different organisms	Calculation of phylogenetic profiles for pathway reconstruction and cellular localization
	Conservation of relative gene position	Operon reconstruction
	Domain fusions	Pathway reconstruction
	Intergenomic conservation of regulatory regions	Identification of coregulated genes
Information in expression data about the relationships between genes		
Clustering genes by their expression profiles	mRNA expression profiles	Identification of coregulated genes and pathway or operon reconstruction
	Spatial expression profiles	Pathway reconstruction
	Protein expression profiles	Pathway reconstruction
Clustering genes by the expression levels of all other genes in one or more experiments	Genomewide expression as a gene phenotype	Pathway reconstruction

Beyond simply coding for genes and their regulatory sequences, genomes are rich in information about the relationships between genes. Analysis of this information allows reconstruction of cellular systems, pathways, and genetic networks. For the last entry, the expression of all other genes is used as the phenotype when the gene in question is disrupted. Genes are then clustered to maximally match their phenotypes.

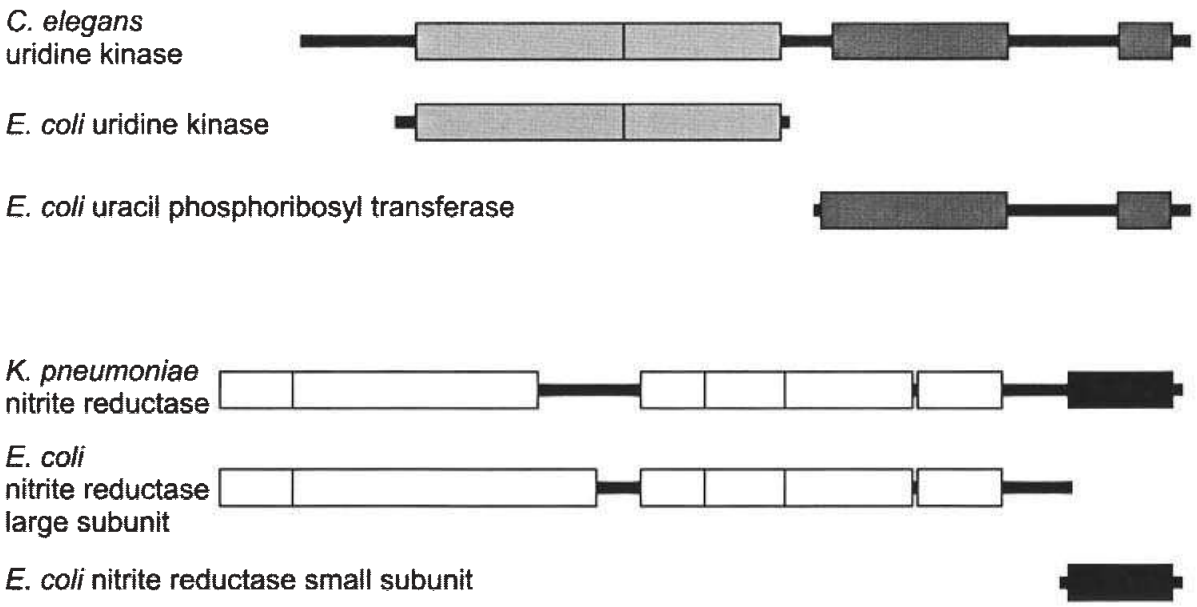


Figure 9.2 Two examples of the domain fusion or Rosetta Stone method of finding functional links. In each example the two lower proteins can be inferred to be functionally linked because of the existence of the top fusion protein. For example, if we did not already know that the *E. coli* nitrite reductase large and small subunits formed a heterocomplex, they could be inferred to be functionally linked after finding the *K. pneumoniae* fusion protein.

polypeptide. Two such examples are shown in figure 9.2. In each of the two examples, the separately encoded *E. coli* proteins are drawn beneath the fusion protein from another organism. In both cases, the *E. coli* proteins are members of the same pathway. In the bottom example, the nitrite reductase proteins physically interact to form an active nitrite reductase enzyme.

In fact, this trend is surprisingly common (Marcotte et al., 1999a; Enright et al., 1999), especially among metabolic proteins (Tsoka and Ouzounis, 2000). Thousands of such fusion events can be found—in yeast, more than 45,000 pairs of proteins can be found as fusion proteins in other organisms (Marcotte et al., 1999a). Almost universally, the cellular functions of the component proteins are very closely related. Searching systematically for these fusion events therefore rapidly generates functional links between proteins. For this reason, the fusion proteins have been called "Rosetta Stone" proteins for their ability to decode the functional links between component proteins (Marcotte et al., 1999a).

Rosetta Stone links are found by aligning a query protein's amino acid sequence against protein sequences from genomes or a large sequence database such as GenBank. The statistically significant hits from this search include sequence homologues and candidate Rosetta Stone proteins. These hits are then used as the query proteins for a second set of searches against the sequence database. The statistically significant hits from this second round of searches are then tested for similarity to the original query protein. Those second-round hits without sequence similarity to the original query protein are proteins with Rosetta Stone links to the original query protein.

Not all fusions convey the same degree of confidence in the resulting functional linkage. Individual domains have different propensities to participate in these gene fusion events, and many cell signaling domains, such as SH3 or tyrosine kinase domains, can be found fused into literally hundreds of different genes. These *promiscuous domains* still can be used to generate functional linkages, but it has been found that limiting the Rosetta Stone analysis to nonpromiscuous domains increases the functional similarity of the linked proteins. This filtering step can be performed either by explicitly forbidding links generated by promiscuous domains (Marcotte et al., 1999a; `http://www.doe-mbi.ucla.edu`) or by requiring strong sequence homology or even orthology between the individual proteins and the Rosetta Stone protein (Enright et al., 1999; Enright and Ouzounis, 2000). Regardless, it is possible to generate thousands of significant links between pairs of proteins in a genome by this method.

Finding Function from Coinheritance

One important consequence of the genomic revolution is the finding that genomes have mosaic compositions, containing genes with widely varying phylogenetic origins. This trend is especially strong among prokaryotes due to processes such as horizontal gene transfer (Jain et al., 1999; Koonin and Galperin, 1997), but is true to a considerable extent in eukaryotes as well (Marcotte et al., 2000). These variable phylogenetic origins of genes are another aspect of gene context for use in these analyses.

This phylogenetic diversity can be explicitly described for each gene by calculating its *phylogenetic profile* (Pellegrini et al., 1999, with related

concepts in Gaasterland and Ragan, 1998; Huynen et al., 1998; Ouzounis and Kyrpides, 1996; and Tatusov et al., 2001). A phylogenetic profile describes the presence or absence of a gene across a set of organisms with sequenced genomes. Genes with similar phylogenetic profiles are therefore always inherited together or absent from the same organisms. This similarity is unlikely to happen by chance if enough species are examined, and such proteins are thus extremely likely to function together. For 30 genomes, there are about 2^{30}, or 10^9, possible phylogenetic profiles, making random matches of profiles unlikely. When enough different species are examined to be statistically significant, genes with similar phylogenetic profiles are inferred to be functionally linked.

