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

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 Baysian 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 archaebacteria. 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.

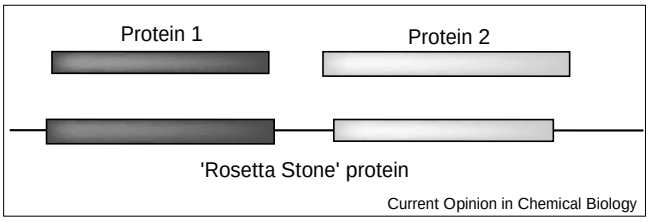
Alignment methods

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.

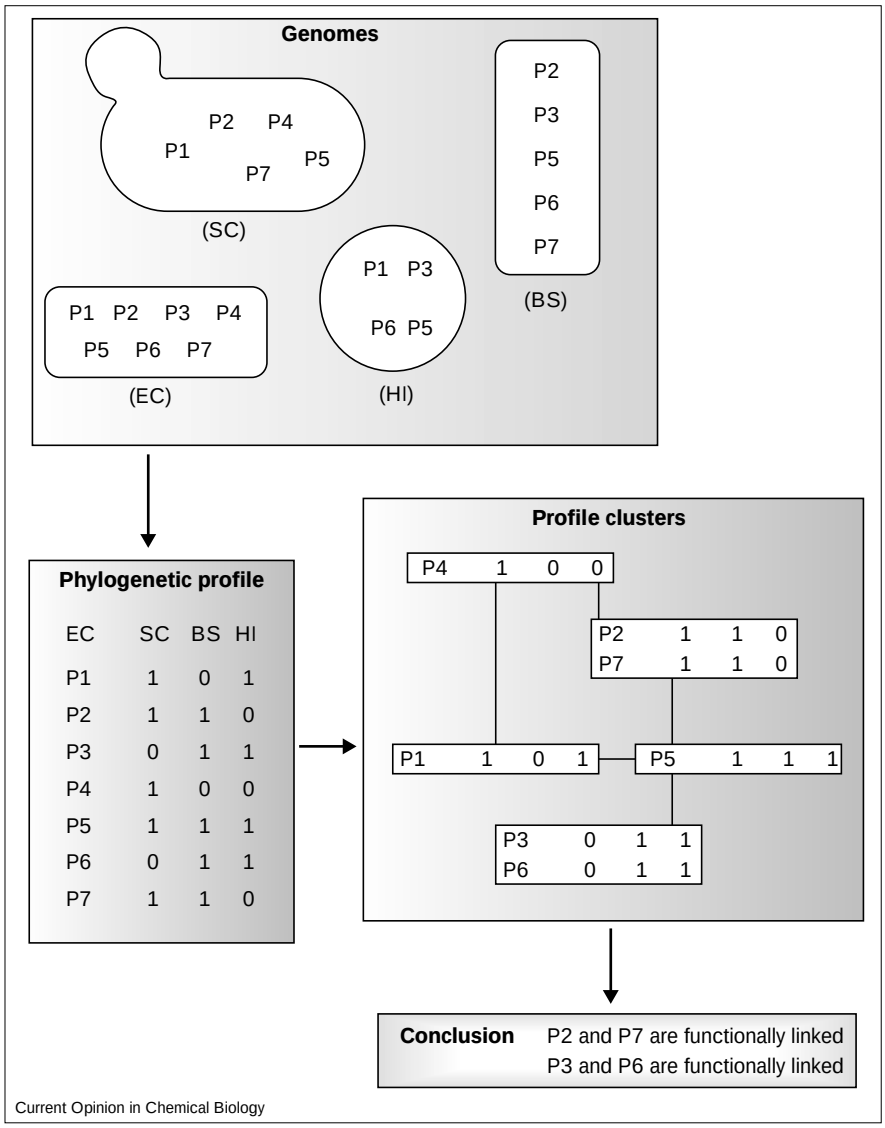
Figure 1



A 'Rosetta Stone' protein is shown schematically as the fusion of two independent proteins (protein 1 and protein 2).

Figure 2

The method of protein phylogenetic profiles illustrated schematically for the hypothetical case of four fully sequenced genomes, in which we focus on seven proteins (P1–P7). For each *E. coli* protein, a profile is constructed, indicating which genomes code for homologs of the protein. The profiles are next clustered to determine which proteins share the same profiles. Proteins with identical (or similar) profiles are boxed to indicate that they are likely to be functionally linked. BS, *Bacillus subtilis*; EC, *Escherichia coli*; HI, *Haemophilus influenzae*; SC, *Saccharomyces cerevisiae*.



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.

