In: Focus on Protein Research
Editor: John W. Robinson, pp. 165-189
ISBN: 1-59033-968-1

Chapter 8

COMPUTATIONAL METHODS FOR PROTEIN FUNCTION ANALYSIS

Matteo Pellegrini, Thomas G. Graeber
Department of Chemistry and Biochemistry, University of California

INTRODUCTION

Computational methods used to analyze protein function can be divided into three broad categories: sequence, expression and interaction based methods. Sequence based methods rely on the ability to construct alignments between protein sequences. These methods are by far the most developed in the field of bioinformatics, with the bulk of the development occurring over the past two decades. Recent innovations in protein sequence alignment methodology include indirect homologies, graph based analysis, Bayesian alignments and protein fusion analysis. One important consequence of new alignment techniques has been the cataloguing of protein domains.

Over the past few years protein sequence alignments methodologies have been extended to utilize genome sequences. Genome based methods exploit the information contained within the full sequence of an organism's genome. As this review is being written over one hundred complete genomes are available for analysis. The bulk of these come from bacteria and archaebacteria. The genomes of eukaryotes such as yeast, fruit fly, *C. elegans*, *Arabidopsis thaliana* and humans have also been completed. The alignment methods covered in this review that use genome sequence information are phylogenetic profile analysis, which searches for the absence or presence of gene families across organisms, and gene neighbor analysis, which searches for gene pairs whose proximity on the genome is preserved across species. Gene neighbor analysis allows one to partially reconstruct the components of operons within bacteria.

The second category of computational methods we will review utilizes the information from mRNA profiling experiments. Messenger RNA concentrations of each expressed gene within a cell may be measured on an array when a fluorescently labeled mRNA hybridizes to a particular spot on the array. Using this technology it is possible to study changes in gene

expression as cells are perturbed. One type of analysis of this data involves clustering it into sets of genes with similar expression levels across multiple experiments. These clustered genes often participate in the same biological process and so this technique may be used to infer the functions of the proteins for which the clustered genes code.

The last technique that we will review involves the study of the function of a protein through an analysis of its interacting partners. Protein interactions may be measured experimentally using a variety of techniques: yeast two hybrid, co-precipitation, and protein fragment complementation assays among others. To date thousands of protein interactions have been reported in the literature and catalogued in protein interaction databases. Using these databases it is possible to identify the interacting partners of a specific protein. This analysis may permit one to obtain a deeper understanding of protein complexes and the biological processes in which proteins are involved.

DEFINITION OF PROTEIN FUNCTION

In order to discuss computational approaches to assign protein function, it is first necessary to briefly review what is meant by protein function. Borrowing from the Gene Ontology Consortium (Gene Ontology), protein function may be understood at two levels: molecular function and biological process.

The molecular function of a protein involves the tasks performed by individual gene products; examples are transcription factor, DNA helicase and kinase. The biological process instead involves broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions.

It is important to distinguish between these two function levels because different methodologies shed light on one or the other level. As an example of this distinction let us consider kinases. Within the human genome there are about 500 kinase genes. Most of these genes code for proteins that phosphorylate other proteins on serine, threonine or tyrosine residues. However the substrates of these enzymes vary greatly and include other kinases, small molecules or other proteins such as histones.

Although all kinases share this common molecular function they are involved in very different biological processes. For instance mitogen activated protein kinases participate in signal transduction regulating cell division, glukokinases phosphorylate glucose and are part of the glycolytic pathway and histone kinases that phosphorylate histones are involved in transcriptional regulation and chromatin remodeling.

Typically, one can learn more about the molecular function of a protein by using protein alignment techniques. These approaches group together regions of proteins with similar sequences and similar molecular functions. For instance, a multiple alignment of all human protein kinases reveals the highly conserved core kinase domain, with small differences between serine/threonine kinases and tyrosine kinases. Additionally subsets of the kinase family may be aligned to elucidate similarity in other regions such as extracellular domains and receptor tyrosine kinases.

In contrast, to understand the biological process a protein is involved in it is necessary to understand how a protein fits within an interaction network and to look at the proteins it is linked to within the network. Methods that elucidate the biological process that a protein is

involved in include various non-homology alignment techniques discussed below, along with the clustering of genes based on their co-expression and the direct study of protein interaction networks.

ALIGNMENT METHODS

Protein Sequence Alignments

During the course of evolution protein sequences are subject to point mutations and insertions and deletions of sub-sequences. These mutations act to gradually transform the amino acid sequences of proteins in newly evolving species. Although the amino acid sequences are altered, it is still often possible to recognize which sequences evolved from a common ancestor if one has an appropriate evolutionary model. To determine which sequences are homologous, and are likely to have descended form a common ancestral protein sequence, one needs an amino acid substitution matrix and an alignment algorithm.

In many cases, amino acids with similar chemical properties can substitute for each other in a protein without significantly altering the protein's function, while other amino acids make poor replacements. A substitution matrix captures these preferences and models the likelihood of changing one amino acid into another during the course of evolution. Typically the entries of these matrices are the log odds of these substitutions. Over the years many different substitution matrices have been developed such as BLOSUM (Henikoff 1992), PAM (Schwartz 1978) and Gonnet (Gonnet 1994), each one measuring these probabilities using different sets of starting alignments that have been manually created.

Armed with a substitution matrix it is now possible to determine which positions in one protein sequences have remained unchanged or have mutated in another sequence. To determine the correspondence of the positions in two protein sequences one must find the correct alignment of the two sequences. Several methods that find the optimal alignment (that maximizes the log odds scores from the substitution matrix) between two protein amino acid sequences have been developed over the past few decades (e.g. Needleman 1970,Smith 1981). More recently Bayesian statistics have been applied to rigorously compute optimal alignments (Zhu 1998). However, in general these approaches are computationally intensive, and thus not always applicable to large scale homology searches.

The most commonly used alignment method to date is BLAST (Altschul 1999), which speeds up the optimal searches by limiting the space of all possible alignments to those that contain an exact small sequence match, without significantly compromising the results. This allows users to search for homologous sequences to a query protein in protein databases that contain millions of sequences. Using this method, searching a protein sequence against the full non-redundant protein sequence database requires less than a minute on a typical computer.

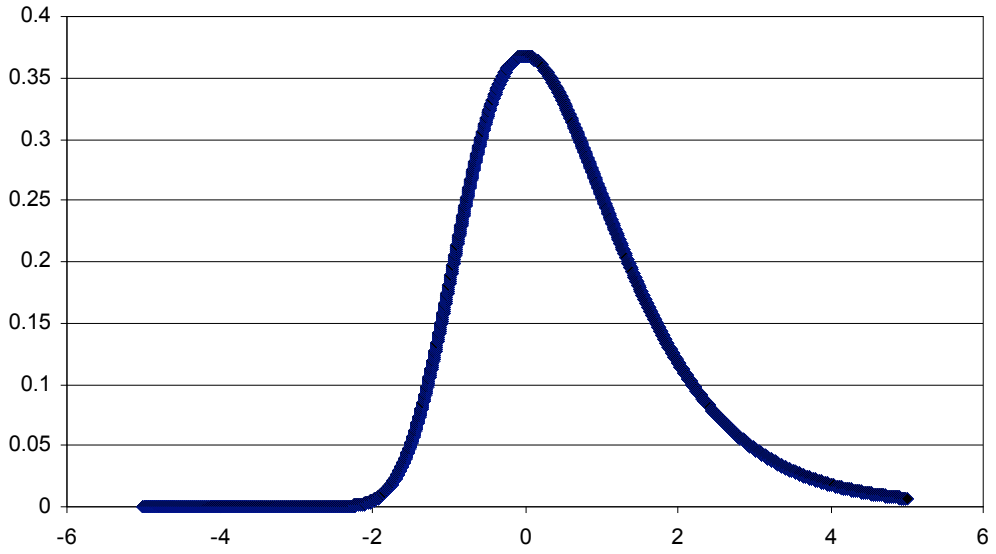
In practice it is important not only to compute the optimal alignment between two sequences, but also to estimate the statistical significance of the alignment. The simplest method to compute the probability of an alignment is to repeatedly randomly permute one of the sequences, align it to the other sequence and measure how often it produces a score that is

higher than the alignment score between the two real sequences. It has been found that the distribution of alignment scores between two randomly permuted sequences approximates an extreme value distribution (see figure 1). The probability of observing a score greater than the actual alignment score between two sequences is given by

$$P(S > S_0) = 1 - \exp(Kmne^{-\lambda S_0})$$

where m and n are the lengths of the two sequences and K and λ are two parameters that describe the width and mean of the distribution and are specific for a particular set of sequences.