Constructing a phylogenetic profile for a gene requires performing sequence alignments between that gene and all genes from each of the fully sequenced genomes. Because thousands of sequence alignments must be calculated, rapid alignment algorithms like BLAST (Altschul et al., 1997) are typically used. The phylogenetic profile of a gene is then calculated as a vector in which each entry represents a measure of sequence similarity between that gene and the most similar sequence match in a given genome. This measure of sequence similarity $S_{i,j}$ can be as simple as a binary code: $S_{i,j} = 1$ if a sequence homologue of gene i is present in genome j and $S_{i,j} = 0$ if no homologue exists. Alternatively, the measure of sequence similarity can be a real, valued measurement reflecting the degree of sequence similarity present. One such measure that has empirically been shown to work satisfactorily is $S_{i,j} = -1/\log(E)$, where E represents the expectation value from the sequence alignment between gene i and the top-scoring sequence match in genome j (Marcotte, 2000). Real-valued phylogenetic profiles calculated in this fashion are shown in figure 9.3.

Once a phylogenetic profile is calculated for each of the genes in a genome, functional links can then be inferred between genes with similar phylogenetic profiles. The simplest approach is to treat phylogenetic profiles as coordinate vectors positioning genes in a high-dimensional space, then calculating distances between genes, using such distance metrics as the Manhattan, Euclidean, or Mahalonobis distance. Genes positioned close together in space can be inferred to be coinherited, and therefore functionally linked. Another approach is to apply a statistical test such as a Fisher exact test on the binary phylogenetic profile vectors to identify coinherited genes.

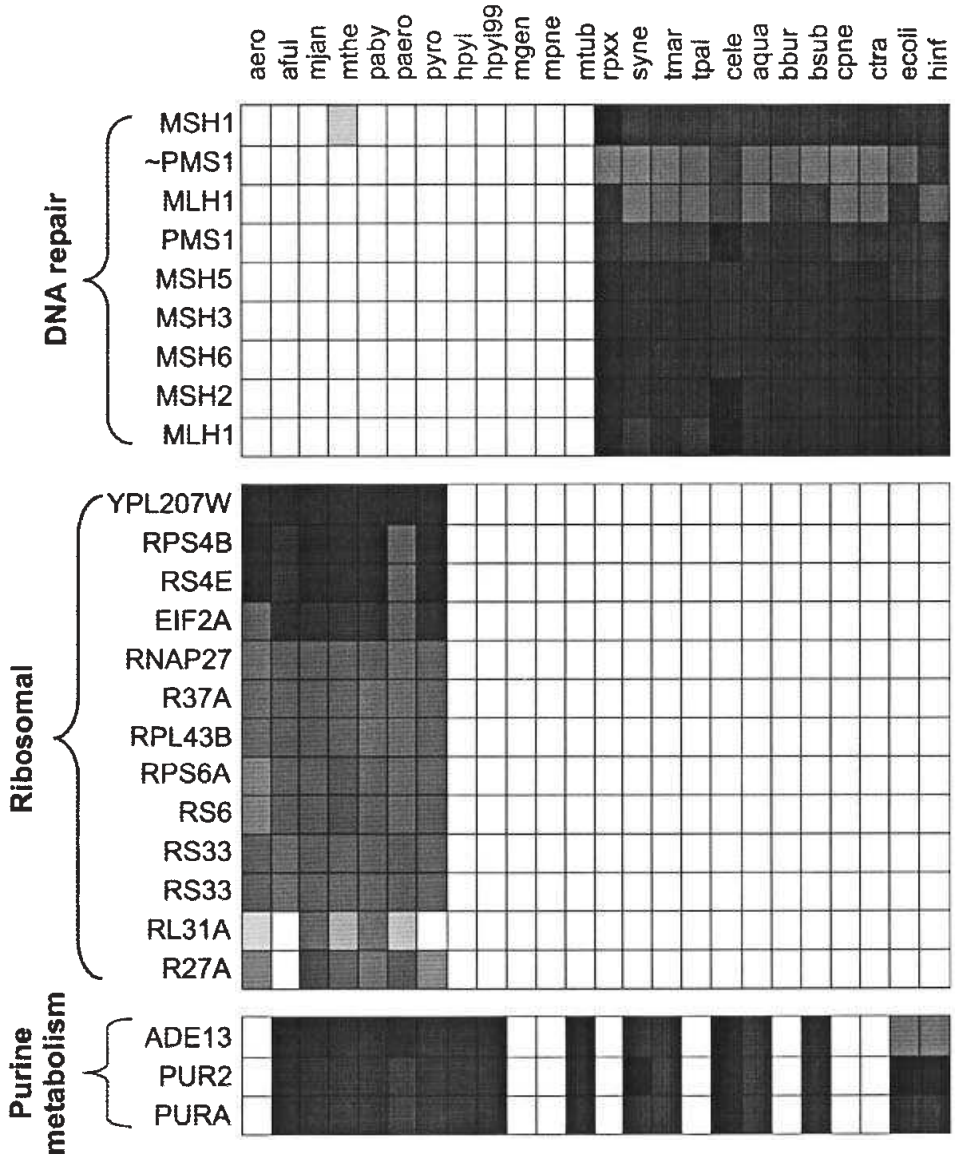


Figure 9.3 Examples of phylogenetic profiles for a number of yeast proteins. Each profile, drawn horizontally, indicates the degree of sequence similarity of a protein—for example, MLH1—to the most similar protein in each of the fully sequenced genomes (listed as abbreviations across the top.) Where there is no sequence homologue, the profile has a white square, and where there is a statistically significant sequence homologue, the square is colored to indicate the degree of homology, with black being most similar. Three functional classes of proteins are profiled; profiles are shared within a functional class but are quite distinct between classes.

