CMLS, Cell. Mol. Life Sci. 60 (2003) 2637–2650
1420-682X/03/122637-14
DOI 10.1007/s00018-003-3114-8

CMLS Cellular and Molecular Life Sciences

Review

Automatic prediction of protein function

B. Rost [a,b,c,*], J. Liu [a,c,d], R. Nair [a,e], K. O. Wrzeszczynski [a] and Y. Ofran [a,f]

[a] Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, New York 10032 (USA), Fax: + 1 212 305 7932, e-mail: rost@columbia.edu
[b] Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, New York 10032 (USA)
[c] Northeast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, New York 10032 (USA)
[d] Department of Pharmacology, Columbia University, 630 West 168th Street, New York, New York 10032 (USA)
[e] Department of Physics, Columbia University, 538 West 120th Street, New York, New York 10027 (USA)
[f] Department of Biomedical Informatics, Columbia University, 630 West 168th Street, New York, New York 10032 (USA)

Received 26 March 2003; received after revision 15 May 2003; accepted 12 June 2003

Abstract. Most methods annotating protein function utilise sequence homology to proteins of experimentally known function. Such a homology-based annotation transfer is problematic and limited in scope. Therefore, computational biologists have begun to develop ab initio methods that predict aspects of function, including subcellular localization, post-translational modifications, functional type and protein-protein interactions. For the first two cases, the most accurate approaches rely on identifying short signalling motifs, while the most general methods utilise tools of artificial intelligence. An outstanding new method predicts classes of cellular function directly from sequence. Similarly, promising methods have been developed predicting protein-protein interaction partners at acceptable levels of accuracy for some pairs in entire proteomes. No matter how difficult the task, successes over the last few years have clearly paved the way for ab initio prediction of protein function.

Key words: Genome analysis; protein function prediction; ab initio prediction; neural networks; multiple alignments; sequence analysis; subcellular localization; post-translational modifications; protein-protein interactions; bioinformatics.

Introduction

‘Protein function’ is an operational concept

Proteins perform most important tasks in organisms, such as catalysis of biochemical reactions, transport of nutrients, recognition and transmission of signals. The plethora of aspects of the role of any particular protein is referred to as its ‘function’. However, protein function is not a well-defined term; instead, function is a complex phenomenon that is associated with many mutually overlapping levels: biochemical, cellular, organism-mediated, developmental and physiological. These overlapping levels are intertwined in complex ways; for example, protein kinases can be related to different cellular functions (such as cell cycle) and to a chemical function (transferase). The same kinase may also ‘misfunction’, thereby causing disease. Here we use the generalised, operational notion that ‘function is everything that happens to or through a protein’.

* Corresponding author.

Sequence-structure and sequence-function gaps

The first entire genome (DNA) sequence of a free-living organism, *Haemophilus influenzae*, was published in 1995 [1]. Now we know the genomes for more than 100 organisms; for more than 60, the data is publicly available and contributes about 250,000 protein sequences, that is, one-fourth of all known protein sequences [2–5; J. Liu and B. Rost, unpublished]. This explosion of sequence information has widened the gap between the number of protein sequences and the number of experimentally characterised proteins [4, 6–8]. Computational biology plays a central role in bridging this gap [9–14]. For about 10–40% of all sequences, we can deduce structure from homology to known structures [4, 15–20]. For about 40–60% of all sequences from current genome projects, sequence homology suggests some aspects of function [6, 21–23]. However, a firm conclusion about function is not always clear, as predictions can be anything from cellular function (e.g., adenosinetriphosphatase (ATPase) or ion channel) to details about cofactor binding sites (e.g., ATP binding sites).

Transfer of function based on sequence homology

Querying MEDLINE [24] with 'predict protein function' retrieves over 1000 papers from one year. The vast majority describes single-case studies in which experts combine many tools to guess aspects of function for a particular protein or protein family. Recently, James Whisstock and Arthur Lesk have focused on these aspects in an excellent, comprehensive review [25]. Here we focus mainly on ab initio methods that predict function in the absence of experimental annotations for homologues. We discuss some problems of homology transfer. We ignore methods that successfully identify functionally important residues from multiple alignments and/or protein structures [26–33]. Arguably the most successful approaches combine tools from artificial intelligence (neural networks, Hidden Markov Models (HMMs), Support Vector Machines (SVMs)) with evolutionary information contained in multiple alignments and aspects of protein structure.