Figure 1. The Distribution of Alignment Scores between Two Randomly Permuted Sequences Approximates an Extreme Value Distribution



Note: This distribution is given by the equation: $P(x) = e^{-x}e^{e^{-x}}$

In contrast to the normal distribution the extreme value distribution has an exponentially decaying tail for large values of x. Therefore in an extreme value distribution a large value of x is more probable than in a normal distribution.

The program BLAST does not report the probability of observing a more significant alignment between two sequences but the expectation of observing a particular alignment score if one is aligning a sequence to a large database of protein sequences. Starting with the extreme value distribution of alignment scores and the above equation for the probability of observing a score greater than S_0, it is possible to calculate the expectation of observing an alignment score greater than S_0:

$$E(S > S_0) = Kmne^{-\lambda S}$$

where the parameters K and λ are computed a priori and held fixed for the entire search between a sequence and a database. Usually when the expectation is less than 0.001 the alignment is considered significant, but this threshold varies depending the particular application.

Based on experimental evidence, we assume that the molecular function of the query protein is likely to be similar to that of the proteins to which it aligns significantly. Therefore by looking at the annotations of the aligned proteins we can rapidly get clues about the molecular function of the query protein.

In some cases BLAST only produces a limited number of significant homologs, as measured by the expectation of observing an alignment score by chance. In these cases it is possible to incorporate additional information in order to extend the ability of alignments to recognize distant homologs. This has been accomplished, in part, by using transitive sequence comparisons (Park 1997,Gerstein 1998). Each sequence may be homologous to a limited number of others. These in turn may be homologous to other sequences not in the original set. It is reasonable to postulate that the original query protein is also homologous to the homologs of its homologs. Several researchers have shown that such a procedure may reveal a distant relationship between proteins that are known to be structurally similar, and hence may have a related function, that could not be found by conventional alignment techniques. Other techniques, such as multiple alignments based on an initial starting sequence (see PSI-BLAST discussion below), can also be used to find distant homologs.

Multiple Alignments and Domains

Proteins are composed of both structural domains, regions that form stable 3-D structure by themselves, and loop regions, less ordered short regions that connect structural domains. Typically, domain sequences are more conserved across homologous proteins than loop regions, since domains are usually the most functionally and structurally important regions of the protein. The pattern of conservation of protein segments across families of homologous proteins may be revealed by extending the above pair-wise alignment techniques to construct multiple sequence alignments and identifying the highly conserved regions.

Constructing multiple sequence alignments is more complex than computing pair-wise alignments between proteins, since there is not a straightforward technique to identify the optimal multiple alignment between groups of related sequences. However, many techniques have been developed that find approximate solutions to this problem by combining pair-wise alignments. A commonly used program to perform such multiple alignments is ClustalW (Higgins 1994).

The study of multiple sequence alignments and domains not only elucidates aspects of a protein's structure but is also helpful to understand a protein's function. Protein domains often perform a specific functional task. They may for instance contain the functional residues for an enzymatic reaction. Proteins are often composed of multiple domains that perform disparate functions. Therefore, by assigning molecular functions to domains it is possible to enumerate all the molecular functions that a protein may perform.

One of the more popular tools for constructing multiple alignments, and studying domain functions is a variant of the BLAST program called PSI-BLAST (Altschul 1999). Unlike the standard BLAST program that uses a single substitution matrix this program constructs a position specific substitution matrix specific for the query sequence. To construct this matrix one first generates a multiple alignment using the regular BLAST results by aligning all sequences with significant homology to the query sequence using pair-wise alignments. This multiple alignment is then used to compute the amino acid distribution at each position of the alignment. From this information the position specific matrix is created that describes the likelihood of finding one of the twenty amino acids at each position of the query sequence. The alignment of the query sequence against the whole database is then repeated using the new position specific matrix. Scoring alignments using the position specific substitution matrices can detect distant homologs to the query sequence that might be missed using a position independent substitution matrix. Since domains are more strongly conserved than loop regions, alignments from an exhaustive calculation using PSI-BLAST between all known protein sequences have been clustered and used to define protein domains (and their boundaries) in the PRODOM database (Corpet 2000).

More sensitive multiple alignment protocols have also been developed. For instance, hidden Markov models (HMM) for protein families are able to detect remote homologs that may be missed by simpler multiple alignment techniques (Krogh 1994). HMM's consist of Markov chains, sequential conditional probabilities that describe the likelihood of inserting any one of the twenty amino acids or a gap at the next position in the chain. The many parameters that are used to define a HMM are trained against sets of homologous proteins. The models are then used to identify other sequences in a database that score well against the model. HMM's have been created for most protein families and are compiled within the PFAM database (Bateman 2000). Other efforts to construct hidden Markov models for protein families include SMART (Letunic 2002), TIGRFAMs (Haft 2003), PIR SuperFamily (Wu 2003) and SUPERFAMILY (Gough 2001).

Other approaches search for shorter motifs within domains that are conserved within protein families. In contrast to a domain sequence, which represents a structurally stable 3-D region of a protein, a motif pattern is generally very short including just a few amino acids. Often motifs reflect a small recognition surface on the protein. For instance it is known that potential N-glycosylation sites are specific to the consensus sequence Asn-Xaa-Ser/Thr. In this case the motif includes just three amino acids, the first of which is an asparagine, the site of N-glycosylation, the second any amino acid and the third either a serine or a threonine. It must be noted however that the presence of this motif is not sufficient to conclude that an asparagine residue is glycosylated, due to the fact that the nearby folding of a protein plays an important role in the regulation of N-glycosylation and that there is a non-dismissible probability that such a short motif may occur in a protein sequence by random chance.

Other motifs contain highly conserved residues that are involved in the active sites of enzymes, commonly seen repeats or entire domain sequences. A large set of protein motif sequences are described in the PROSITE database using regular expressions and position-specific matrices (Flaquet 2002). Regular expressions enumerate which residues are allowed to be present at a specific location of a motif. For instance, the regular expression corresponding to the glycosylation motif mentioned above is:

N-{P}-[ST],

where N implies asaparagine at position 1, {P} any amino acid at position 2 and [ST] a serine or a threonine at position 3. An extension of PROSITE called PRINTS (Attwood 2003) builds domain signatures based on groups of sequence motifs.

All of the above domain databases have been merged into the InterPro database (Mulder 2003). This database provides a compendium of domain definitions that are all cross referenced and associated with a single protein indexing scheme. As such it represents an invaluable resource for enumerating protein domains, and studying the molecular functions of protein sequences.

A slightly different approach to the ones mentioned above involves the use of graph analysis applied to a database of alignments. This approach allows one to rapidly cluster protein families and decompose proteins into their respective domains (Enright 2000) based on the analysis of pathways through the resulting graph.

CO-EVOLUTION OF NON-HOMOLOGOUS PROTEINS

Protein Fusions

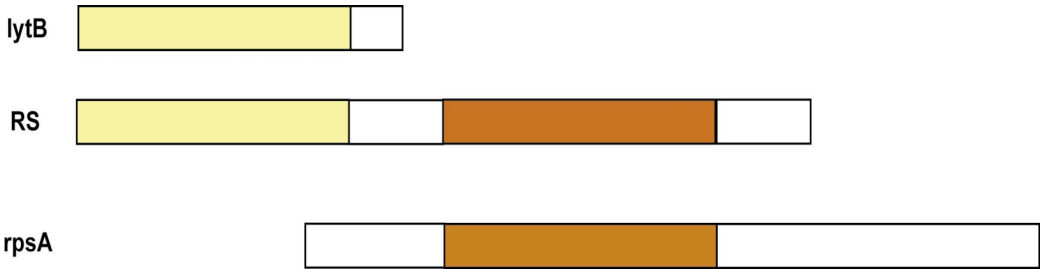
All the methods discussed so far study the evolution between homologous proteins. However, over the past few years various techniques have been developed to study the evolution, and functional relationship, between non-homologous proteins. The first example of these techniques we discuss is the search for protein fusion events that may have occurred between non-homologous proteins.