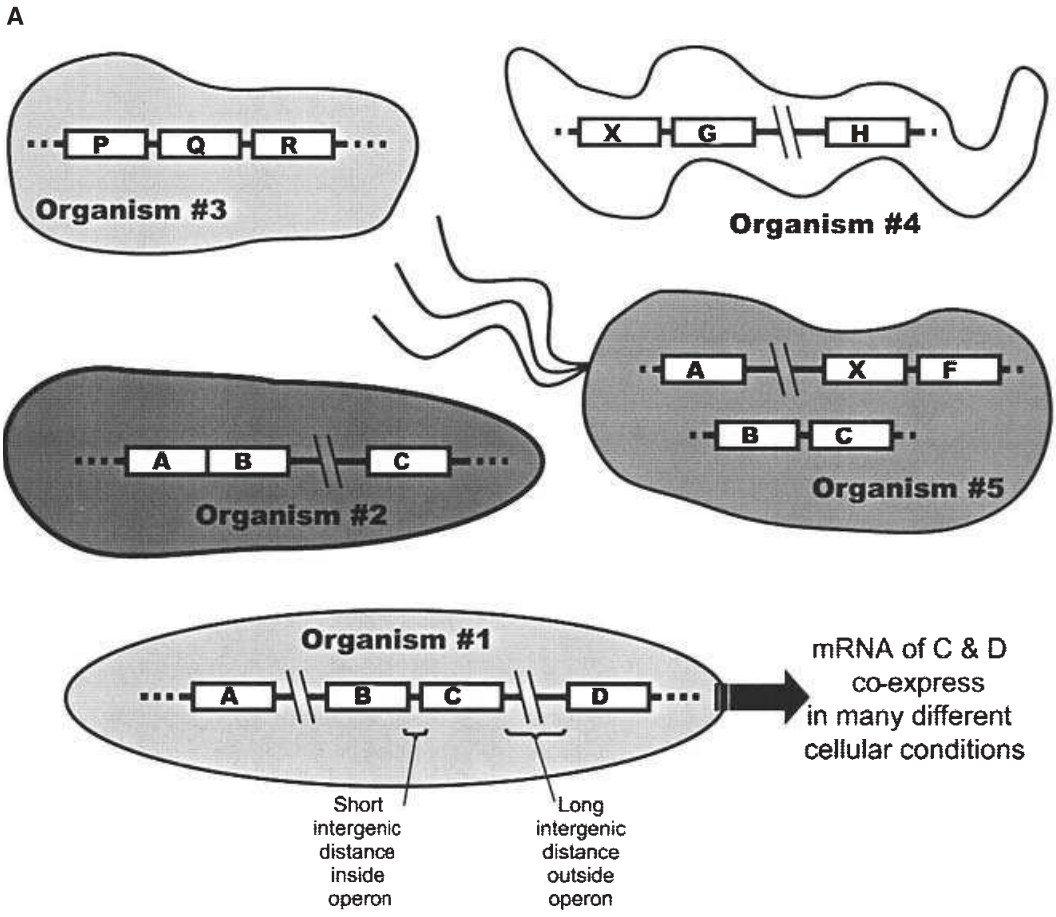
As an alternative to calculating pairwise links, the genes can simply be clustered into coinheritance groups on the basis of similarity between their phylogenetic profiles. Many such clustering approaches have been developed in computer science and statistics for clustering points in high-dimensional spaces, such as *k*-means clustering, which are appropriate for this task.

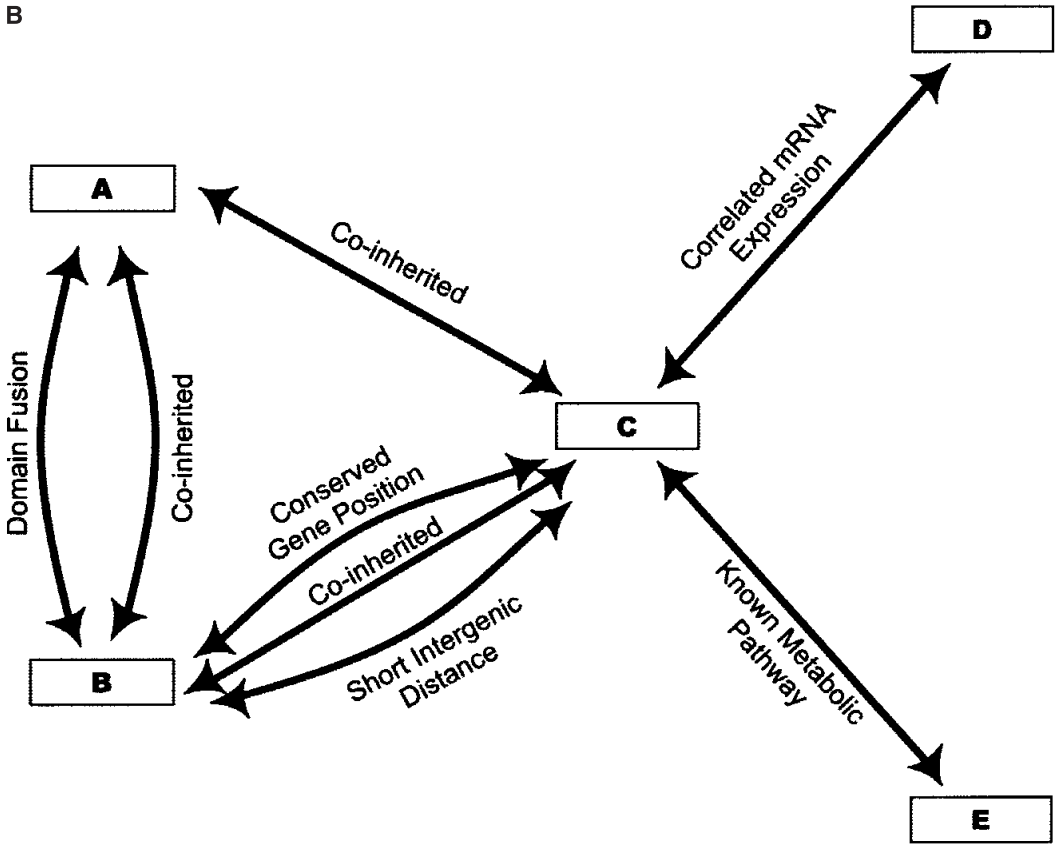
Finding Function from Relative Gene Position

Another powerful method for finding functionally linked genes comes from examining the conservation of relative positions of genes in genomes (Dandekar et al., 1998; Tamames et al., 1997; Overbeek et al., 1999). As with the two previous methods, this aspect of gene context can be analyzed in a straightforward fashion. The essence of the method is that the order of genes in genomes tends to randomize over time. Therefore, if two genes have similar positions relative to one another in several genomes, the genes are likely to be functionally linked. In the simplest case, this means that the genes are immediate neighbors in several genomes, but the method could theoretically be extended to any separation between the genes. On-line tools for investigating the genomic neighbors of a gene include the Entrez genome (`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome`) and WIT (`http://wit.mcs.anl.gov/WIT2`) databases.