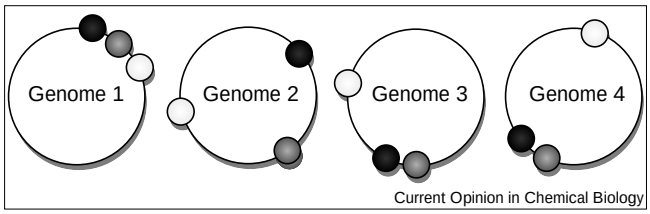
Phylogenetic analysis can sometimes be used to deduce the function of a protein in more detail [12]. The phylogenetic tree of a protein family may reveal that a single family in fact clusters into several subfamilies. The members of different subfamilies often have distinct functions within a cell. Phylogenetic 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

Figure 3



A schematic of the position of three genes (shown as black, dark gray and light gray circles) and their homologs across four genomes. The figure illustrates that the genes shown in black and dark gray are found nearby on multiple genomes and therefore are likely to have related functions. In contrast, the position of the light gray gene does not correlate strongly with that of the black and dark gray genes, and therefore probably has an unrelated function.

are typically binary arrays constructed for each protein that encode whether a homolog of the protein is present in any of the fully sequenced genomes (see Figure 2). It has been shown that proteins with similar phylogenetic profiles are often members of the same protein complex or cellular pathway. Thus, these profiles may be used to assign approximate functions to proteins that are not homologous to characterized proteins.

A similar approach utilizes the positions of genes on the genome [17,18,19••]. If two genes are found nearby in multiple genomes, it is likely that they are members of a conserved operon (see Figure 3). 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 [20].

As a result of these and related techniques, it has become possible to precisely describe the operon organization of *Escherichia coli* [21]. In the near future, this in-depth transcriptional knowledge is likely to be deciphered for all fully sequenced microbes.

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 [22,23] 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 [24].

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 [25,26]. 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 [27•,28]. 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 [29,30]. Co-occurring genes are likely to be functionally related, or possibly interacting. In one study, genes associated with prostate cancer were grouped using this technique ([29]; Figure 4).

Clearly, the ability to measure all gene expression levels within a cell offers biologists an entirely new approach to studying the cellular functions of proteins without relying on sequence homology. These methods more directly probe transcriptional regulation, however, and may therefore be used to reconstruct transcriptional networks. In particular, one may use this data to reconstruct the common upstream DNA motifs that effect transcriptional regulation among co-expressed genes [31•].

Combined methods

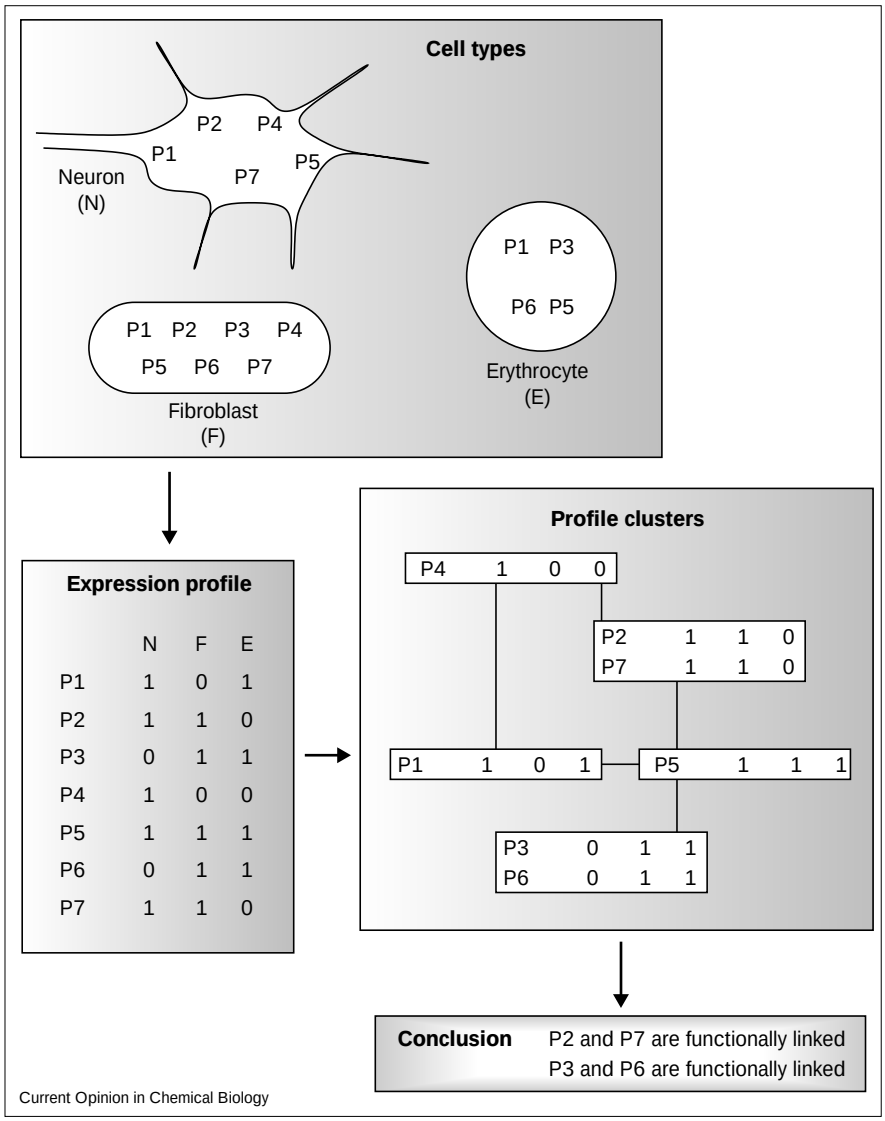
The methods described above exploit different properties of proteins to gain functional insights. Often, these properties generate information on different sets of proteins. It is therefore useful to combine these methods to gain a more complete picture of protein function [32–36].