In the past few years protein sequence alignments have been used to search for protein fusions (Marcotte 1999a, Enright 1999). During the course of evolution two genes encoding distinct proteins may fuse together to encode a single polypeptide. In certain cases such a fusion event may confer an evolutionary advantage to the organism by bringing into physical proximity domains that perform related functions. For instance, fusing together two enzymes that perform sequential enzymatic reactions in a metabolic pathway may generate a single enzyme that catalyzes the two reactions more efficiently. It has been shown that in general two proteins that undergo fusion events are more likely to interact or participate in the same biological process than two random proteins.

It should be mentioned, however, that we do not necessarily know whether the fused protein or the separate proteins are the ancestral protein species. In other words, two genes may have fused or a single gene may have split into two separate genes coding for separate proteins. When enough evolutionary information exists for the organisms involved, a fusion event may be distinguished from a splitting event. However, since it is often difficult to distinguish between these two scenarios, for simplicity we will refer to both of them as fusion events.

To identify fused proteins we seek two non homologous proteins that align to different regions of another protein (see Figure 2). In other words, the sequences of the two proteins are essentially fused into a single longer polypeptide chain. The longer protein has been dubbed the ``Rosetta Stone" protein, because it often reveals that the two unfused proteins are interacting or involved in the same biological process.

Figure 2. The Rosetta Stone Technique Searches for Non-Homologous Protein Sequences that have Fused into a Single Polypeptide



Note: In this figure we see an example of two *E. coli* proteins, lytB and rpsA that are fused into a single protein in the organism *Carboxydothermus hydrogenoformans*. Regions with homology are shown in the same color in two sequences. Regions with no homology to the other two sequences are shown in white.

The completed sequencing of many whole genomes has increased the power of this approach by greatly increasing the number of identifiable Rosetta Stone proteins. Additionally, this method is useful in the initial characterization of each newly sequenced genome. This is accomplished by identifying pairs of fused proteins where one member of the pair is characterized and the other is not. In this case, we associate the function of the characterized protein to the uncharacterized one.

However, this approach may also yield a significant number of spurious fusion events due to the fact that, as we saw in the previous section, protein sequences contain conserved domains. The presence of domains makes protein sequences inherently modular and many of the modules are repeated hundreds of times throughout a protein sequence database. For example, finding a protein with a kinase domain that aligns to a Rosetta Stone protein is less likely to represent a true fusion event since many proteins within the human genome contain kinase domains.

To estimate which protein fusion events are more likely to be real and hence to link together proteins that are likely to interact it has recently been suggested that one may use the hypergeometric distribution (Marcotte 2002):

$$P(k \mid n,m,N) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$$

where k represents the number of Rosetta Stone proteins found between two non-homologous proteins A and B, n the number of homologs of protein A, m the number of homologs of protein B and N the total number of proteins within the database. According to this function, proteins that have many homologs are less likely to be involved in real protein fusion and more likely to "appear" fused because they contain a commonly found domain just as it would be more likely to draw two non-face cards than two face cards since a deck of cards contains more non-face than face cards. In the end one cannot computationally distinguish a

true fusion from an apparent but false fusion, but it is nonetheless helpful to be able to rank results based on their probability of being true. In fact, it has been shown that pairs with more significant P values are more likely to participate in the same biological process than pairs with less significant ones (Marcotte 2002).

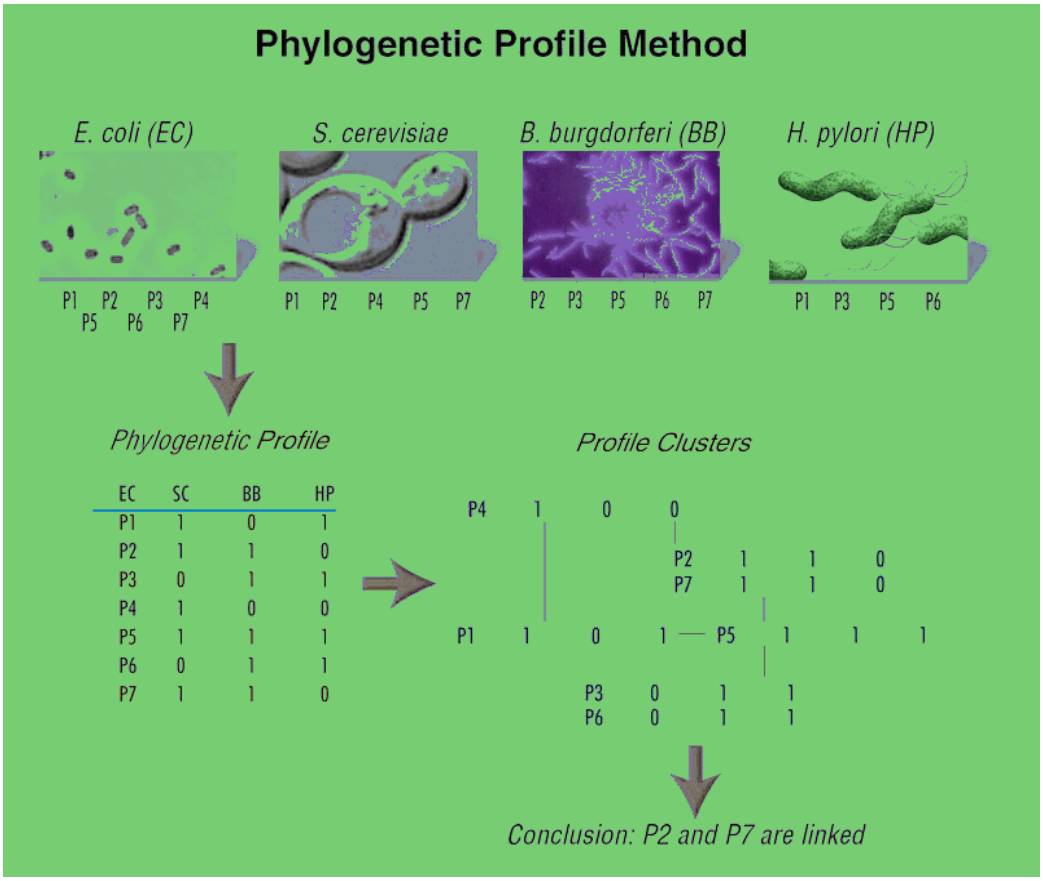
Phylogenetic Analysis of Alignments

Over the past few years the numbers of fully sequenced genomes has grown dramatically to include over one hundred organism, including human. The analysis of this data is already yielding significant information about protein function. At the simplest level, it is now possible to classify proteins into clusters of orthologous groups (COGs) (Tatusov 1997, Tatusov 2000). Orthologs are defined as protein homologs found in different organisms that have descended from a common ancestor through speciation. It is important to identify orthologous proteins because they typically perform equivalent functions. The COG methodology uses a graph based analysis to construct highly connected sets of orthologous proteins. The proteins within a COG are all assumed to perform equivalent functions.

Operationally orthologs may be identified by finding pair-wise best hits: two proteins that are their closest homologs when two organisms are compared (Overbeek 1999). In other words if the closest homolog of protein A from genome 1 is protein B in genome 2, and the closest homolog of protein B in genome 2 is protein A in genome 1, then it is likely that proteins A and B are orthologs.

From the analysis of orthologs across genomes it is also possible to construct phylogenetic profiles (Pellegrini 1999). These are binary arrays computed for each protein that encode whether an ortholog of the protein is present in any of the fully sequenced genomes (see figure 3). In practice, one may construct phylogenetic profiles using both the presence of orthologs or simply homologs as the criteria. Proteins with similar phylogenetic profiles are effectively co-evolving, since they are often found together in organisms. It is not surprising therefore to find that they are usually members of cellular complexes or proteins that participate in the same biological process.