Annotations and annotation transfer of protein function

Molecular biology databases with functional information

Information about protein sequences is stored in public databases such as SWISS-PROT and TrEMBL (table 1). SWISS-PROT [34] is a curated database of protein sequences that also contains annotations about function added by a team of experts who extract this information primarily from journal publications [35]. TrEMBL [34] consists of entries that are derived from the translation of all coding sequences in the EMBL nucleotide sequence database [36] and are not in SWISS-PROT. Unlike SWISS-PROT records, those in TrEMBL are awaiting manual annotation. SWISS-PROT currently contains 122,564 sequence entries, while the TrEMBL database contains over 821,014 sequence entries [34]. Many databases of protein families are derived from these original resources [6, 12, 37–43]. An issue that becomes increasingly important is the redundancy in original and derived databases. Such redundancy causes problems for database search techniques (alignments) and complicates estimates for the accuracy of annotation transfer [44]. A few resources address this problem by maintaining non-redundant subsets like KIND [45], CluSTr [46] or BLOCKS+ [47] databases; others provide tools to address the problem [48–50].

Transfer of annotation through homology

Experimentally determining protein function continues to be a laborious task that may take enormous resources. For instance, more than a decade after its discovery, we still do not know the precise and entire functional role of the prion protein [51]. The automatic elucidation of protein function is therefore an appealing challenge [25, 37, 52, 53]. Bioinformatics exploits that two proteins with similar sequence often have similar function. Albeit this concept appears straightforward, in practice, there are many hurdles to overcome. First, function is not well defined; hence, it is very difficult to create controlled vocabularies [54–56]. Second, the precise values for thresholds of significant sequence similarity (T) are actually specific to particular aspects of function and have to be re-established for any given task [44, 55, 57–64; K. O. Wrzeszczynski and B. Rost, unpublished] (fig. 1). The problem of annotating function was illustrated immediately after the release of the first genome [1]: 148 amendments were published a few weeks after the original publication [65]. Similar amendments followed most papers presenting entirely sequenced genomes [66–68]. Several pitfalls in transferring annotations of function have been reported, for example, having inadequate knowledge of thresholds for 'significant sequence similarity', using only the best database hit or ignoring the domain organisation of proteins [67–73]. However, Iyer et al. turned the issue of annotation transfer errors around by collecting a few examples for which subsequent experiments showed that theoretical predictions had been more accurate than previous experiments [74].

Problem 1: Multiple levels of description

Several groups and associations have ventured to introduce numerical schemata to define function. The first attempt was the introduction of enzyme classification num-

Table 1. Web sites of major databases and genome resources.