One recent approach to combining these methods treats every prediction as a link between two proteins [37•]. 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 [38]. 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.

Figure 4

Different cell types express different sets of genes. By clustering together genes that are always expressed together in multiple cells we can often infer a functional coupling between the encoded proteins. The process of deducing functional links between proteins is shown schematically for three different cell types and seven distinct proteins.



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

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

2. Smith TF, Waterman MS: Identification of common molecular subsequences. *J Mol Biol* 1981, 147:195-197.

3. 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 Res* 1999, 25:3389-3402.

4. 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.

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

6. Zhu J, Liu J, Lawrence CE: Bayesian adaptive sequence alignment algorithms. *Bioinformatics* 1998, 14:25-39.

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

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

9. •• 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.
Protein fusions are used to gain functional information about proteins with no homology to characterized proteins.

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

11. • Enright AJ, Ozounis CA: GeneRage: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* 2000, 16:451-457.
Graph-based analysis is used to cluster together protein families and to decompose proteins into their respective domains.

12. Eisen JA: Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* 1998, 8:163-167.

13. • 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.
Interacting pairs of proteins may be discovered by searching for complementary phylogenetic trees.

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

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

16. •• 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 criterion for clustering proteins. The members of these clusters tend to share similar functions.

17. 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.

18. 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.

19. •• 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 nearby 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.

20. 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 Natl Acad Sci USA* 2000, 97:3509-3514.

21. 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 ***Escherichia coli*** K-12. *Nucleic Acids Res* 2000, 28:65-67.

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

23. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M *et al.*: Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 2000, 18:630-634.

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

25. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998, 95:14863-14868.

26. 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.

27. • Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD *et al.*: Functional discovery via a compendium of expression profiles. *Cell* 2000, 102:109-126.
The expression levels of yeast genes are measured in response to gene knockouts and drugs. Clustering the genes according to their expression levels across multiple experiments reveals which genes may potentially be interacting with the drugs.

28. 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 USA* 1999, 96:12833-12838.

29. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T: Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res* 1999, 9:1198-1203.

30. Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie JM: Large-scale statistical analysis of rice ESTs reveal correlated patterns of gene expression. *Genome Res* 1999, 9:950-959.

31. • 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.

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

33. Galperin MY, Koonin EV: Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol* 2000, 18:609-613.

34. Teichman SA, Mitchison G: Computing protein function. *Nat Biotechnol* 2000, 18:27.

35. 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.

36. Aravind L: Guilt by association: contextual information in genome analysis. *Genome Res* 2000, 10:1074-1077.

37. • 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.
Phylogenetic profiles, Rosetta Stone and expression links are combined to uncover the function of uncharacterized genes.

38. Pavlidis P, Grundy WN: Combining microarray expression data and phylogenetic profiles to learn gene functional categories using support vector machines. In *Columbia University Computer Science Department Technical Report* 2000, CUCS-011-00:1-11.

39. Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC *et al.*: E-Cell: software environment for whole-cell simulation. *Bioinformatics* 1999, 15:72-84.

40. von Dassow G, Meir E, Munro EM, Odell GM: The segment polarity network is a robust developmental module. *Nature* 2000, 406:188-192.

41. • Edwards JS, Palsson BO: The ***Escherichia coli*** MG1655 ***in silico*** metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA* 2000, 97:5528-5533.
The conservation of metabolites across multiple metabolic reactions allows the authors to model metabolism in *E. coli* and predict which knock-outs inhibit growth.