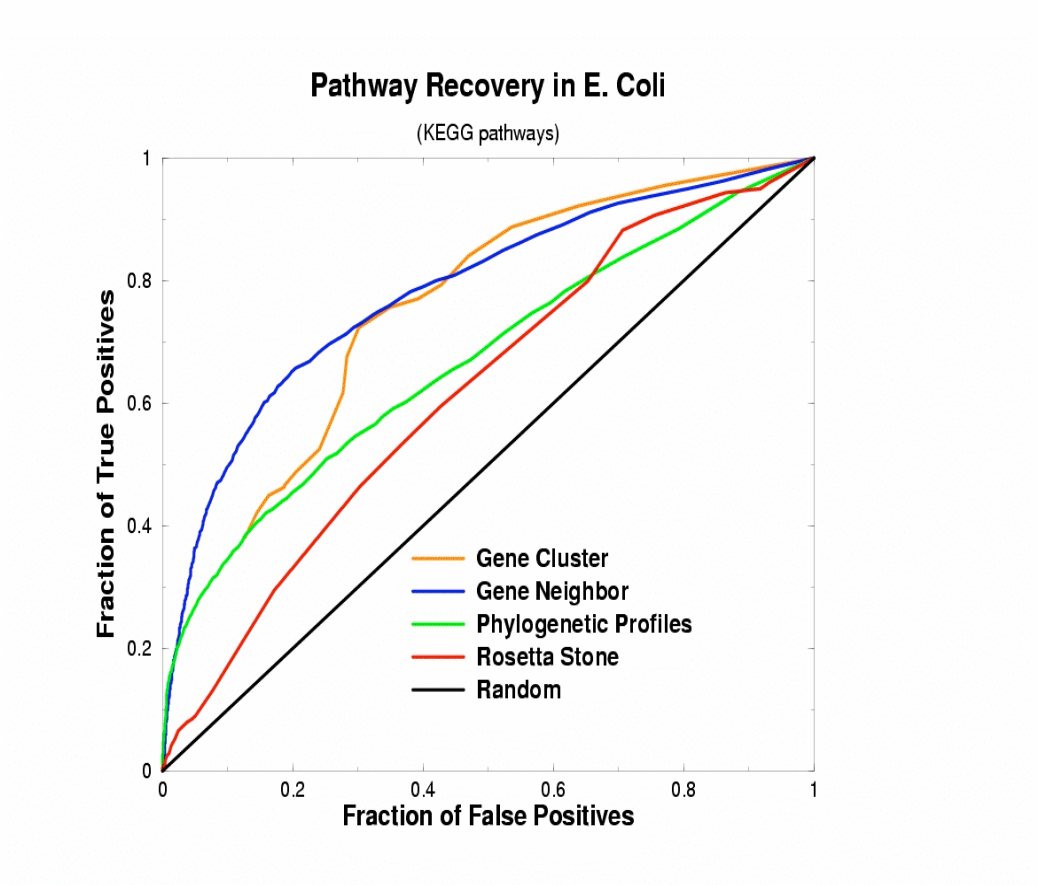
Figure 3. Phylogenetic Profiles Provide a Method for Identifying Co-evolving Genes



Note: The figure displays a schematic of the construction of phylogenetic profiles from four genomes. Homologs of the genes in the starting genome (*E. coli*) are identified in the three other genomes (*S. cerevisiae, B. burgdorferi* and *H. pylori*). The presence of a homolog is represented with a one and the absence with a zero. The phylogenetic profiles are then clustered and pairs of genes with similar profiles are identified. In this example *E. coli* proteins P2 and P7 are identified as being linked, and likely co-evolving, because of their similar profiles.

Several metrics are available to measure the similarity of two phylogenetic profiles. The simplest is the computation of the hamming distance: the number of bits that differ in two binary profiles. However, it has been shown that one obtains more accurate estimates of profile similarity by using either the hypergeometric distribution (Wu 2003b) or mutual information (Date 2003). To establish the accuracy of a metric one typically uses a receiver operator characteristic (ROC) curve to measure the number of true positive versus false positive interactions using a set of known interactions as a benchmark (see figure 4). In this case, proteins that are known to participate in the same biological process are considered true positives while those that are not are false positives.

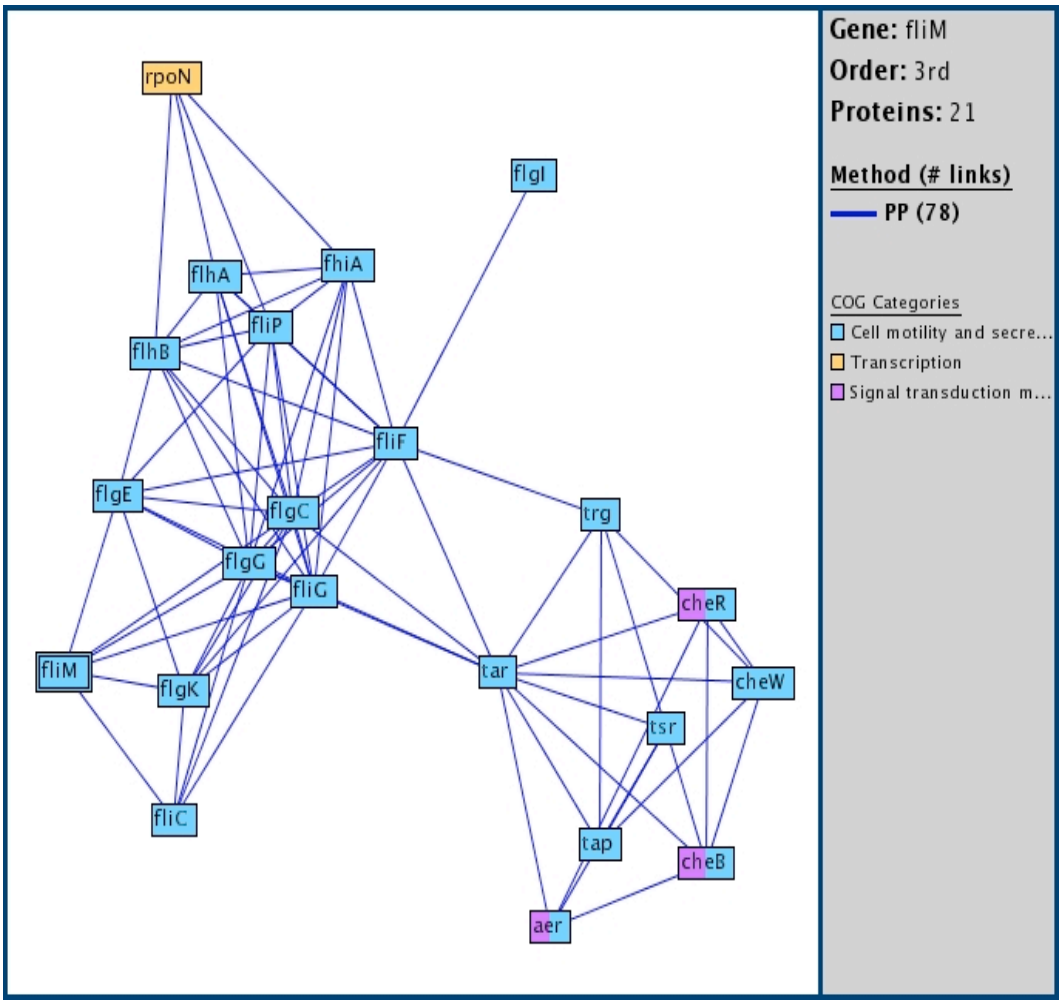
Figure 4. Receiver Operator Characteristic (ROC) Curves are Used to Assess the Accuracy of the Various Methods that are Used to Find Pairs of Functionally Linked Proteins



Note: For each of the four methods we first rank order protein pairs by their statistical significance (see main text). We then measure the number of true positive versus false positive pairs as a function of rank using a set of known interactions as a benchmark. In this figure, true positive pairs are those for which both proteins are annotated in the same KEGG pathway and false positive are those annotated in different pathways. The diagonal line represents a random selection of protein pairs. The fact that the measured curves are above the diagonal random curve suggests that the methods are detecting more true positive pairs than false positive ones.