Name	Description	URL
General databases		
SWISS-PROT	annotated protein sequences	www.ebi.ac.uk/swissprot/
TrEMBL	translated protein sequences	www.ebi.ac.uk/trembl/
Gene Ontology (GO)	ontology of protein function	www.geneontology.org/
MIPS	annotation and ontology of function	mips.gsf.de/
Ensembl	proteins from human and mouse	www.ensembl.org/
Post-translational modification		
RESID	database of post-translational modifications	www.nbrf.georgetown.edu/pirwww/dbinfo/resid.html
PROSITE	database of protein motifs	www.expasy.ch/prosite/
PlantsP	database of phosphorylation for plants	plantsp.sdsc.edu
NetPhos	predict protein phosphorylation	www.cbs.dtu.dk/services/NetPhos/
NetOGlyc	predict O- a-GlcNAc glycosylation	www.cbs.dtu.dk/services/NetOGlyc/
DictyOGlyc	predict O-GalNAc glycosylation	www.cbs.dtu.dk/services/DictyOGlyc/
YinOYang	predict O-b-GlcNAc glycosylation and Yin-Yang sites	www.cbs.dtu.dk/services/YinOYang/
GPI-predict	predict GPI-anchored proteins	mendel.imp.univie.ac.at/gpi/gpi_prediction.html
The Sulfinator	predict tyrosine sulfation	us.expasy.org/tools/sulfinator/
Subcellular localisation		
HMMTOP	predict transmembrane helices	www.enzim.hu/hmmtop/
TMHMM	predict transmembrane helices	www.cbs.dtu.dk/services/TMHMM/
PHDhtm	predict transmembrane helices	cubic.bioc.columbia.edu/predictprotein/
PredictNLS	nuclear localisation signals	cubic.bioc.columbia.edu/predictNLS/
LOC3d	localisation for eukaryotic structures	cubic.bioc.columbia.edu/db/LOC3d/
PSORT II	predict localisation	psort.nibb.ac.jp/
NNPSL	predict localisation	www.doe-mbi.ucla.edu/cgi/astrid/nnpsl_mult.cgi
TargetP	combination of signal, chloroplast, and mitochondrial targeting signals	www.cbs.dtu.dk/services/TargetP/
	predict localisation of yeast proteins	bioinfo.mbb.yale.edu/genome/localize/
ProtComp	predict localisation for plants	www.softberry.com/berry.phtml?topic=proteinloc
Predotar	predict mitochondrial and plastid targeting	www.inra.fr/Internet/Produits/Predotar/
Processing, degradation and antigen presentation		
MEROPS	database of proteases	www.merops.co.uk
IMGT	immunogenetics database	imgt.cines.fr/
FIMM	database of functional immunology	sdmc.krdl.org.sg:8080/fimm/
MHCPEP	database of MHC-binding peptides	wehih.wehi.edu.au/mhcpep/
SYFPEITHI	database of MHC ligands and peptide motifs; also includes the prediction service	syfpeithi.bmi-heidelberg.com/scripts/MHCServer.dll/home.htm
BIMAS	predict HLA peptide binding	bimas.dcrt.nih.gov/molbio/hla_bind/
NetChop	predict human proteasome cleavage sites	www.cbs.dtu.dk/services/NetChop/
Functional class		
ProtFun	predict cellular, enzyme and GO class	www.cbs.dtu.dk/services/ProtFun/
Meta servers		
Pedant	proteome predictions and analysis	pedant.gsf.de/methods.html
PEP	predictions for entire proteomes	cubic.bioc.columbia.edu/db/PEP/

Note: URLs are given without the standard tag 'http://', e.g., http://www.geneontology.org/

bers (EC) [75]; this classification uses four digits to classify enzymatic activity [55]. The MIPS database attempts to extend this idea to a wider perspective of more proteins and more roles through their classification catalogue [76]. Another characterisation of protein function originates from the Gene Ontology (GO) Consortium [54]. GO distinguishes three levels of protein function: (i) molecular function, where the protein can, for example, catalyse a metabolic reaction and recognize or transmit a signal; (ii) biological process, in which a set of many cooperating proteins is responsible for achieving broad biological goals, for example, mitosis or purine metabolism, or signal transduction cascades; and (iii) cellular component, which includes the structure of subcellular compartments, the localisation of proteins, and macromolecular complexes. Examples include the nucleus, telomere, and origin recognition complex. The subcellular localisation of a protein is an essential attribute for this level. The totality of the physiological subsystems of the cell and their interplay with various environmental stimuli determines properties of the phenotype, the morphology and physiology of the organism, and the organism's behaviour. Although not complete, GO constitutes the best set of definitions available today.

Problem 2: Functional information not machine-readable

Nearly all databases present the protein sequence in formats that are more or less straightforward to parse by computers. However, annotations are mostly written in free text using a rich biological vocabulary that often varies in different areas of research. Such annotations are primarily meant for the eyes of human experts; hence, they are not machine-readable [77]. Another problem that hampers automatic annotations is the quality of database annotations: only a few database groups attempt quality control of curated annotations [78].

Establish accuracy of homology transfer

The reliability of transfer by homology depends on the particular feature of function/structure considered. In order to estimate the accuracy in transferring function given a particular threshold in sequence similarity, we have to complete the following three steps (fig. 1, top sketch): (i) build data sets that have experimental annotations about the presence (true, e.g., all proteins experimentally known to be nuclear) and absence (false, e.g., all proteins experimentally known not to be nuclear) of a certain aspect of function; (ii) to avoid estimates that are incorrectly biased by the distribution of today's experimental information [44], extract a representative subset of proteins from the true data and align it against all proteins in

