Bioinformatics in the post-sequence era

Minoru Kanehisa[1] & Peer Bork[2]

doi:10.1038/ng1109

In the past decade, bioinformatics has become an integral part of research and development in the biomedical sciences. Bioinformatics now has an essential role both in deciphering genomic, transcriptomic and proteomic data generated by high-throughput experimental technologies and in organizing information gathered from traditional biology. Sequence-based methods of analyzing individual genes or proteins have been elaborated and expanded, and methods have been developed for analyzing large numbers of genes or proteins simultaneously, such as in the identification of clusters of related genes and networks of interacting proteins. With the complete genome sequences for an increasing number of organisms at hand, bioinformatics is beginning to provide both conceptual bases and practical methods for detecting systemic functional behaviors of the cell and the organism.

The exponential growth in molecular sequence data started in the early 1980s when methods for DNA sequencing became widely available. The data were accumulated in databases such as GenBank, EMBL (European Molecular Biology Laboratory nucleotide sequence database), DDBJ (DNA Data Bank of Japan), PIR (Protein Information Resource) and SWISS-PROT, and computational methods were developed for data retrieval and analysis, including algorithms for sequence similarity searches, structural predictions and functional predictions. Such activities of computational biology—or 'bioinformatics' as it is now called—were already apparent in the 1980s, although they mainly involved DNA and protein sequence analysis and, to a small extent, the analysis of three-dimensional (3D) protein structure.

The impact of the genome projects of the past 10 years is not simply an increased amount of sequence data, but the diversification of molecular biology data. A genome sequence presents not only a complete set of genes and their precise locations in the chromosome, but also gene similarity relationships within the genome and across species. Automatic sequencing has had an enormous impact as it has been at the forefront of the high-throughput generation of various biological data—expressed-sequence tags (ESTs) and single-nucleotide polymorphisms (SNPs) among others. Experimental technologies have been developed, notably DNA microarrays for systematically analyzing gene expression profiles and yeast two-hybrid systems and mass spectroscopy for detecting protein-protein interactions. Initiatives on structural genomics are not in the large-scale production phase as yet, but they will certainly result in an increased amount of protein 3D structural data.

In addition to the continual development of experimental technologies for accumulating divergent molecular biology data, the past decade saw developments in informatics technologies. The single most important event was the arrival of the Internet, which has transformed databases and access to data, publications and other aspects of information infrastructure. The Internet has become so commonplace that it is hard to imagine that we were living in a world without it only 10 years ago. The rise of bioinformatics has been largely due to the diverse range of large-scale data that require sophisticated methods for handling and analysis, but it is also due to the Internet, which has made both user access and software development far easier than before.

There is no doubt that the lists of 'molecular parts' obtained by systematic experiments and revealed using bioinformatics tools have helped to advance genetics and other biological sciences. As compared with physics or chemistry, however, biology is still an immature science in the sense that we cannot make predictions based on general principles. This situation is bound to change through our expanding body of large-scale data and accumulated knowledge and through bioinformatics, which is gradually altering course to become a more fundamental discipline. The ultimate goals of bioinformatics will be to abstract knowledge and principles from large-scale data, to present a complete representation of the cell and the organism, and to predict computationally systems of higher complexity, such as the interaction networks in cellular processes and the phenotypes of whole organisms. From this perspective, as illustrated in Fig. 1, we present a chronological account of the rise of bioinformatics.

From sequence to high-throughput data analysis

The first big breakthrough in the past decade was the introduction of the rapid sequence database search tool BLAST[1]. This search tool was not only more efficient than FASTA[2], which had been developed in the 1980s, but also based on different principles. A database search involves pair-wise comparison of the query sequence against each sequence contained in the database.