In figure 5 we show an example of a protein network that is constructed using phylogenetic profiles. In this case the similarity of the phylogenetic profiles was established using the hypergeometric distribution and only those pairs that were deemed statistically significant were drawn as edges. The network represents a subgraph of all proteins that are no more than three edges removed from fliM (in box with double lines), a component of the *Escherichia coli* flagella.

Figure 5. The Figure Displays a Network of Genes with Similar Phylogenetic Profiles



Note: The network is constructed starting with the *E. coli* fliM gene and then adding all other genes that are no more than three edges removed. The network includes many flagellar genes as well as genes involved in the chemosensing signaling cascade, all of which are involved in cell motility or chemotaxis.

This graph illustrates how proteins that participate in the same biological process tend to have similar phylogenetic profiles. We see that the proteins with profiles most similar to fliM are the other components of the *E. coli* flagella. We also see in the lower right part of the graph that several genes that are part of the environmental sensing apparatus (chemosensing) are also connected to theses genes. It is not surprising that proteins that are responsible for chemosensing would co-evolve with flagellar proteins since they are all involved in the process of cellular motility, or chemotaxis.

The clustering of proteins on the basis of the similarity between their phylogenetic profiles is significantly different from the clustering of proteins based on their sequence similarity, because the proteins within a phylogenetic profile cluster may share no sequence

similarity between themselves. Thus, these profiles can be used to link together proteins that are not homologous to each other, yet participate within the same biological process.

Predicting Interactions of Paralogous Proteins

As we saw in the previous section, the study of protein co-evolution yields clues about a protein's interactions and functions. However, using phylogenetic profiles it is difficult to discern subtle differences in the evolution of paralogs. Paralogs are homologous proteins that have emerged by duplication within a species and would necessarily have very similar phylogenetic profiles because of their homology, even though they may have evolved to perform slightly different functions. In order to study the subtle differences in interactions and function between paralogs it is therefore necessary to more completely describe their evolution using standard evolutionary distance estimation techniques.

In order to estimate the evolutionary distance within a group of homologous proteins, one must first construct a multiple sequence alignment. As we noted in section 3.2, this may be accomplished using the ClustalW program (Higgins 1994). Once the multiple alignment has been built it is possible to estimate the evolutionary distance between any two sequences using the alignment score. One use of this information is to deduce the different interacting partners of paralogs by comparing the distance matrices of two protein families that are known to contain interacting pairs.

One might imagine that if two proteins interact, the evolution of one might be correlated with the other. For instance, mutations that occur on a ligand might be compensated by mutations to its receptor in order to maintain the ligand-receptor binding affinity. This phenomenon has in fact been demonstrated in the case of chemokines and their associated receptors (Goh 2000, Goh 2002). By correctly aligning the distance matrices of ligands and ligand receptors using Monte Carlo techniques it is possible to partially reconstruct which ligand is likely to bind which receptor (Ramani 2003).

Gene Neighbors

Genome sequences tell us not only which genes are coded within them, but also where on the genome the gene is located. Knowledge of the position of genes on a genome allows us to identify which pairs of genes are frequently coded next to each other across multiple bacterial genomes. We will refer to these pairs as gene neighbors. It has been observed that gene neighbors often code for proteins that are involved in the same biological process and therefore the computation of gene neighbors complements the use of phylogenetic profiles and protein fusions to study protein functions (Tamames 1997, Dandekar 1998, Overbeek 1999).

The likely reason that two genes are found nearby in multiple genomes is that they are members of a conserved operon. Operons are sequential genes on the same DNA strand that are transcribed as a single unit and are primarily found in prokaryotes. The genes within a bacterial operon are therefore transcriptionally co-regulated. Typically bacterial operons have

evolved to contain genes that participate within the same biological process, allowing the organism to coordinately regulate this process at the level of transcription.

Because of the conservation of operons across bacteria, and the functional relationships between genes within an operon, it is possible to use the analysis of gene neighbors to study protein functions. It has been shown that the gene neighbor analysis, when combined with conventional homology based methods, yields functional information on the biological process of the vast majority of genes encoded in newly sequenced genomes (Selkov 2000).

The study of gene neighbors can be used to both study gene functions and to reconstruct the operon structure of particular bacteria. Several groups have applied these methodologies to reconstruct in great detail the full operon structure of *E. coli* (Salgado 2000, Ermolaeva 2001). In the near future, this in depth knowledge of operon structures is likely to be deciphered for all fully sequenced microbes, leading to more sophisticated models of bacterial transcriptional regulation.

EXPRESSION METHODS

Expression Platforms

As a result of remarkable developments during the past few years, it is now possible to measure the concentrations of nearly every mRNA within a cell. There are two primary ways to accomplish this. The first is by sequencing short fragments of mRNA (Velculescu 1995, Brenner 2000) and counting the number of copies of a particular gene. Although these techniques are useful for identifying which genes are being transcribed within a cell they are not optimal for estimating transcript abundances.

The second and more popular technique involves hybridizing fluorescent mRNA to complementary sequences that are arrayed on a chip, and then estimating the concentrations by the fluorescence intensity. Over the past few years gene chips have become a standard tool in genomics research. Several different techniques have been used to manufacture gene chips including photolithography to synthesize short oligonucleotides on an array (Fodor 1993), ink jet printers to attach cDNA to a glass slide (Schena 1995) and ink jet printers to synthesize short oligonucleotides on an array (Hughes 2001). In each case, it is now possible to measure 10,000 or more different genes per chip, making expression microarrays the most advanced form of molecular profiling to date.

Clustering Techniques

Typically, the expression levels of the genes within a cell are measured under varying conditions. For instance, one may measure the concentrations of yeast genes at different times during the cell division cycle (Spellman 1998), in different yeast strains where certain genes have been knocked out (Hughes 2000) or when drug-like compounds are added to a cell (Wilson 1999). The result is that each gene has an associated expression vector that describes its concentrations in the cell under different experimental conditions. Various techniques

discussed below have been developed to use this information to study the function of genes and the proteins for which they code.

In other cases the different experimental conditions of measured samples may not be fully known. For example, gene array data may be collected for various patients whose disease status is initially unknown. By analyzing the data it is possible to categorize the samples into disease and non-disease states or disease sub types. This type of diagnostic has been shown to be useful in diagnosing the cancer subtype or disease aggressiveness for individual patients (Shipp 2002).

One of the most common forms of analysis to study gene functions and experimental categories involves the clustering of gene expression vectors. It has been found that genes that cluster together are likely to code for proteins that function in the same biological process. Therefore, gene expression clustering may be used in a manner very similar to that of the co-evolutionary techniques discussed above to assign approximate functions to all the proteins within a cell using the observation that co-clustered genes often have similar functions.

Similarly one may also cluster experiments rather than genes. Just as in the case of genes, experiments within the same cluster are likely to share common properties. For instance a cluster of experiments collected from diverse individuals may contain patients with a similar disease sub type.