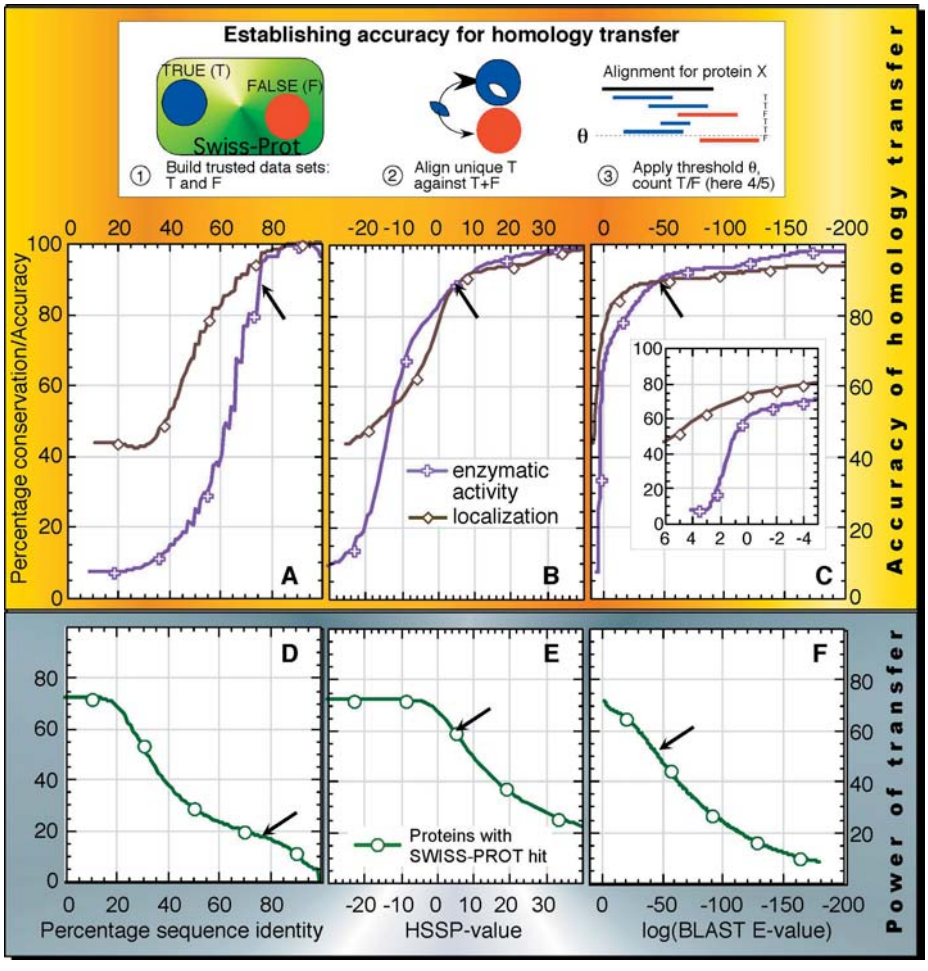

Figure 1. Accuracy and power of homology transfer. Thresholds for sequence similarity implying functional similarity depend on the particular aspect of function that we want to infer. For example, transfer of annotations for enzymatic activity (thick lines with open plus signs A–C) requires higher levels of similarity than transfer for annotations about subcellular localisation (thin lines with diamonds A–C). Even at levels above 80% pairwise identity, or for PSI-BLAST expectation values $<10^{-150}$, we still make mistakes in transferring EC numbers. For which fraction of entirely sequenced organisms can we transfer annotations? An upper limit is provided by the fraction of proteins that have sequence similarity to proteins from SWISS-PROT (E–G). If we want the transfer at error levels <10% (arrows A–C), maximally 60% of all proteins from 62 entirely sequenced organisms can be annotated (arrow F). This estimate provides an upper limit, since its two basic assumptions are likely overly optimistic: (i) not all SWISS-PROT proteins have reliable and detailed experimental annotations about function and (ii) the accuracy of homology transfer for details of the functional role may be much lower for mechanisms that are less local than enzymatic activity.