[1]*Bioinformatics Center, Kyoto University, Uji, Kyoto 611-0011, Japan.* [2]*European Molecular Biology Laboratory, 69012 Heidelberg, Germany. Correspondence should be addressed to M.K. (e-mail: kanehisa@kuicr.kyoto-u.ac.jp).*

Traditionally, this comparison was treated as an optimization query that searches for the optimal sequence alignment maximizing either the number of matched letters or the similarity score using an amino acid mutation matrix. When gaps are allowed, there are an enormous number of possibilities for aligning two sequences; however, the optimal alignment can be found rigorously by the 'dynamic programming' algorithm[3], which systematically 'prunes' the branches of the search tree containing all possible alignments. Unfortunately, this algorithm requires much computation time and is not readily applicable to large-scale databases. Thus, the strategy taken in FASTA is first to do a rapid, rough search for matched areas using a data structure called 'hash' and then to apply the dynamic programming algorithm in the neighborhood of those areas.

Whereas FASTA follows the tradition of combinatorial optimization, BLAST is based on mathematical statistics coupled with human intuition. For example, if we humans were to compare two sequences by eye, we would not dare to examine all possible alignments; rather, we would look for common patterns shared by two sequences and try to extend these to obtain longer matches, because we know that related sequences tend to contain conserved sequence motifs. This is the strategy taken in BLAST, which incorporates a sound mathematical foundation to calculate the statistics of high-scoring segment pairs (HSPs)—ungapped local alignments whose scores cannot be improved by extension or trimming. The probability of finding an HSP with score S is known to follow an extreme value distribution, and this probability or so-called 'E value' can be estimated for a given combination of the query sequence, the database to be searched and the scoring system. The E value is now used widely as a standardized measure to estimate statistical significance of sequence similarity, in other words, how often one would expect to observe a particular database hit by chance alone.

At about the same time that BLAST was developed, researchers started to collect a different type of data—the gene-based sequence-tagged sites[4] or ESTs[5]—that would subsequently influence the nature of sequence databases. The mass collection of data containing single-pass (low-quality and fragment) sequences is a quick way to capture a complete repertoire of genes expressed in specific cells or tissues. In this approach, BLAST is the method of choice to search for similarities against existing databases and to do all-against-all comparisons within the data set for identifying clusters of similar sequences.

The mid-1990s saw the collection of another, qualitatively different type of mass sequence data—the whole genomes of cellular organisms. Starting with the small bacterium *Haemophilus influenza*[6], progressing to yeast[7] and the more-complex multicellular organisms *Caenorhabditis elegans*[8] and *Drosophila melanogaster*[9] and ultimately, the draft sequence of the human genome[10,11]. The completely sequenced genomes of more than 100 organisms are now available and many more are in progress, including a finished version of the human genome.

An increase in the amount of large-scale sequence data does not necessarily lead to an increase in biological knowledge unless it is accompanied with new or improved tools for sequence analysis[12]. Elaborate methods have been developed to increase the sensitivity of sequence similarity searches by a factor of three[13], the most successful ones being PSI-BLAST[14] and hidden Markov models (HMMs)[15]. PSI-BLAST is an extremely sensitive method to detect weak similarities and is based on an iterative procedure that makes successive improvements to the position-specific scoring matrix (or 'profile') initially generated by a standard BLAST search.

Similar to profiles, HMMs are constructed from multiple sequence alignments, such as those produced by ClustalW and X[16,17], but they explicitly incorporate the probabilities of insertions and deletions and enable a search against an HMM library to detect subtle sequence features. Other successful methods for sequence analysis have been based on neural networks, resulting in considerable improvement in, for example, protein secondary structure predictions[18], and on rule-based systems that predict various functional features of proteins, for example, the prediction of protein localization by PSORT (ref. 19; a comprehensive survey of prediction accuracy in different areas of sequence analysis is given in ref. 20).