This method exploits the trend for prokaryotic genes to be organized into operons, in which genes with a related function are clustered close together on the genome to allow coordinate transcription and translation of the genes. Operons seem to be uncommon in most eukaryotes, occurring mainly in unusual gene families such as the cadherins (Wu and Maniatis, 1999). However, in prokaryotes, operons are virtually the norm, and where genes from an operon are conserved in multiple species, this method allows very reliable functional links to be inferred. In fact, it has been shown that the observation of two genes as immediate neighbors in two reasonably unrelated organisms is sufficiently statistically significant to infer a functional link between the proteins (Overbeek et al., 1999), as diagrammed in figure 9.4A and B.

A rough calculation of significance goes as follows. Given two adjacent genes in a genome, we would expect by random chance to find the genes adjacent in a second genome of n genes, with all genes, but not gene order, conserved between the two genomes, only two times out of


A
P
Q
R
Organism #3
X
G
H
Organism #4
A
X
F
B
C
Organism #5
A
B
C
Organism #2
Organism #1
A
B
C
D
mRNA of C & D
co-express
in many different
cellular conditions
Short
intergenic
distance
inside
operon
Long
intergenic
distance
outside
operon


B
D
A
Co-inherited
Correlated mRNA
Expression
Domain Fusion
Co-inherited
C
Conserved
Gene Position
Co-inherited
Short Intergenic
Distance
Known Metabolic
Pathway
B
E

$n - 1$. So, given a typical bacterial genome of $n = 4000$ genes, we would expect to find the two genes adjacent with a random probability $p = 2/(3999)$, or 5×10^{-4}. Because of the ubiquity of operons and the low random likelihood of conserved neighbors, the coverage of this method can be quite high, and thousands of pairwise functional links can be generated (Huynen et al., 2000a).

Finding Function from Intergenic Distances

A second promising method has been described that analyzes gene position to find functional links between proteins. This method is explicitly formulated to detect operons and works by analyzing the number of nucleotides separating neighboring genes (Salgado et al., 2000). Genes organized in bacterial operons are cotranscribed on a single mRNA and translated in a coordinated fashion. This coordination of transcription and translation for the genes in an operon probably places a selective pressure on keeping genes close together, as compared to an absence of selection for adjacent genes not in the same operon. Thus, adjacent genes with short intergenic distances tend to be in the same operon; adjacent genes with long intergenic distances tend not to be.

One advantage of this method is that it can be performed for genes unique to a genome—no gene conservation is required for the method to operate. This ability to work on ORFans, the genes found only in a

Figure 9.4 (A) Comparing the genome of organism #1 with the genomes of several other organisms allows a number of inferences to be drawn about the relationships between the genes of organism #1. In the figure, genes, depicted as labeled white boxes, are arranged on the genomes, drawn as heavy black horizontal lines. First, the genes A and B can be found fused in organism #2, suggesting that A and B are functionally linked. Second, the genes A, B, and C are found in the same set of organisms (#1, 2, and 5) and are absent from the same set of organisms (#3 and 4). This coinheritance suggests A, B, and C are functionally linked. Third, genes B and C are neighbors in more than one genome, suggesting a selective pressure to maintain their relative positions. Likewise, the intergenic distance between genes B and C is much smaller than the typical intergenic distance, suggesting that B and C may belong to an operon. Fifth, the mRNA of genes C and D are coexpressed in many different experiments, suggesting C and D are coregulated or function together. Each of these inferences can be conceptualized as generating a functional linkage between two proteins. (B) The resultant network of functional links. Predicted networks can be compared can complemented by experimental networks, such as the experimentally derived link between C and E.

single genome (Fischer and Eisenberg, 1999), sets this method apart from the other genomic methods described above. In this respect, the analysis of intergenic distance has more in common with analysis of expression data, which can also provide functional links for ORFans.

Finding Function from Regulatory Regions

The last obvious contextual property of genes useful for assigning protein function is the presence of regulatory regions found outside of gene coding regions. These regulatory sites in DNA are recognized and bound by transcription factors, enhancers, and repressors to control the transcription of the neighboring genes. Because genes with related functions are often coregulated, it would seem reasonable to create functional links between genes with similar regulatory regions.

Unfortunately, regulatory regions are notoriously difficult to identify. Judging the similarity between them is equally difficult. Although consensus sequences have been identified for most major regulatory sites (e.g., see the Eukaryotic Promoter Database, `http://www.epd.isb-sib.ch`), the sites recognized by a given transcription factor or polymerase are often quite varied. Nonetheless, progress has been made in identifying shared regulatory regions upstream of coexpressed genes (Roth et al., 1998) and upstream of coinherited genes, Rosetta Stone linked genes, and genes coconserved in operons (Manson McGuire and Church, 2000). Since genes in these categories are often functionally related, it seems likely that the inverse process, clustering genes by their regulatory regions, will also yield functional information. An attempt at this process (Pavlidis et al., 2001) shows that such functional information is available, although the method is not currently as powerful as methods exploiting other sorts of contextual information. However, such analyses of regulatory regions are likely to improve dramatically with the explicit knowledge of transcription factor binding sites generated from DNA microarray mapping of transcription factor specificities (Iyer et al., 2001).

DISCOVERING PROTEIN FUNCTION FROM EXPRESSION DATA

The genomic analyses discussed above examine static genomes and draw inferences from the state of the genomes at one point in time. However, genomes and cells are dynamic systems, and considerable

information can be gleaned about cellular systems by analyzing these dynamics. We might argue that genomes are dynamic on two time scales: the evolutionary and the immediate. The methods discussed above analyze events on the evolutionary time scale. Now we turn to events on the more immediate time scale.

Finding Function from mRNA Expression Patterns

DNA microarrays and EST sequencing have produced literally millions of discrete measurements of gene expression. This flood of data has in turn stimulated many analyses of gene expression profiles. In general, the analyses share the following form: A set of measurements of the expression of a number of genes under different conditions is available, from DNA microarrays (e.g., as in Lashkari et al., 1997), serial analysis of gene expression (SAGE; Velculescu et al., 1995), or expressed sequence tags (EST; Adams et al., 1991). Expression vectors are then constructed for the genes, each vector describing the expression of a given gene under a range of cellular conditions, cell types, genetic backgrounds, and so on. These expression vectors are then clustered to find genes with similar expression patterns (Eisen et al., 1998). Given enough independent experiments (>100) with sufficient variation in the conditions, genes clustered in this fashion tend to be functionally related (Marcotte et al., 1999b). Fortunately, unlike complete genome sequences, data of this sort are readily generated. It is possible to perform large numbers of microarray experiments, producing enough expression data to find statistically significant functional links. Many expression data sets are publicly available from sites such as the Stanford Microarray Database (`http://genome-www4.stanford.edu/MicroArray/SMD`).