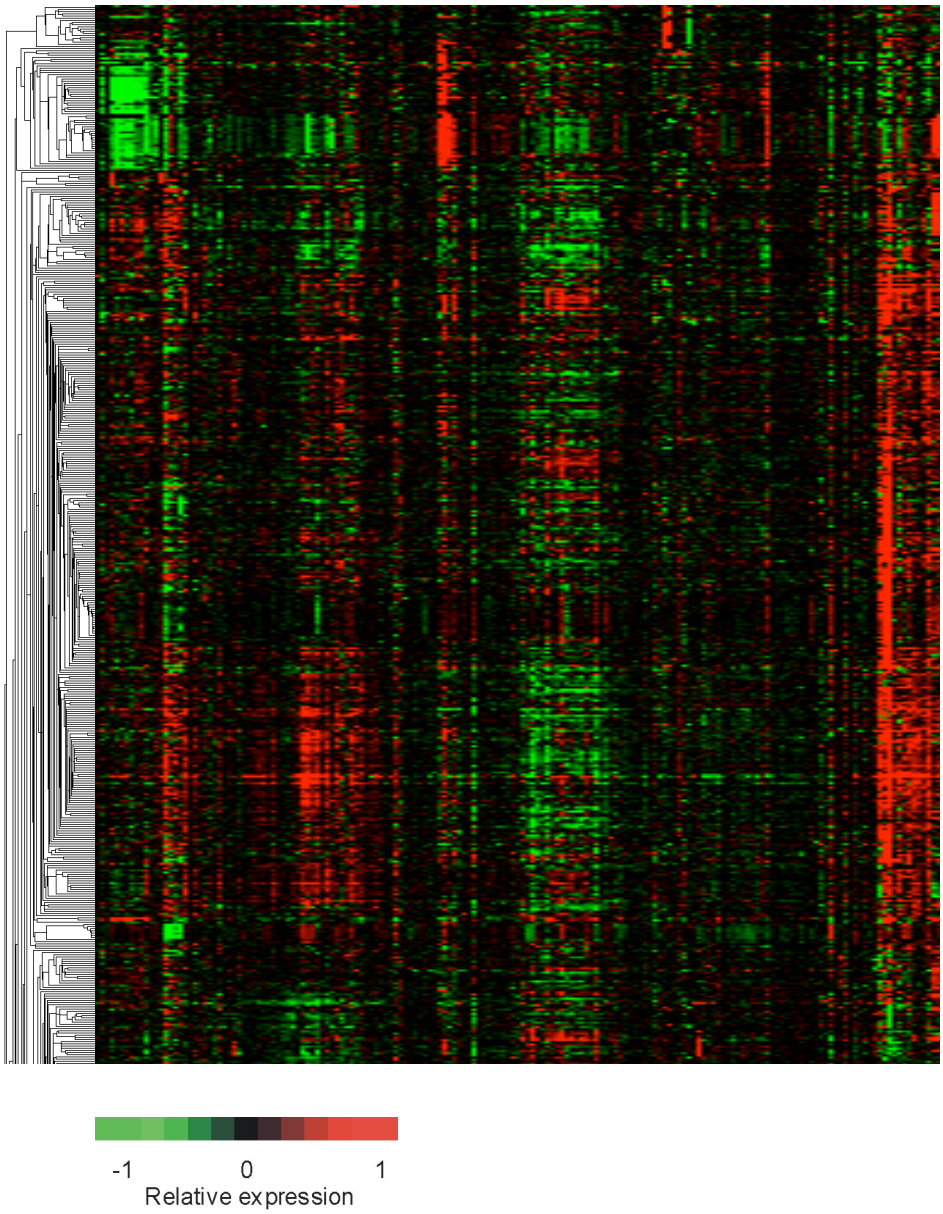
In order to cluster expression data a measure of similarity between the expression vectors must first be determined. The most common metric used is the Pearson correlation coefficient,

$$r(x,y) := \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N-1}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}},$$

although one may also use Euclidian distance or mutual information among others.

A variety of approaches have been developed to cluster expression data based on the degree of similarity between expression vectors. The most popular is hierarchical clustering (Eisen 1998). This technique generates a dendrogram of gene relationships that may be visualized as a hierarchical tree (see figure 6). The dendrogram is generated in an iterative fashion by first selecting the two most similar expression vectors, linking them, and then treating them as a single expression vector when searching for the next most similar pair. This process is repeated until all expression vectors have been included in a single dendrogram on which the lengths of the branches represent the distance between vectors. Once the tree has been constructed, clusters are generated by selecting all groups in the dendrogram that are separated by no more than a threshold distance.

Figure 6. One of the Most Popular Techniques for Visualizing Gene Expression Data from Microarrays is Known as Hierarchical Clustering



Note: In this figure we see a hierarchical cluster of data from yeast (Hughes 2001). The rows in the graph represent genes, while the columns represent experiments. The dendrogram on the right is constructed as described in the main text by successively linking together the two most similar genes. Therefore, neighboring rows contain genes with similar expression profiles. Hierarchical clustering is also performed on the experiments, so neighboring columns contain similar experimental conditions. Genes that are closely linked in the dendrogram tend to participate in the same biological process.

Another approach is K means clustering, where the data is partitioned into a predetermined number of clusters. In this method the cluster means are first selected at random. Each gene is assigned to the cluster mean that is most similar to itself. Once all the genes have been assigned to a cluster, new cluster means are computed. This process is iterated until the clusters no longer change.

These techniques have been used countless times over the past few years to identify previously uncharacterized components of cellular pathways. As a typical example, microarray clusters were used to identify the components of a system for phosphate accumulation and polyphosphate metabolism (Ogawa 2000). In this case yeast cells were grown in various conditions that varied in their P_i concentrations. Genes that showed differential expression across these conditions were clustered. An analysis of these clusters revealed several genes that were previously unknown to participate in phosphate pathways, but proved to be components of these pathways once they were further experimentally characterized. This analysis also revealed insights into the role of polyphosphate metabolism that had previously been overlooked. Through the construction of yeast strains that lacked some of the key phosphate genes, the authors demonstrated that polyphosphate metabolism is an important mechanism by which yeast cells store intracellular phosphate.

The ability to measure all gene expression levels within a cell offers biologists an entirely new approach to studying the cellular function of proteins without relying on sequence homology. As databases of gene expression arrays grow, this data will provide an important contribution to our efforts to study protein functions.

Efforts are also underway to develop technologies to measure protein concentration levels in a high throughput fashion. However, it is more difficult to measure individual protein concentrations from a complex mixture of proteins than it is for mRNA. This is in part due to proteins having a diverse range of chemical properties whereas mRNA sequences have very similar hybridization properties independent of their particular nucleotide sequence.

Transcriptional Networks

Cells typically contain hundreds of transcription factors that affect the rate of transcription of most genes. Each of these factors may affect many genes that could in turn code for other transcription factors. The reverse engineering of a cell's transcriptional network remains one of the major challenges for 21st century biology.

There are various techniques that have been developed in attempts to study transcription networks. Since transcription factors typically bind to promoter regions of genes that contain nucleotide sequence motifs specific for each transcription factor, it is in principle possible to study the network by searching for all motifs in a genome. Several programs have been developed to search for these motifs, among which are MEME (Bailey 1995) and the Reverse Gibbs Sampler (Thompson 2003). However, because the regulatory DNA motifs to which a transcription factor binds are generally short and can occur at random throughout a genome, not all promoter sequences matching a motif are functional. A functional motif is one that when mutated effects the transcription of nearby genes. It has proven difficult to computationally predict which motifs are functional and thus the elucidation of transcriptional

networks generally requires additional experimental information than just the promoter sequences.

A complementary approach to understanding transcription networks involves the experimental identification of which promoters are actively bound by specific transcription factors. These interactions have recently been mapped within the yeast *Saccharomyces cerevisiae* (Lee 2002). The strategy used to map the biding sites of yeast transcription factors consisted of first adding myc epitope tags into the genomic sequence of the COOH terminus of each regulator. Chromatin immunoprecipitation was then used to identify the DNA sequences bound to the transcription factors. In total the authors were able to identify 3985 high confidence interactions, which reveal a complex network of transcriptional regulation.

Since gene expression arrays directly probe transcriptional regulation, they can also be used to aid in the reconstruction of transcriptional networks. A first step involves asking whether the co-expressed genes share common cis regulatory sequences in their upstream promoters. Clustering genes and aligning their respective promoters has allowed scientists to both verify known motifs as well as identify previously unknown ones (Hughes JD 2000, Bussemaker 2001).

A more complex approach to the reconstruction of transcriptional networks from expression data involves the use of Bayesian networks (Friedman 2000). These networks model the multidimensional probability distribution of gene expression levels by determining the dependencies between genes in the form of a directional acyclic network. Once these models are constructed they allow one to qualitatively study the connections inherent in the transcriptional network as well as in principle to predict the outcome of perturbing a component of the network.