Biological knowledge obtained from the analysis of primary databases containing sequences and 3D structures was stored in secondary databases for further reference purposes, and this trend has continued (Table 1). In particular, HMMs and PSI-BLAST facilitated the development of protein domain databases[21–27] that allow identification of the modular architecture of proteins and their functional units. The similarity search against primary sequence databases can be used for functional prediction of a gene or a protein as long as the database is well annotated and curated, although it is increasingly more difficult to maintain an up-to-date, well-annotated sequence database owing to the varying quality and ever-increasing amounts of sequence data. Thus, it has become customary to rely more on the secondary databases, which contain signatures of protein domains and functional sites—rather like dictionaries containing knowledge on the 'words' and 'sentences' of the 'sequence language'.

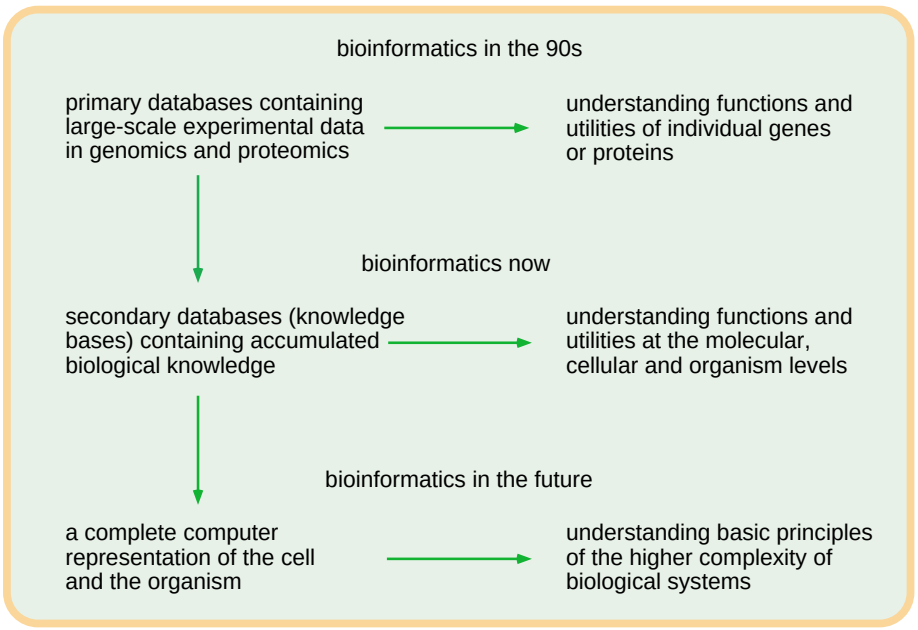

Fig. 1 A view of bioinformatics past, present and future.

Table 1 • Databases of organized biological knowledge

Knowledge	Database	URL	Refs.
protein functional sites	PROSITE	http://www.expasy.ch/prosite/	21
	BLOCKS	http://www.blocks.fhcrc.org/	22
	PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/	23
	ProDom	http://prodes.toulouse.inra.fr/prodom/doc/prodom.html	24
	Pfam	http://pfam.wustl.edu/	25
	SMART	http://smart.embl-heidelberg.de/	26
	TIGRFAMs	http://www.tigr.org/TIGRFAMs/	27
protein 3D folds	SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	52
	CATH	http://www.biochem.ucl.ac.uk/bsm/cath_new/	53
transcription factors	TRANSFAC	http://transfac.gbf.de/TRANSFAC/	54
protein interactions	BIND	http://www.bind.ca/	55
	DIP	http://dip.doe-mbi.ucla.edu/	56
protein pathways	KEGG	http://www.genome.ad.jp/kegg/	42
	EcoCyc	http://www.ecocyc.org/	43
ortholog groups	COG	http://www.ncbi.nlm.nih.gov/COG/	30
controlled vocabulary	GO	http://www.geneontology.org/	41

With the increasing number of complete genomes available for comparative studies, different types of function prediction concepts have been developed, notably 'gene context' and 'gene content' analyses (reviewed in refs. 28,29). If the genome is viewed as a string of genes, gene context represents the positional association of genes, such as an operon in prokaryotic genomes. Gene context analysis, involving the comparison of gene orders and gene fusions across different genomes, can detect the functional association of proteins, such as physically interacting subunits, members of the same pathway and an enzyme and its regulator.