A variation of this approach involves analysis of SAGE or EST libraries collected from various tissues and cell conditions (e.g., the dbEST database: `http://www.ncbi.nlm.nih.gov/dbEST/index.html`). In this approach, mRNAs from cells are reverse transcribed into cDNAs and sequenced. Since many thousands of mRNAs are typically sequenced, the EST or SAGE library is a fairly representative selection of the mRNAs present under those cellular conditions, and thus can serve in a fashion analogous to microarray expression measurements. EST and SAGE libraries vary widely in size and completeness, so calculations of expression vectors with their data are not entirely

straightforward. However, related analysis can be performed, such as Guilt-by-Association, which essentially creates functional links between genes based on their copresence and coabsence from EST libraries (Walker et al., 1999).

Coexpression analyses have advantages and disadvantages in regard to genomic data for functional predictions. The primary disadvantage, beyond having to collect additional data, is that the functional inferences from coexpression are relatively weak until a large body of expression data is collected (Marcotte et al., 1999b). However, this is more than compensated for by the advantage of learning information about any gene for which expression can be detected, regardless of its conservation in other species. Coexpression analysis and the prediction of operons using intergenic distances (Salgado et al., 2000) are currently the only two computational methods capable of generating functional information for ORFans (Fischer and Eisenberg, 1999).

Finding Function from Spatial Expression Profiles

The expression methods discussed above typically give no information about the intracellular location of the expressed molecules. However, spatial expression data should be useful for pathway reconstruction, since we expect functionally linked proteins to be found at similar subcellular locations. Therefore, the converse will often be true: proteins that are always expressed at the same locations probably function together. This approach to finding protein function is quite technically demanding, but in spite of the difficulty, one group has collected such spatial expression data for more than 1750 genes expressed in *Xenopus* oocytes (Gawantka et al., 1998). More recently, the data have been incorporated into a database and methods to measure similarity between mRNA spatial expression patterns have been developed (Pollet et al., 2001). Although considerable work remains, this work establishes the viability of this method for generating functional information.

Finding Function from Protein Expression Profiles

Gathering expression data for an entire proteome, or all of the proteins encoded by a genome, is only now becoming feasible, due largely to the development of high-throughput mass spectrometric analyses of pro-

teins (Shevchenko et al., 1996; Hunt et al., 1986; Gygi et al., 1999; Jensen et al., 2000). Although such protein expression data are not yet widely available, they will be a valuable complement to the mRNA expression data from EST libraries and chips. What is not yet clear is how well protein expression data will correlate with mRNA expression data.

Early results comparing protein expression by mass spectrometry and mRNA expression by SAGE suggested that mRNA and protein expression patterns are quite different (Gygi et al., 1999). Recent developments with DNA chips have allowed quantification of mRNAs being actively translated through analysis of polysomal mRNA fractions. These chip-based measurements of protein expression show strong correlation with chip-based measurements of mRNA expression (Joe DeRisi, personal communication). Nonetheless, it is likely that the protein expression patterns will hold considerable value for inferring protein function.

In theory, protein expression data can be analyzed similarly to mRNA expression data. It is likely that expression data will be collected for many of the proteins in a proteome over many different cellular conditions. As with mRNA expression data, these protein expression data will compose expression vectors that can be clustered and analyzed much as the mRNA data are.

However, protein expression data may contain an additional element absent from mRNA expression data: mass-spectrometric methods have the capability not only to measure protein expression levels but also to identify protein modifications. Posttranslational modifications of proteins are widespread in cells, both spontaneous unregulated events such as oxidation, and enzymatic modifications such as lipidation, phosphorylation, and ADP ribosylation. Such modifications often modify the activity or localization of the proteins. Thus, it seems likely that protein expression profiles will catalog not only expression patterns but also protein states, such as on, off, activated, repressed, and so on. These protein state vectors will provide a rich source of data for protein function prediction.

MEASURING PROTEIN FUNCTION AND TESTING PREDICTIONS

Before testing any of these predictive methods, one must develop a *metric* for measuring protein function. At first glance, protein function

would seem difficult to quantify. However, several metrics have been developed that perform quite well, allowing optimization and calibration of the methods.

Perhaps the most obvious metric is that of testing that the methods recover known functional relationships. Using a database of known pathways, such as the KEGG (Kanehisa and Goto, 2000; `http://www.genome.ad.jp/kegg/kegg2.html`) or EcoCyc (Karp et al., 2000; `http://ecocyc.pangeasystems.com/ecocyc`) database of metabolic pathways, or the DIP database of protein interactions (Xenarios et al., 2001; `http://dip.doe-mbi.ucla.edu`), each method is evaluated by its coverage, the fraction of experimental links correctly predicted by the algorithm, and by its accuracy, the fraction of predicted links that are verified by an experimental link.

Unfortunately, the measurement of accuracy cannot be very exact, since our knowledge of experimental pathways is limited and few pathways are known completely. Thus, absence of a link from the experimental database does not necessarily mean the link is wrong. Due to this limited knowledge, we can measure false negative predictions accurately (failure to predict an experimental link), but cannot evaluate false positive predictions (prediction of a functional link where none exists). To some extent, the accuracy measurement, while not correct in an absolute sense, can be treated as a relative value for optimization and for comparisons between algorithms.

A second metric that performs well in practice is that of *key word* recovery or category matching (Marcotte et al., 1999b). For this approach, genes of known function are first classified into a limited set of functional categories. Many databases have such categorizations incorporated, sometimes explicitly (as in the MIPS database of yeast proteins; Mewes et al., 1998) and sometimes implicitly (as in the key words associated with proteins in the SWISSPROT protein sequence database; Bairoch and Apweiler, 2000). Testing predictions is then reduced to checking for agreement between the predicted and known key words or categories for each characterized protein, and finding the average agreement over all characterized proteins. An example is calculating

$$\langle \text{key word recovery} \rangle = \frac{1}{A} \sum_{i=1}^{A} \sum_{j=1}^{x} \frac{n_j}{N},$$

where x is the number of key words known for the protein i being

tested, N is the number of key words predicted for the protein, and n_j is the number of times key word j from the protein's known annotation appears in the predicted key word list. The average key word recovery is calculated for all A characterized proteins.

For many of the predictive methods, the prediction is not of a given functional category but of a link between two proteins. In these cases, for all predicted protein pairs involving proteins of known function, the overlap between the key words or categories of the two proteins can be calculated with a function such as the Jaccard coefficient:

$$\langle \text{key word overlap} \rangle = \frac{1}{P} \sum_{i=1}^{P} \left(\frac{k_1 \cap k_2}{k_1 \cup k_2} \right),$$

where for each of the P pairs of linked, characterized proteins, the k_1 key words of one protein are compared against the k_2 key words of the linked protein partner. The number of key words in the intersection is divided by the number of key words in the union to give a normalized measure of the overlap between the two sets of key words. The value of this overlap averaged over all P pairs gives a measure of the accuracy of the prediction algorithm. To optimize and compare prediction algorithms, this measurement of method accuracy can be combined with the measured coverage of known pathways.

