Computational methods for protein function analysis
Matteo Pellegrini

Two recent advances have had the greatest impact on protein function analysis so far: the complete sequences of genomes and mRNA expression level profiles. The former has spurred the development of novel techniques to study protein function: phylogenetic profiles and gene clusters. The latter has introduced a method, not based on sequence homology, that enables one to group together functionally related genes.

Addresses
Protein Pathways, 1145 Gayley Avenue, Suite 304, Los Angeles, CA 90024, USA; e-mail: matteope@proteinpathways.com

Current Opinion in Chemical Biology 2001, 5:46–50
1367-5931/01/$ — see front matter © 2001 Elsevier Science Ltd. All rights reserved.

Introduction
Computational methods used to analyze protein function can be divided into three broad categories: alignment, genome and expression methods. Alignment methods rely directly on the similarity of amino acid sequences between proteins. These methods are by far the most developed in the field of bioinformatics, but the bulk of the development has occurred over the past two decades. Recent innovations in alignment methods have had less impact than the introduction of genome and expression methods over the past two years. Nevertheless, a few alignment-method innovations will be included in this review, such as indirect homologies, graph-based analysis and Bayesian alignments.

Genome-based methods exploit the information contained within the full sequence of an organism’s genome. As this review is being written, approximately sixty complete genomes are available for analysis. The bulk of these come from bacteria and archaeabacteria. The first eukaryotic genomes, from yeast, fruit fly and Caenorhabditis elegans, are also available. The human genome will also soon be completed. The methods covered in this review include 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 third category of computational methods utilizes the information from mRNA profiling experiments. mRNA concentrations of each expressed gene within a cell may be measured by direct sequencing or by spotting arrays of genes and measuring hybridization through fluorescence. Typically, these experiments may be repeated multiple times with only small modifications to the cell, to see how expression levels are perturbed by external stimuli. Once this data is collected, it is usually clustered into sets of genes with similar expression levels across multiple experiments. These clustered genes often share some common functional properties, and so this technique may be used to infer the function of the clustered genes. However, these techniques are likely to be more useful in elucidating transcriptional regulation within a cell, rather than protein function.

Alignment methods
Methods that align the sequences of amino acids in proteins have been developed over the past few decades [1,2]. The most commonly used method to date is probably **BLAST** [3]. This allows users to rapidly search for homologous sequences in large protein databases. Searching a protein sequence against the full protein-sequence database requires less than a minute on a typical computer.

More sensitive and accurate alignment protocols have also been developed. For instance, hidden Markov models for protein families are able to detect remote homologs that may be missed by simple sequence alignment techniques [4]. These models have been created for most protein families and are compiled within the PFAM database [5]. More recently, Bayesian statistics have been applied to rigorously compute optimal alignments [6]. In general, however, these approaches are extremely computationally intensive, and thus not applicable to all homology searches.

A newer approach is to extract additional information from the rapid alignments produced by BLAST. For instance, one common goal is to extend alignments to recognize distant homologs. This has been accomplished, in part, by using transitive sequence comparisons [7,8]. Each sequence may be homologous to several hundred 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 that could not be found by conventional alignment techniques.

**Figure 1**

A ‘Rosetta Stone’ protein is shown schematically as the fusion of two independent proteins (protein 1 and protein 2).
In a similar manner, BLAST alignments have been used to search for protein fusions [9••,10]. In this technique, one seeks two proteins, which are not homologous to each other, that align to different regions of a third protein. In other words, these 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 fused protein are interacting or functionally related (see Figure 1).

Graph analysis applied to databases of alignments enables one to rapidly cluster protein families and decompose proteins into their respective domains [11•]. In the future, these techniques will probably be used to rapidly annotate the growing populations of proteins.

Phylogenetic analysis can sometimes be used to deduce the function of a protein in more detail [12]. The phylogenetic tree of a protein family may reveal that a single family in fact clusters into several subfamilies. The members of different subfamilies often have distinct functions within a cell. Phylogenetic trees have also been utilized creatively to match proteins with their interaction partners [13•], assuming that the two must have co-evolved.

**Genomic methods**

Over the past few years, the number of fully sequenced genomes has grown dramatically. 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 orthologous groups [14,15]. This graph-based analysis constructs highly connected sets of orthologous proteins, which are then classified according to their function.

From the analysis of orthologs across genomes it is also possible to construct phylogenetic profiles [16••]. These
Expression methods

As a result of remarkable developments during the past few years, it is now possible to measure the concentrations of every mRNA within a cell. There are two primary ways to accomplish this. The first is by sequencing the mRNA and counting the number of copies of a particular gene. The second is by hybridizing the mRNA to complementary sequences that are attached to a chip, and then estimating the concentrations by fluorescence.

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. The result is that each gene has an associated expression vector that measures its concentrations in the cell during a time series. These vectors may then be clustered. It was found that genes that cluster together are likely to have similar functions. Therefore, gene expression clustering may be used similarly to phylogenetic profiles to assign approximate functions to all the genes within a cell by assuming that co-clustered genes often have similar functions.

Instead of measuring varying mRNA concentrations during cellular cycles, it is also possible to induce expression changes by adding specific compounds to a cell, or knocking out a gene. This method probes the response of specific genes to external stimuli and more directly infers that the function of the varying genes relates to the stimulus.

The above discussion has concentrated on yeast, but similar methods may also be effectively applied to multicellular organisms. In this case, one may ask whether two genes are expressed within the same tissues of a certain organism or if they are neighbors on multiple genomes, it is likely that they are members of a conserved operon. Once again, one is able to use this information to group together genes that are likely to be part of a single pathway or complex and thus extract functional information for genes that have not been characterized. It has been shown that such an analysis, when combined with conventional homology-based methods, yields functional information on the vast majority of genes encoded in newly sequenced genomes.

As a result of these and related techniques, it has become possible to precisely describe the operon organization of Escherichia coli. In the near future, this in-depth transcriptional knowledge is likely to be deciphered for all fully sequenced microbes.

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.

One recent approach to combining these methods treats every prediction as a link between two proteins. 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. Using support vector machines, the authors show that these combined data structures are able to recover functional information for a greater number of genes than any one of the methods alone.
Conclusions
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, which also allows us to gain insights on protein function.

In the near future, these insights will be used by computational biologists to model cellular pathways in great detail [39]. It is already possible to begin to model developmental pathways [40] and metabolic pathways [41•], 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 cellular biology.

Acknowledgements
The author would like to thank Jennifer Kelly, Michael Thompson and Ken Goodwill for their comments on the manuscript.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

* of special interest
** of outstanding interest


9. Marcotte EM, Pellegrini M, Ng HL, Yeates TO, Eisenberg D: ** Detecting protein function and protein–protein interactions from genome sequences. Science 1998, 285:751-753. Protein fusions are used to gain functional information about proteins with no homology to characterized proteins.


15. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999, 96:4285-4288. The presence or absence of genes across multiple genomes is used as a cri‐ terion for clustering proteins. The members of these clusters tend to share similar functions.


18. Overbeek R, Fonstein M, D’Souza M, Pusch GD, Maltsev N: The use • of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 1999, 96:2896-2901. Genes that are found near on multiple genomes are likely to be part of conserved operons. This method generalizes the concept to search for genes that are likely to participate in common pathways.


30. Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 2000, 296:1205-1214. Expression and function data is used to cluster together co-regulated genes. The authors then proceed to uncover the DNA motifs to which transcription factors bind.