By contrast, gene content analysis is a comparison of gene repertoires across different genomes. When two genes are present or absent in a correlated manner among many organisms, there may be a functional link between these genes. A prerequisite to such an analysis is to establish orthologous relationships—that is, functionally identical genes that have descended from a common ancestor. In practice, orthologs are defined by sequence similarities, often by the criterion of bidirectional best hits in pair-wise genome comparisons. The COG database[30] was one of the early and most prominent efforts to organize knowledge about orthologous groups among completely sequenced genomes.

From molecular to higher order function

The availability of different types of high-throughput experimental data in the late 1990s has expanded the role of bioinformatics and facilitated the analysis of higher order functions involving various cellular processes. Notably, an oligonucleotide array[31] or a cDNA microarray[32] containing every gene in the genome is a powerful tool to measure gene expression across a whole cell or tissue under different conditions. In addition to sequence similarity and proximity on the chromosome, two genes can now be related by their similarities in expression profiles, either at specific time points or in other controlled conditions. Clusters of co-regulated genes can be detected from gene expression data—a process that is similar in essence to detecting clusters of orthologous genes in COG or clusters of positionally correlated genes in gene context analysis—and these expression clusters can identify potential members of gene groups responsible for specific physiological processes. Complex gene expression data also have stimulated applications of informatics technologies, including self-organizing maps[33] and support vector machines[34], for the extraction of inherent biological features.

Protein-protein interactions represent another type of experimental data. High-throughput two-hybrid system analyses have been performed for whole sets of protein pairs encoded in the yeast genome[35,36], and components in purified protein complexes in yeast have been identified systematically by mass spectroscopy[37,38]. These data sets add an additional layer of information about proteins (protein interactions) to the existing genome (sequence similarity and gene context) and transcriptome (expression similarity) data sets. All of these data sets can be treated as collections of binary relationships—that is, relationships between two objects—which allow an integrated analysis that can extract biological features more accurately. When different data sets in yeast were combined, commonly found pairs were likely to be biologically more meaningful[39,40]. This implies that data on higher order function usually have higher error rates, that the interpretation can contain many pitfalls, and that rigorous benchmarking is therefore required.

Until recently, there was no common terminology for the different aspects of function. The first steps towards a common vocabulary for protein function have been taken by the Gene Ontology Consortium[41] so that functional features can be compared and described better. The Gene Ontology Consortium categorizes currently accumulated and dynamically changing knowledge into three systematic terminologies or 'ontologies': the 'molecular function' of an individual protein; the 'biological process' in which a protein is involved; and the 'cellular component' in which a protein functions.

To increase our understanding of cellular processes from genome information, pathway database, for example KEGG[42] and EcoCyc[43], have been created in the past decade (Table 1). Whereas most databases concentrate on molecular properties (for example, sequences, 3D structures, motifs and gene expressions), these databases tackle complex cellular properties, such as metabolism, signal transduction and cell cycle, by storing the corresponding networks of interacting molecules in computerized forms, often as graphical pathway diagrams. Inevitably, it is necessary to collect data and knowledge from published literature accumulated over many years from traditional studies of biology. At least for metabolic pathways, the past knowledge is relatively well organized in these databases, providing a reference data set for annotating genomes (the 'metabolic reconstruction') and for screening microarray and other high-throughput experimental data.