ASSIGNING PROTEINS TO FUNCTIONAL CATEGORIES

One simple way to implement these methods is to test if proteins belong to given functional categories (Pavlidis et al., 2001; Marcotte et al., 2000). To do this, an algorithm is trained to recognize the characteristics of proteins in a given functional category. Such a discrimination algorithm requires a set of quantitative features for each protein. Effectively, these features are treated as coordinates mapping the protein into a high-dimensional feature space. When the features are chosen appropriately, proteins belonging to a given functional category fall in a distinct region of this feature space and proteins from other functional categories fall in other regions.

Many of the contextual properties of genes can be interpreted as features. For example, the phylogenetic profile of a protein is a vector in which each element describes the degree of similarity of the protein to the most similar sequence in a given genome. When interpreted as a list

of features, the phylogenetic profile describes the mapping of a protein into a *phylogenetic space*. The attributes of this space are the following: It is an n-dimensional space, where n is the number of genomes used to calculate the phylogenetic profile. The axes of the space are not orthogonal—some genomes are quite similar to each other, so some axes are more correlated than others. (If we choose, we can orthogonalize the space—for example, by applying a whitening transformation.) Last, proteins are not evenly distributed in this space. Certain systematic biases occur in the types of proteins encoded by a genome, and these in turn introduce biases in the genes' locations in phylogenetic space. For example, each genome contains a fraction of genes unique to that species; these genes all map to the same region of phylogenetic space. Likewise, certain genes are broadly conserved among only eukaryotes or prokaryotes—again, these genes all map to the same general region of phylogenetic space. However, proteins with a related function cluster in this space, as do eukaryotic proteins localized to similar cellular compartments (Marcotte et al., 2000).

A discrimination algorithm defines a set of boundaries in this high-dimensional space that separate proteins with the desired function from all other proteins. Numerous algorithms have been adapted from statistics and computer science for this purpose, including Bayesian classifiers (elegantly described in Mosteller and Wallace, 1984), support vector machines (Pavlidis et al., 2001), neural network discriminators, and linear discrimination functions (Marcotte et al., 2000). The advantage of this method of predicting function is that one can test for very specific functions, as well as calculate the degree of confidence in the results.

INTEGRATING METHODS TO DISCOVER PROTEIN FUNCTIONAL AND INTERACTION NETWORKS

The discrimination algorithms described above work under the assumption that a set of functionally related proteins is known, and more proteins with the same function are desired. In this approach, the algorithms must be trained on a set of positive examples, proteins whose functions are known to match the desired function, as well as on a set of negative examples.

However, a naive approach can be useful to discover what trends are in the data and to look for naturally occurring clusters. The naive

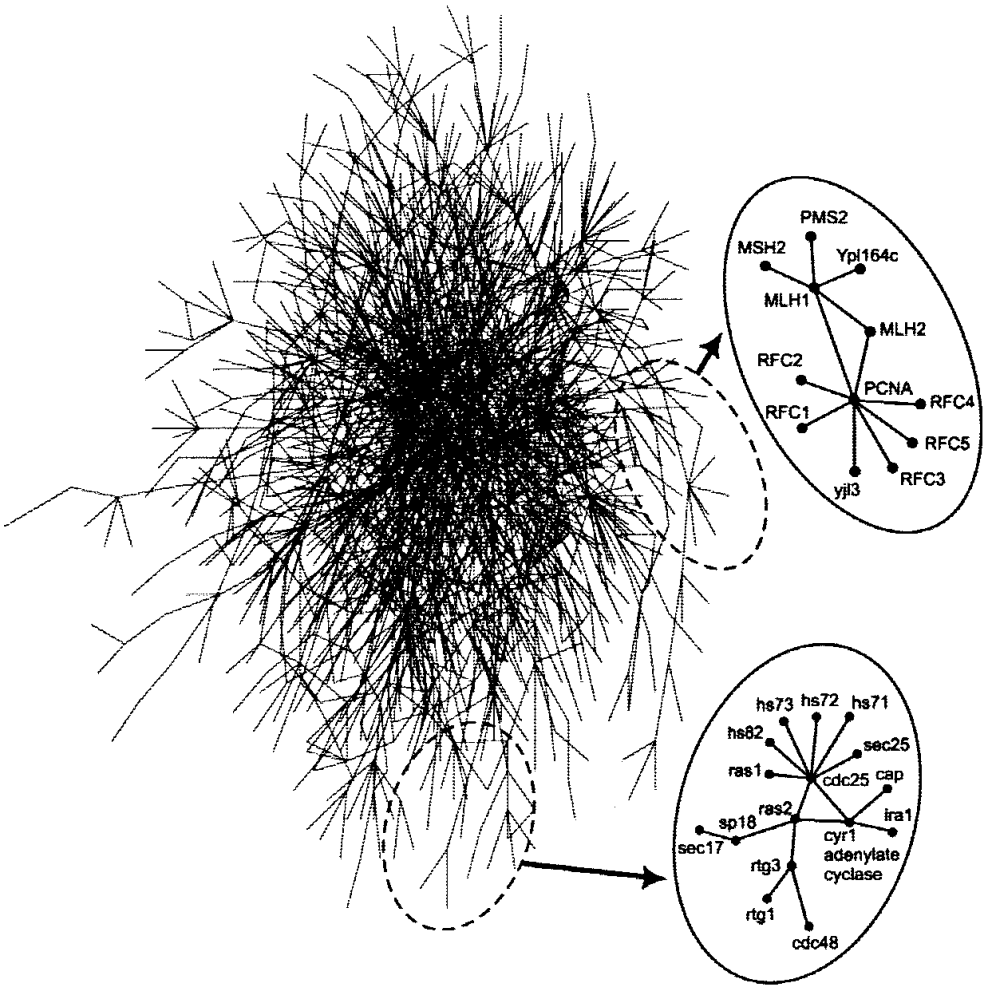


Figure 9.5 The proteins of yeast interact in an extensive network. Here, the vertices of this graph are 1722 yeast proteins participating in 2612 experimentally observed interactions, drawn as edges connecting the interacting partners. Two regions are expanded to show an interaction network involving the ras protein and an interaction network involving several DNA replication factors (RFC1–5). Many experimental techniques are represented, including high-throughput two-hybrid interaction screens (Uetz et al., 2000; Ito et al., 2000), mass spectrometry, and co-immunoprecipitation. The interactions are available courtesy of Ioannis Xenarios, curator of the Database of Interacting Proteins (Xenarios et al., 2001).

approach is also biologically motivated: it is now becoming apparent that proteins are organized into large interaction networks in the cell. One such experimentally derived protein interaction network is shown for the proteins of yeast in figure 9.5, derived from high-throughput measurements of protein interactions (Uetz et al., 2000; Ito et al., 2000) and from mining biological literature for all previously known yeast

protein interactions (Xenarios et al., 2001). Such networks reinforce the notion that proteins never work alone. Ideally, the predictive methods should reveal exactly these sorts of networks.