the true set (minus the representative subset) and false set; (iii) for all alignments, count how many true and false we find at every given threshold for sequence similarity. How is sequence similarity measured? The most popular way is the level of pairwise sequence identity, that is, the percentage of residues that are identical in an alignment of two proteins (R on R → 1, R on K → 0). The major problem with such a score in the context of automatic annotations is that it does not reflect the length of the alignment. For example, peptides with 11 identical residues may differ in both function and structure [44, 59, 79]. On the other hand, levels of pairwise sequence identity such as 33% for alignments longer than 100 residues or 22% for alignments longer than 250 residues imply similarity in structure [79]. This observation is used to compile an empirical threshold for significant sequence similarity as a function of alignment length [79–81]. We refer to this threshold as the HSSP-value; it is empirically chosen such that any pair of proteins A, B have similar structure if HSSP-value (A,B) > 0. Another measure of sequence similarity is the expectation value built into the popular PSI-BLAST [82] alignment program. An important point to realise for BLAST and PSI-BLAST users is that the expectation value depends on the size of the database used to search for related proteins. This implies the following: assume we align proteins A and B by pairwise BLAST in two ways, (i) by searching with A against SWISS-PROT and (ii) by searching with A against SWISS-PROT + PDB (Protein Data Bank) [83]. Even if the resulting alignments between A and B are identical, the expectation values may differ significantly because of the difference in size of the two databases. Unfortunately, the accuracy of transferring different aspects of function differs substantially (fig. 1A–C illustrates this for the case of localisation and enzymatic activity).

Most annotations of function are through homology transfer

In general, the inference of function is reliable only for very high levels of sequence similarity [44, 58, 59]. Although some perceive the estimate that 30% of the annotations may contain errors as particularly high [71], our analysis of the sequence conservation of enzymatic activity suggested that this value may be overly optimistic [44]: if we want to transfer the full enzymatic activity with less than 30% errors, we have to require levels of >60% pairwise sequence identity, and for errors below 10%, >75% sequence identity (fig. 1A, arrow). For the same error rate (<10%), the HSSP-value must be >5 (fig. 1B) and the PSI-BLAST expectation value $<10^{-48}$ (fig. 1C). How many proteins from entire proteomes can we annotate at such a level of accuracy? We aligned all proteins from 62 entirely sequenced organisms [3] by PSI-BLAST (protocol described in detail elsewhere [84] but basically three iterations at 10^{-10} thresholds) against a database containing all proteins from SWISS-PROT, TrEMBL and PDB. Then we monitored at which level of sequence similarity we found the most similar protein in SWISS-PROT or PDB. If we assume that all proteins in SWISS-PROT and PDB have complete annotations about function, and that the accuracy of homology transfer for all aspects of function is similar to that for enzymatic activity, then we simply have to mark the points of 90% accuracy (fig. 1D–F, arrows). Maximally, when using the HSSP-value for annotation, we can thus transfer annotation for about 60% of all proteins in the 62 proteomes. When we require less than 5% errors, the number drops to about 35% of all proteins, and when permitting 40% errors, it rises to above 70% of all proteins. The latter (70%) also constitutes the saturation: for about 25–30% of the proteins from proteomes, we find no protein in SWISS-PROT or PDB even at thresholds of sequence similarity that are far too permissive to transfer annotations (fig. 1). These estimates are likely to constitute upper limits since the assumption that all proteins in SWISS-PROT and PDB are fully annotated experimentally is overly optimistic. Nevertheless, we currently know more than 1.4 million protein sequences. Even if we pessimistically expect the ratio of reliable transfers to be only 10%, we still conclude that most annotations about function result from homology transfer. Furthermore, all these numbers ignore the capability of experts, who can increase accuracy by combining many resources to annotate families [25], as realised, for instance, in Pfam-A [85] and TIGRFAMs [The Institute for Genomic Research (TIGR) families] [86].

Subcellular localisation

Basic concept

Bacterial cells generally consist of a single intracellular compartment surrounded by a plasma membrane. In contrast, eukaryotic cells are elaborately subdivided into functionally distinct, membrane-bound compartments. Most eukaryotic proteins are encoded in the nuclear genome and synthesised in the cytosol, and many need to be further sorted into other subcellular compartments. The sorting signals that direct the movement of a protein through the cell, and thereby determine its eventual subcellular localisation, are contained in its amino acid sequence [87, 88]. Proteins that remain in the cytosol do not have sorting signals. Many others, however, have specific sorting signals that direct their transport from the cytosol into the nucleus, endoplasmic reticulum (ER), mitochondria, plastids (in plants) or peroxisomes. Sorting signals can also direct the transport of proteins from the ER to other destinations in the cell [89]. Proteins must be localised in the same subcellular compartment to cooperate

toward a common physiological function. Thus, the native subcellular localization of a protein is one indicator of protein function. Aberrant subcellular localisation of proteins has been observed in the cells of several diseases, such as cancer and Alzheimer disease. Attempts to predict subcellular localisation have become a central task in bioinformatics [77, 90]. The main methods are based on homology transfer, motif recognition or the correlation between sequence features and localisation (fig. 2).