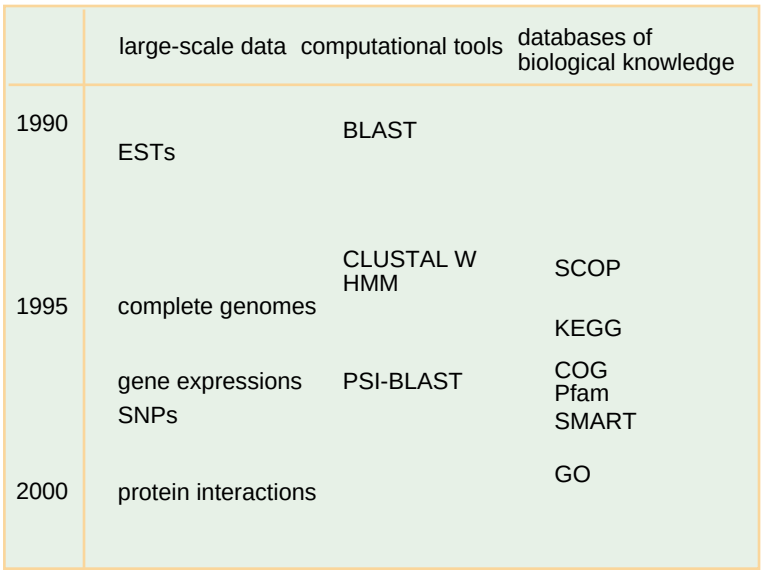

Fig. 2 Bioinformatics developments of the past decade. Shown are principal landmarks in the generation of large-scale data by high-throughput experiments, the development of computational tools for data analysis, and the creation of databases of biological knowledge.

In contrast to the sequence, which is a simple one-dimensional object, the network of interacting molecules is represented as a complex graph object. Mathematically, a graph is a set of nodes and edges and, depending on what is to be taken as a node, different types of graphic objects can be defined. For example, the protein sequence is a graph object consisting of amino acids (nodes) connected by peptide bonds (edges), whereas the protein 3D structure is a graph object consisting of atoms (nodes) and atomic interactions (edges). To understand higher functions, it is necessary to consider 'higher' graph objects: the KEGG database consists of three such objects called 'protein network', 'gene universe' and 'chemical universe', for which the nodes are proteins, genes and chemical compounds, respectively. These databases of higher graph objects have paved the way for developing graph algorithms, such as those for detecting local graph similarities[44] among pathways, expression profiles and genomic contexts.

The concept of the *E* value in the BLAST search is based on the view that the database is a collection of independent objects (sequences). By contrast, the KEGG database or any interaction network database contains graph objects that are collections of nodes (proteins, genes or chemical compounds) that are inherently related by different types of edge. Thus, the statistics of similarities and other features of the graph must be understood and converted to a new *E* value for the network analysis to become more intelligent and efficient—in other words, a FASTA-to-BLAST type of development.

Toward this end, recent analyses on global statistical features of network topology may have relevance. Various interaction networks of complex systems such as the Internet and social networks, as well as metabolic networks, gene regulatory networks and other biological networks, share common properties of network topology. One property is the 'small world' network[45], in which any two nodes can be connected by a few steps because of the intermediate topology between completely regular and completely random graphs. Another property is the 'scale-free' network, in which the node connectivity distribution follows a power law[46], which suggests the existence of highly connected nodes (hubs). In the continually expanding Internet and social networks, these properties have been related to the tendency of new nodes to be linked to larger hubs. In the biological networks, the implications must be pursued in functional and evolutional perspectives. For example, the scale-free property seems to be correlated to the network stability against random errors—a feature that is favored in evolution. Although different types of complex network share global properties, they seem to be distinct when simple structural elements (network motifs) are examined[47].

Certainly, the complexity of network topology arises from complex patterns of connections (interactions) and not simply from the size of the network (measured by the number of nodes). This may have biological implications, especially in view of the surprisingly few genes found in the human genome[10,11]. Graphs and patterns of node connections are static in nature. Predicting network dynamics is far more difficult than simply predicting connection patterns, as has been accomplished in metabolic reconstruction. Here again, by designing high-throughput experiments that systematically perturb dynamic environments and collecting enough experimental data, network dynamics may become computable, at least for dynamic changes in response to small environmental perturbations[48].