To discover such networks, the predictive methods can be applied to produce functional links between pairs of proteins. Although the links are generated in a pairwise fashion, extensive networks of proteins result when links are calculated for all of the genes in a genome. Different types of networks are calculated, depending on the method used. For example, mRNA expression links may produce coexpression networks, and phylogenetic profiles will produce coinheritance networks. However, networks provide a logical framework for combining methods. Using the metrics described earlier, each method can be optimized to link proteins with a comparable degree of confidence. Then, links generated by each method can be combined to create a functional interaction network. A simplified example is diagrammed in figure 9.4B, derived from the gene context information for one the organisms (#1) in figure 9.4A.

Actual networks calculated for all of the proteins encoded in a genome are much more complicated. Figure 9.6 shows such a predicted functional network for 2240 proteins of yeast. Inspection of the network shows that it has considerable diversity in its structure, with many highly connected subnetworks. Examination of such subnetworks shows reasonable correspondence to many known pathways (e.g., see Pellegrini et al., 1998; Marcotte, Pellegrini, Ng, et al., 1999; and Marcotte, Pellegrini, Thompson, et al., 1999). Uncharacterized proteins can therefore be assigned function by linking them with known pathways. This approach allowed preliminary assignment of functions to more than half the uncharacterized proteins of yeast (Marcotte, Pellegrini, Thompson, et al., 1999; `http://www.doe-mbi.ucla.edu/`) and to 10% of the genes of *M. genitalium* (Huynen et al., 2000b).

Analysis of these predictive networks and their correspondence to metabolic, signaling, and interaction networks is an ongoing area of study. Open topics of study include defining subnetworks, cliques, and network properties; determining which functional links correspond to physical interactions and which have other interpretations; and dynamic models of networks. Predictive networks can be incorporated into metabolic pathway models, such as those discussed in chapter 10 or those incorporated into the E-cell project, described in chapter 11.

Figure 9.6 A network of predicted functional links between yeast proteins. As in figure 9.5, proteins are drawn as vertices of the graph, and functional links are drawn as edges between functionally linked proteins. In all, 2240 proteins are shown participating in 12,012 functional links, as calculated from phylogenetic profiles (adapted from Marcotte, 2000).

For these models, predictive networks may be especially useful for completing input pathways known only partially from experiment.

DISCOVERING NEW METABOLIC SYSTEMS

An especially tantalizing aspect of the study of protein networks is the discovery of novel cellular systems. Molecular biology has until recently generated knowledge about proteins one at a time, each researcher studying the system of his or her desire. The overall effect has been a somewhat random patterning of knowledge over the proteome.

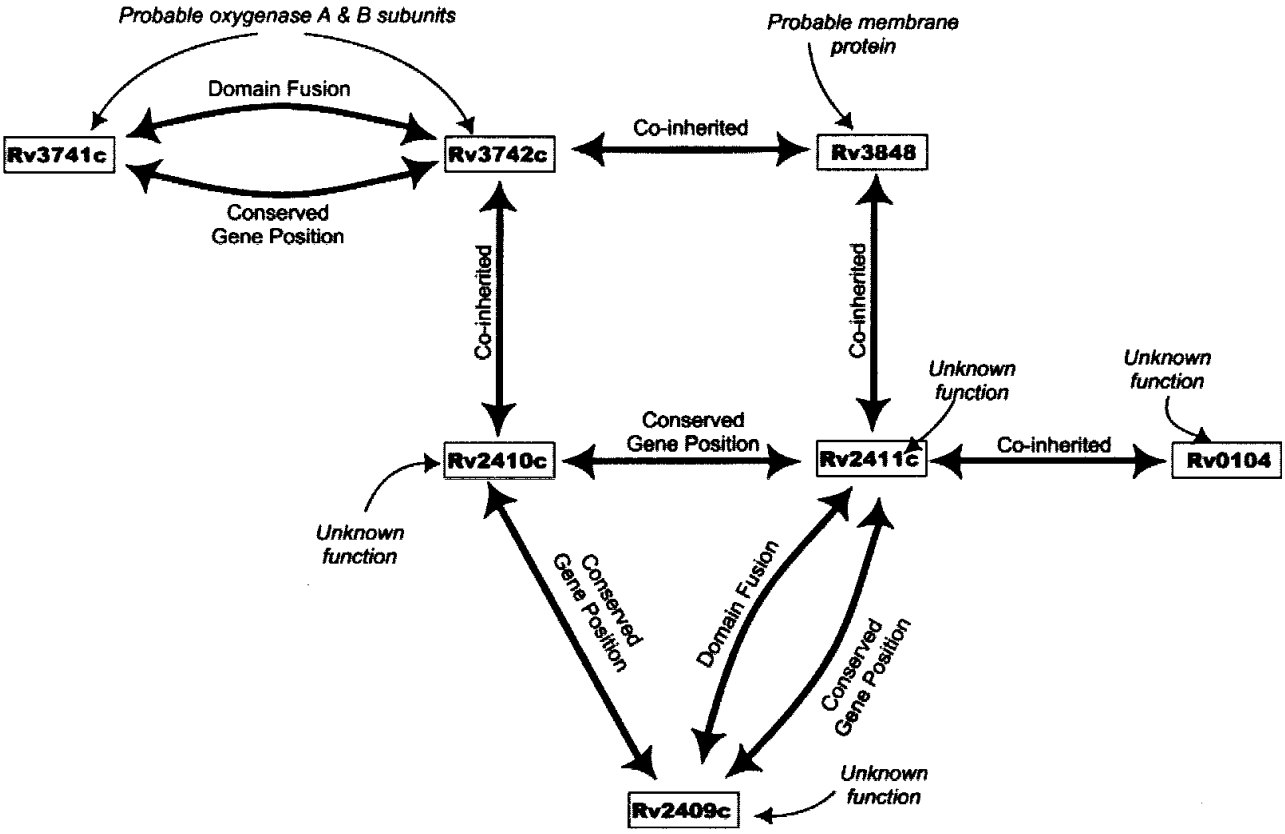


Figure 9.7 Novel pathways are revealed in computationally predicted networks. Shown here is a network of *M. tuberculosis* genes linked together by a combination of predictive methods. Multiple methods support each other in linking the genes, increasing confidence that the proteins participate in the same pathway. At the time of this writing, functions were unknown for all of the genes, with the exception of homology of Rv3741c and Rv3742c to oxygenase subunits. This homology suggests that the genes in the network may participate in a novel metabolic pathway in *M. tuberculosis*.