PROTEIN INTERACTION METHODS

Experimental Techniques to Measure Protein Interactions

Functional interactions between proteins can be defined in many ways. For example, two proteins that sequentially modify a metabolite are functionally related. Of course, many functional interactions also involve direct physical interaction.

Several experimental techniques have been developed to directly probe protein interactions within a cell in a high throughput fashion. The two-hybrid technique is based on the construction of a bait and a prey protein that are fused to two halves of a transcription factor (Fields 1989). If the bait and the prey protein interact the transcription factor is reconstituted and its activity is measured though the activation of the transcription of a reporter gene. This approach is a specific example of a general class of protein fragment complementation assays (PCA). As in the two-hybrid approach, in PCAs half of a reporter protein is fused to protein A and the other half to protein B. If protein A and B interact, the two halves of the reporter protein reconstitute to restore its activity. The assay then reads out the activity of the reporter protein (Pelletier 1998).

Protein interactions may also be directly monitored using various co-purification techniques. A protein may be directly purified using a specific antibody or the protein may be

tagged with another protein or a small molecule tag and then affinity purified. In all cases, if the selected protein interacts with other proteins these will be co-purified. The identity of the interacting partners may be deduced using mass spectrometry among other techniques (Gavin 2002).

Protein microarrays are also emerging as a promising tool to directly observe protein interactions in a parallel fashion. Protein chips are the protein counterpart of DNA chips which are widely used to measure gene expression levels. In a protein chip each spot consists of a different purified protein. By studying the binding of proteins labeled with fluorescent molecules to these chips it is possible to reconstruct protein interaction patterns (Zhu 2001).

Protein Interaction Databases

The data reported from experiments that probe direct protein-protein interactions has been catalogued within various databases such as the Database of Interacting Proteins (DIP) (Xenarios 2002, see Table 1) and the Biomolecular Interaction Network Database (BIND) (Bader 2003). These databases contain interactions measured in many different organisms, however the majority of these interactions involve yeast proteins. Currently there are about 15,000 interactions between *Saccharomyces cerevisiae* proteins reported in these databases. When viewed as a network these relationships represent a comprehensive view of protein interactions within yeast, and thus far involve about two thirds of the yeast proteome.

Table 1. Statistics for the Database of Interacting Proteins

Number of proteins	7141
Number of organisms	104
Number of interactions	18670
Number of distinct experiments describing an interaction	22918
Number of articles used to build database	2507

Although it is not known how many direct physical interactions between yeast proteins will ultimately be measured, it is found that in comprehensive two-hybrid screens each protein engages in an average of only three or four interactions (Uetz 2000). If this is in fact an accurate estimate of the true number of interactions per protein we conclude that the current catalogue will not grow by many multiples in the future.

Prediction of Protein Function from Interaction Networks

Knowledge of a protein's interactions can shed a great deal of light on its function. It may, for instance, allow one to understand the substrates of an enzyme or the complex in which a protein functions.

Protein interaction networks may also be used to assign proteins to broad biological processes. The simplest algorithm that has been implemented to exploit protein networks for functional annotation is a voting scheme: a protein is assigned to the biological process that is

most often present among its interacting partners (Schwikowski 2000, Hishigaki 2001). This simple scheme has been found to assign proteins to biological processes with approximately 50% accuracy.

More complex algorithms for assigning proteins to biological processes using protein interaction networks have also been recently proposed. One example involves the minimization of a score function:

$$E = -\sum_{i,j} J_{ij}\delta(\sigma_i,\sigma_j) - \sum_i h_i(\sigma_i)$$

where J_{ij} is the adjacency matrix for the protein interaction network and the σ's are the binary biological process vectors with entries of 1 if the protein is involved in a particular biological process and 0 otherwise (Vazquez 2003). Minimizing this energy function has been shown to yield improved predictions with respect to the simpler voting approach.

COMBINED METHODS

The methods described above exploit different properties of proteins to gain functional insights. Often, these properties generate information on different sets of proteins. It is therefore useful to combine these methods to gain a more complete picture of protein function (Eisenberg 2000, Galperin 2000, Teichman 2000, Huynen 2000, Aravind 2000).

Databases have been constructed to combine the evolutionary based methods described in section 4 for deducing protein couplings (Pellegrini 2001, Mellor 2002, von Merring 2003). These methods include phylogenetic profiling, protein fusion analysis, gene neighbor analysis and the reconstruction of operons. To date these have been mostly applied to bacteria where the existence of operons and numerous fully sequenced genomes renders the methods more successful. It has been shown that these networks may be used to accurately assign biological processes to uncharacterized genes.

One approach to combine the above techniques with expression data treats every pair-wise prediction as a link between two proteins (Marcotte 1999b). That is, proteins are linked if they have similar phylogenetic profiles or expression profiles, or if they are neighbors on multiple genomes or if they are fused within a Rosetta stone protein. By studying the graph of links for yeast, it is possible to infer approximate functions for most of the uncharacterized genes coded by this genome.

Another approach combines the phylogenetic profiles and expression profiles into a single data structure (Pavlidis 2000). By concatenating the binary phylogenetic profiles with expression vectors one may construct a single vector for each gene. Support Vector Machines, trained on annotated genes, are then used to classify these vectors into functional categories. The results of this analysis demonstrate that these combined data structures are able to recover functional information for a greater number of genes than any one of the methods alone.

CONCLUSION

The advent of whole-genome sequencing and mRNA profiling, has created new opportunities for computational biologists. It is now possible to utilize information from comparative genome analysis to reconstruct a protein's evolution, and hence gain insights into its function. The ability to probe the expression levels of every gene within a genome is also revolutionizing our ability to understand transcriptional regulation and the function of co-regulated proteins.

In the future, these insights will be used by computational biologists to model cellular pathways in great detail (Tomita 1999). It is already possible to begin to model developmental pathways (Von Dassow 2000), metabolic pathways (Edwards 2000) and signal transduction pathways (Schoeberl 2002) and compare the predictions of these models to experimental results. In the next few years there will undoubtedly be exciting new approaches that combine genome wide experimental measurements with complex mathematical modeling, to gain an unprecedented understanding of protein function and cellular biology.

REFERENCES

[1] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W. & Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Research 1999, 25:3389-3402.

[2] Aravind L: Guilt by Association: Contextual Information in Genome Analysis. Genome Res. 2000, 10:1074-1077.

[3] Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C: PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. 2003, 31:400-402.

[4] Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 2003 Jan 1;31(1):248-50.

[5] T. L. Bailey and C. Elkan. *The Value of Prior Knowledge in Finding Motifs with MEME*. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology* (ISMB'95), pp. 21-29. Cambridge, England, July 1995

[6] Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: The Pfam protein families database. Nucleic Acids Res. 2000, 28:263-266.

[7] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL: The Pfam Protein Families Database. Nucleic Acids Res. 2002, 30:276-280.

[8] Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using correlation with expression. Nat Genet. 2001, 27(2):167-71.

[9] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd D, Johnson D et al.: Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature Biotechnology 2000, 18:630-634.

[10] Corpet F, Servant F, Gouzy J, Kahn D: ProDom and ProDom-CG: Tools for protein domain analysis and whole genomecomparisons. Nucleic Acids Res. 2000, 28:267-269.

[11]Dandekar T, Snel B, Huynen M, Bork P: Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 1998, 23:324-328.

[12]Date SV, Marcotte EM: Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol. 2003, 21(9):1055-62.

[13]Edwards JS, Palsson BO: The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc. Natl. Acad. Sci. USA 2000, 97:5528-33.

[14]Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998, 95:14863-14868.

[15]Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: Protein function in the post-genomic era. Nature 2000, 405:823-826.

[16]Enright AJ, Iliopoulos I, Kyrpides N, Ouzounis CA: Protein interaction maps for complete genomes based on gene fusion events. Nature 1999, 402:86-90.

[17]Enright AJ, Ozounis CA: GeneRage: a robust algorithm for sequence clustering and domain detection. Bioinformatics 2000, 16: 451-457.

[18]Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes. Nucleic Acids Res. 2001 Mar 1;29(5):1216-21.