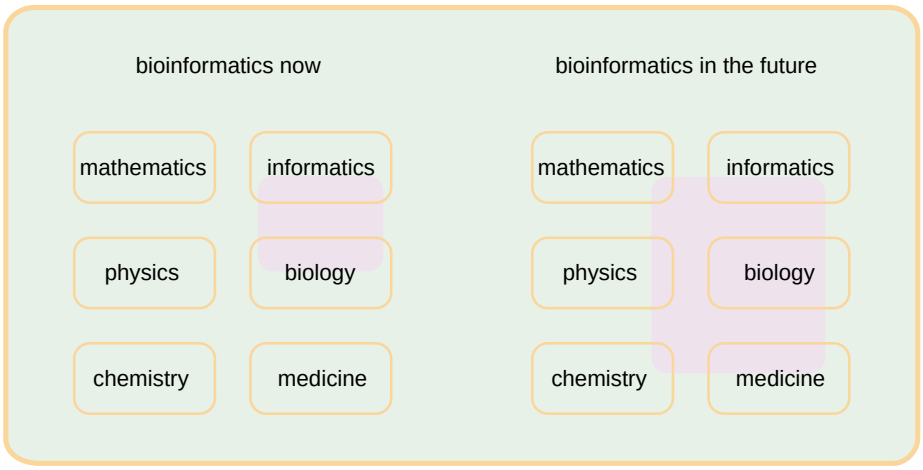

Fig. 3 Bioinformatics now and in the future. Bioinformatics (shaded area) is currently an interdisciplinary field of biology and informatics that focuses on practical applications of informatics technologies to genomics and other areas of high-throughput biology. As data-driven biology is replaced by principle-driven biology in the future, bioinformatics will become a more-fundamental discipline encompassing mathematics, physics and chemistry.

From data-driven to principle-driven biology

In the past decade, bioinformatics was characterized by the development of innovative computational methods to help generate and analyze various large-scale data and by the creation of new databases of biological knowledge as a direct result of the large-scale analyses. (Fig. 2) We consider that this is only the beginning of the path to our ultimate goal of understanding the basic principles underlying the complexity of living cells and organisms (Fig. 1). Enumeration in biology is no longer limited to the lists of molecular parts such as genes (genome), mRNAs (transcriptome), proteins (proteome) and metabolic compounds (metabolome). More extensive lists include the interactome[36], which incorporates sets of protein–protein interactions, and the localizome[49], which describes the subcellular localizations of proteins. The repertoire of different lists will continue to grow as high-throughput experimental methods are further elaborated and expanded.

Of course, on its own the bottom-up approach from large-scale data in genomics and proteomics will not be sufficient for understanding the higher complexity of biological systems. Efforts to computerize our knowledge on cellular functions, at present either by the controlled vocabulary of Gene Ontology or by graph representation in KEGG, will both facilitate the computational mapping of genomic data to complex cellular properties and detect any empirical relationships between genomic and higher properties. Although the field is already looking forward to a 'systems biology' approach and to simulations of whole cells, much of the effort must be devoted to capturing even higher properties, such as ontology for human diseases and the computable representation of cellular networks. In addition, the dependence of functionality on the context (such as experimental conditions, cell status and environment) is currently mostly ignored; in other words, several other levels of complexity will have to be considered before we can come to a more basic understanding of life as a series of complex information systems[50].

There are already several molecular biology databases available on the Internet[51], and many more will evolve as genomic data and computational methods are introduced into individual research in biology. Although Internet-based linking among different databases is a convenient way to use this huge resource, more effort is required for the true integration of biological knowledge and data. Integration in this respect does not simply involve methodology, such as links and common interfaces, but rather it involves biology. The ultimate integration of biological databases will be a computer representation of living cells and organisms, whereby any aspect of biology can be examined computationally.