However, the functional and physical interaction networks give the best and most complete estimates of cellular pathways and systems. Examining the networks shows exactly which systems have been well studied and which have been neglected entirely. By searching for such unstudied systems, the network analysis allows systematic discovery of novel pathways.

One such novel system from *Mycobacterium tuberculosis* is diagrammed in figure 9.7. This system was found in a search for tightly functionally linked but unannotated genes. The genes are linked by a number of redundant functional linkages, increasing the likelihood that the genes really function together. Of the seven genes in this putative

pathway, none have a known function, although two are estimated from sequence homology to be subunits of an uncharacterized oxygenase. Here, the network analysis reveals only the cellular functions, but not the molecular functions, of the proteins. We can only speculate, based upon the oxygenase proteins, that this system is a novel metabolic pathway in *M. tuberculosis*. Defining all such new systems is the first step; what follows is perhaps the harder work of characterizing and understanding the new systems.

REFERENCES

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulous, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., et al. (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252: 1651–1656.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. 25: 3389–3402.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*. 28: 45–48.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: A fingerprint of genes that physically interact. *Trends Biochem. Sci*. 23(9): 324–328.

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.

Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* 405: 823–826.

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

Enright, A. J., and Ouzounis, C. A. (2000). GeneRAGE: A robust algorithm for sequence clustering and domain detection. *Bioinformatics* 16: 451–457.

Fischer, D., and Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics* 15: 759–762.

Gaasterland, T., and Ragan, M. A. (1998). Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics* 3: 177–192.

Gawantka, V., Pollet, N., Delius, H., Vingron, M., Pfister, R., Nitsch, R., Blumenstock, C., and Niehrs, C. (1998). Gene expression screening in *Xenopus* identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. *Mech. Dev*. 77: 95–141.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol*. 19: 1720–1730.

Hunt, D. F., Yates, J. R. III, Shabanowitz, J., Winston, S., and Hauer, C. R. (1986). Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* 83: 6233–6237.

Huynen, M., Dandekar, T., and Bork, P. (1998). Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett*. 426: 1–5.

Huynen, M., Snel, B., Lathe, W. III, and Bork, P. (2000a). Exploitation of gene context. *Curr. Opinion Struct. Biol*. 10: 366–370.

Huynen, M., Snel, B., Lathe, W. III, and Bork, P. (2000b). Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Res*. 10: 1204–1210.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA* 97: 1143–1147.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409: 533–538.

Jain, R., Rivera, M. C., and Lake, J. A. (1999). Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. USA* 96: 3801–3806.

Jensen, P. K., Pasa-Tolic, L., Peden, K. K., Martinovic, S., Lipton, M. S., Anderson, G. A., Tolic, N., Wong, K. K., and Smith, R. D. (2000). Mass spectrometric detection for capillary isoelectric focusing separations of complex protein mixtures. *Electrophoresis* 21: 1372–1380.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28: 27–30.

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucleic Acids Res*. 28: 56–59.

Kim, S. H. (2000). Structural genomics of microbes: An objective. *Curr. Opinion Struct. Biol*. 10: 380–383.

Koonin, E. V., and Galperin, M. Y. (1997). Prokaryotic genomes: The emerging paradigm of genome-based microbiology. *Curr. Opinion Genet. Dev*. 7: 757–763.

Lashkari, D. A., De Risi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* 94: 13057–13062.

Manson McGuire, A., and Church, G. M. (2000). Predicting regulons and their cis-regulatory motifs by comparative genomics. *Nucleic Acids Res*. 28: 4523–4530.

Marcotte, E. M. (2000). Computational genetics: Finding protein function by non-homology methods. *Curr. Opinion Struct. Biol*. 10: 359–365.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751–753.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature* 402: 83–86.

Marcotte, E. M., Xenarios, I., van der Bliek, A. M., and Eisenberg, D. (2000). Localizing proteins in the cell from the phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 97: 12115–12120.

Mewes, H. W., Hani, J., Pfeiffer, F., and Frishman, D. (1998). MIPS: A database for protein sequences and complete genomes. *Nucleic Acids Res*. 26: 33–37.

Mosteller, F., and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer-Verlag.

Ouzounis, C., and Kyrpides, N. (1996). The emergence of major cellular processes in evolution. *FEBS Lett*. 426: 1–5.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* 96: 2896–2901.

Pavlidis, P., Furey, T. S., Liberto, M., Haussler, D., and Grundy, W. N. (2001). Promoter region-based classification of genes. *Proc. Pac. Symp. Biocomput.* 6: 151–163.

Pavlidis, P., Weston, J., Cai, J., and Grundy, W. N. (2001). Gene functional classification from heterogeneous data. *Proceedings of the Fifth International Conference on Computational Molecular Biology*. pp. 242–248.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 96: 4285–4288.

Pollet, N., Schmidt, H. A., Gawantka, V., Niehrs, C., and Vingron, M. (2000). *In silico* analysis of gene expression patterns during early development of *Xenopus laevis*. *Proc. Pac. Symp. Biocomput.* 5: 443–454.

Roth, F. P., Hughes, J. D., Estep, P. W., and Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol*. 16: 939–945.

Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J. (2000). Operons in Escherichia coli: Genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA* 97: 6652–6657.

Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., and Mann, M. (1996). Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two-dimensional gels. *Proc. Natl. Acad. Sci. USA* 93: 14440–14445.

Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol*. 44: 66–73.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. 29: 22–28.

Tools for investigating the genomic neighborhood of a gene include the Entrez genome web site: `http://ww.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome` and the WIT database: `http://wit.mcs.anl.gov/WIT2`

The Eukaryotic Promotor database lists the consensus regulatory sequences derived from promotors of many eukaryotic genes: `http://www.epd.isb-sib.ch`

DNA microarray protocols, numerous experiments, and data are available from the Stanford Microarray Database:
`http://genome-www4.stanford.edu/MicroArray/SMD`

Additional measurements of gene expression are available in Expressed Sequence Tag databases such as dbEST: `http://www.ncbi.nlm.nih.gov/dbEST/index.html`

Many known metabolic pathways and networks have been characterized in the KEGG database: `http://www.genome.ad.jp/kegg/kegg2.html`

EcoCyc database: `http://ecocyc.pangeasystems.com/ecocyc`

Database of Interacting Proteins: `http://dip.doe-mbi.ucla.edu`

Last, several sequence databases also provide functional annotation of the genes that can be used in benchmarking programs that predict gene function. Among these annotated databases are

MIPS: `http://www.mips.biochem.mpg.de/`

Swiss-Prot: `http://ca.expasy.org/sprot/`

Gene Ontology Consortium: `http://www.geneontology.org`