[19]Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJA, Hofmann K, Bairoch A: The PROSITE database, its status in 2002. Nucleic Acids Res. 2002, 30:235-238.

[20]Fields S, Song O: A novel genetic system to detect protein-protein interactions. Nature 1989, 340:245-246.

[21]Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL: Multiplexed biochemical assays with biological chips. Nature. 1993, 364(6437):555-6.

[22]Friedman N, Linial M, Nachman I, Pe'er D: Using Bayesian networks to analyze expression data. J Comput Biol. 2000, 7(3-4):601-20.

[23]Galperin MY, Koonin EV: Who's your neighbor? New computational approaches for functional genomics. Nature Biotechnology 2000, 18:609-613.

[24]Gavin AC, Bosche M, Krause R *et. al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002, 415:141-7.

[25]Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 2000, 25: 25-29.

[26]Gerstein M: Measurement of the effectiveness of transitive sequence comparison, through a third intermediate' sequence. Bioinformatics 1998, 14:707-714.

[27]Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: Co-evolution of proteins with their interaction partners. J Mol Biol. 2000, 299:283-293.

[28]Goh CS, Cohen FE: Co-evolutionary analysis reveals insights into protein-protein interactions. J Mol Biol. 2002, 324(1):177-92.

[29]Gonnet G.H., Cohen M.A., Benner S.A.: Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix."; Biochem. Biophys. Res. Commun. 1994, 199:489-496.

[30]Gough J, Karplus K, Hughey R, Chothia C: Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure. J. Mol. Biol. 2002, 313(4):903-919.

[31]Haft DH, Selengut JD, White O: The TIGRFAMs database of protein families. Nucleic Acids Res. 2003, 31:371-373.

[32]Henikoff S and Henikoff JG, Amino acid substitution matrices from protein blocks. Proc. Natl Acad. Sci. USA 1992, 89:10915–10919.

[33] Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994, 22:4673-4680.

[34] Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T: Assessment of prediction accuracy of protein function from protein—protein interaction data. Yeast. 2001, 18(6):523-31.

[35] Hughes JD, Estep PW, Tavazoie S, Church GM: Computational Identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. Journal of Molecular biology 2000, 296:1205-1214.

[36] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R et. al.: Functional discovery via a compendium of expression profiles. Cell 2000, 102:109-126.

[37] Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM,Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS: Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol. 2001, 19(4):342-7.

[38] Huynen M, Snel B, Lathe W, Bork P: Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. Genome Res. 2000, 10:1204-1210.

[39] Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. 1994, 235:1501-1531.

[40] Lee TI, Rinaldi NJ, Robert F *et. al.:* Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 2002, 298:799-804.

[41] Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res. 2002, 30:242-244.

[42] Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: Detecting Protein function and protein-protein interactions from genome sequences. Science 1999, 285:751-753.

[43] Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: A combined algorithm for genome-wide prediction of protein function. Nature 1999, 402:83-86.

[44] Marcotte CJV and Marcotte EM. Predicting functional linkages from gene fusions with confidence Applied Bioinformatics 2002: 1(2)93-100

[45] Mellor JC, Yanai I, Clodfelter KH, Mintseris J, Delisi C. Predictome: a database of putative functional links between proteins. Nucleic Acids Res 2002, 30:306-9.

[46] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJA, Vaughan R, Zdobnov EM: The InterPro Database, 2003 brings increased coverage and new features. Nucl. Acids. Res. 2003, 31:315-318.

[47] Needleman SB, Wunsch CD: A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol Bio. 1970, 48:443-453.

[48]Ogawa N, DeRisi J, Brown PO. New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. Mol Biol Cell. 2000 Dec; 11(12): 4309-4321.

[49]Overbeek R. Fonstein M, D'Souza M, Pusch GD, Maltsev N: The use of gene clusters to infer functional coupling. Proc. Nat. Acad. Sci. USA 1999, 96:2896-2901.

[50]Park J, Teichmann SA, Hubbard T, Chothia C: Intermediate sequences increase the detection of homology between sequences. J Mol Biol. 1997, 273:349-354.

[51]Pavlidis P, Grundy WN: Combining microarray expression data and phylogenetic profiles to learn gene functional categories using support vector machines. Columbia University Computer Science Department Technical Report 2000, CUCS-011-00.

[52]Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Nat. Acad. Sci. USA 1999, 96:4285-4288.

[53]Pellegrini M, Thompson M, Fierro J, Bowers P. Computational method to assign microbial genes to pathways. J Cell Biochem 2001, Suppl 37:106-9.

[54]Pelletier JN, Campbell-Valois FX, Michnick SW: Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. Proc Natl Acad Sci USA 1998, 95:12141-6.

[55]Ramani AK, Marcotte EM: Exploiting the co-evolution of interacting proteins to discover interaction specificity. J Mol Biol. 2003, 327(1):273-84.

[56]Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Blattner FR, Collado-Vides J: RegulonDB (version 3.0): transcriptional regulation and operon organization in Escericia coli K-12. Nucleic acids Research 2000, 28:65-67.

[57]Schena M, Shalon D, Davis R, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995, 270:467-470.

[58]Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol. 2002, 20(4):370-5.

[59]Schwartz R.M., Dayhoff M.O.: Matrices for detecting distant relationships. (In) Atlas of Protein Sequence and Structure, 5 suppl. 3:353-358 (1978), Nat. Biomed. Res. Found., Washington D.C.

[60]Schwikowski B, Uetz P, Fields S: A network of protein-protein interactions in yeast. Nat Biotechnol. 2000, 18(12):1257-61.

[61]Selkov E, Overbeek R, Kogan Y, Chu L, Vonstein V, Holmes D, Silver S, Haselkorn R, Fonstein M: Functional analysis of gapped microbial genomes: amino acid metabolism of Thiobacillus ferooxidans. Proc. Nat. Acad. Sci. USA 2000, 97:3509-3514.

[62]Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M,Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA,Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002, 8(1):68-74.

[63]Smith TF, Waterman MS: Identification of common molecular subsequences. Journal of Molecular Biology 1981, 147:195-197.

[64]Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the

yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998, 9:3273-3297.

[65] Tamames J, Casari G, Ouzounis C, Valencia A: Conserved clusters of functionally related genes in two bacterial genomes. J Mol Evol. 1997, 44:66-73.

[66] Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. Science 1997, 278:631-637.

[67] Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 2000, 28:33-36.

[68] Teichman SA, Mitchison G: Computing protein function. Nature Biotechnology 2000, 18:27.

[69] Thompson W, Rouchka EC, Lawrence CE. Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res. 2003 Jul 1;31(13):3580-5.

[70] Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y et. al.: E-Cell: software environment for whole-cell simulation. Bioinformatics 1999, 15:72-84.

[71] Uetz P, Giot L, Cagney G *et al*:. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 2000, 403:623-7.

[72] Vazquez A, Flammini A, Maritan A, Vespignani A: Global protein function prediction from protein-protein interaction networks. Nat Biotechnol. 2003, 21(6):697-700.

[73] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. Science 1995, 270:484-487.

[74] von Dassow G, Meir E, Munro EM, Odell GM: The segment polarity network is a robust developmental module. Nature 2000, 406:188-92.

[75] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 2003, 31(1):258-61.

[76] Wilson M, DeRisi J, Kristensen HH, Imboden P, Rane S, Brown PO, Schoolnik GK: Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. Proc Natl Acad Sci U S A. 1999, 96:12833-12838.

[77] Wu CH, Huang H, Yeh L, Barker WC: Protein family classification and functional annotation. Comput Biol Chem. 2003, 27:37-47.

[78] Wu J, Kasif S, DeLisi C.: Identification of functional links between genes using phylogenetic profiles. Bioinformatics. 2003, 19(12):1524-30.

[79] Xenarios I, Salwinski L, Duan XJ *et. al.*: DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002, 30:303-5.

[80] Zhu H, Bilgin M, Bangham R *et. al.*: Global analysis of protein activities using proteome chips. Science 2001, 293:2101-5.

[81] Zhu J, Liu J, Lawrence CE: Bayesian adaptive sequence alignment algorithms. Bioinformatics 1998, 14:25-39.