Until now, bioinformatics has been a practical discipline through which to meet the needs for informatics technologies in large-scale data production in genomics and other high-throughput areas of biology. But as data are converted to knowledge and empirical rules lead to principles, bioinformatics is bound to become a more fundamental discipline. As illustrated in Fig. 3, bioinformatics in the future will encompass not only biology and practical aspects of informatics (computer science), but also mathematics and theoretical foundations to detect the basic architectures of complex biological information systems, and physics and chemistry to integrate physical and chemical principles with biological principles. When we have a complete computer representation of living cells and organisms and know the principles of how they compute, then, in the words of Sydney Brenner, "computational biology will become biological computation".

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
2. Lipman, D.J. & Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441 (1985).
3. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
4. Olson, M., Hood, L., Cantor, C. & Botstein D. A common language for physical mapping of the human genome. *Science* **245**, 1435–1435 (1989).
5. Adams, M.D. *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
6. Fleischmann, R.D. *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
7. Goffeau, A. *et al*. Life with 6000 genes. *Science* **274**, 546–567 (1996).
8. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
9. Adams, M.D. *et al*. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
10. Lander, E.S. *et al*. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
11. Venter, J.C. *et al*. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
12. Bork, P. & Koonin, E.V. Predicting functions from protein sequences—where are the bottlenecks? *Nat. Genet.* **18**, 313–318 (1998).
13. Park, J. *et al*. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201–1210 (1998).
14. Altschul, S.F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
15. Krogh, A., Brown, M., Mian, I.S., Sjolander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994).
16. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
17. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882 (1997).
18. Rost, B. & Sander, C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. USA* **90**, 7558–7562 (1993).
19. Nakai, K. & Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14**, 897–911 (1992).
20. Bork, P. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res.* **10**, 398–400 (2000).
21. Falquet, L. *et al*. The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**, 235–238 (2002).
22. Henikoff, J.G., Greene, E.A., Pietrokovski, S. & Henikoff, S. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* **28**, 228–230 (2000).
23. Attwood, T.K. *et al*. PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res.* **30**, 239–241 (2002).
24. Corpet, F., Servant, F., Gouzy, J. & Kahn, D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* **28**, 267–269 (2000).
25. Sonnhammer, E.L., Eddy, S.R., and Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420 (1997).
26. Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. USA* **95**, 5857–5864 (1998).
27. Haft, D.H. *et al*. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29**, 41–43 (2001).
28. Huynen, M., Snel, B., Lathe, W. III & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1204–1210 (2000).
29. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
30. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
31. Pease, A.C. *et al*. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* **91**, 5022–5026 (1994).
32. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metablic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
33. Tamayo, P. *et al*. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).
34. Brown, M.P. *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97**, 262–267 (2000).
35. Uetz, P. *et al*. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
36. Ito, T. *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
37. Gavin, A.C. *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 147 (2002).
38. Ho, Y. *et al*. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
39. von Mering, C. *et al*. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
40. Edwards, A.M. *et al*. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **18**, 529–536 (2002).
41. Ashburner, M. *et al*. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
42. Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **13**, 375–376 (1997).
43. Karp, P.D., Riley, M., Paley, S.M. & Pelligrini-Toole, A. EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* **24**, 32–39 (1996).
44. Ogata, H., Fujibuchi, W., Goto, S. & Kanehisa, M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.* **28**, 4021–4028 (2000).

45. Barabasi, A.L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
46. Watts, D.J. & Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
47. Milo, R. *et al*. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
48. Ideker, T. *et al*. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
49. Kumar, A. *et al*. Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719 (2002).
50. Kanehisa, M. *Post-Genome Informatics* (Oxford Univ. Press, Oxford, 2000).
51. Baxevanis, A.D. The molecular biology database collection: 2002 update. *Nucleic Acids Res.* **30**, 1–12 (2002).
52. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
53. Orengo, C.A. *et al*. CATH—a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
54. Wingender, E. *et al*. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*. **28**, 316–319 (2000).
55. Bader, G.D. *et al*. BIND—the biomolecular interaction network database. *Nucleic Acids Res*. **29**, 242–245 (2001).
56. Xenarios, I. *et al*. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).