Prediction of localisation through sequence motifs

One means for predicting localisation is the identification of local sequence motifs such as signal peptides or nuclear localisation signals (NLSs). A number of neural network-based tools identify signal peptides that target proteins to the secretory pathway and the mitochondria [91, 92]. In a recent benchmark study [93], these tools predicted signal peptide cleavage site at >80% accuracy. A particular problem for methods detecting N-terminal signals is that start codons are predicted with less than 70% accuracy by genome projects [94, 95]. We have collected a data set of experimental and potential NLS motifs that predict nuclear localisation at 100% accuracy [96, 97]. The downside of this look-up library is that it is not complete: most proteins have no known NLS. Either the motif remains to be discovered or the protein is imported into the nucleus through binding to another protein that has an NLS. Overall, known and predicted sequence motifs enable annotating about 30% of the proteins in six eukaryotic proteomes [3, 15].

Ab initio methods predict localisation for all proteins at lower accuracy

Another approach to predicting localisation has been suggested by the observation that the total amino acid composition correlates with the subcellular localisation [98–103]. This observation has led to the development of a variety of prediction methods based solely on composition [94, 104–106]. With the availability of large numbers of completely sequenced genomes, phylogenetic profiles have been employed to identify subcellular localisation [107]. So far, this approach has been much less accurate in predicting localisation than methods based

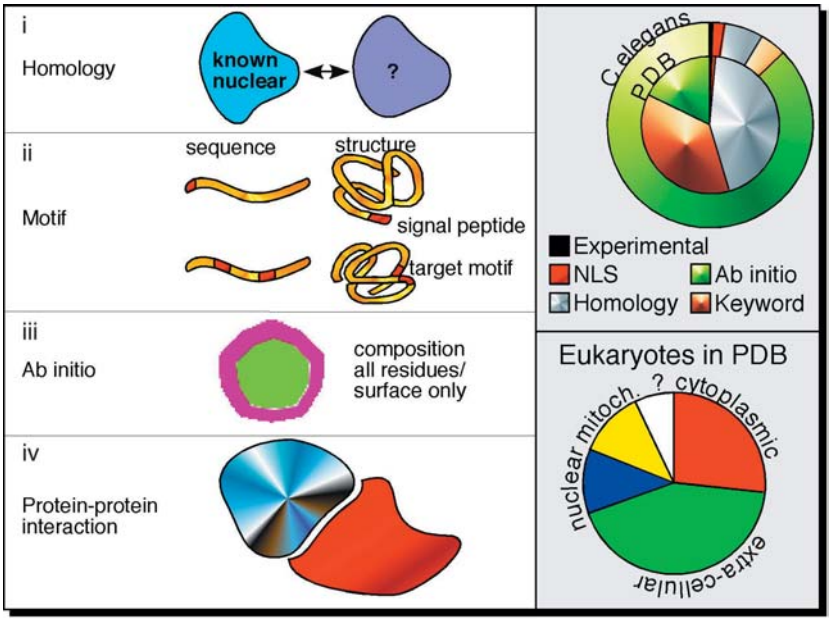

Figure 2. Methods predicting subcellular localisation. Four types of methods currently predict subcellular localisation. (i) Transfer by homology: if we know that protein A is nuclear and we find protein B very similar in sequence to A, we can usually infer that B is also nuclear (fig. 1A–C, thin lines). (ii) Identification by motifs: many proteins are shuttled between different compartments by carrier proteins that recognise short sequence motifs. Some of these motifs are consecutive in sequence (signal peptide, nuclear localisation signal), while others are discernible only from the folded structure (lysosomal retention signals). (iii) Ab initio methods exploit the correlation between sequence features and localisation. (iv) Protein-protein interactions are another mechanism to shuttle proteins between compartments. Assume that two interacting proteins A and B are nuclear and that A has a nuclear localisation signal that is recognised by an importin that carries A into the nucleus; B could be imported into the nucleus by binding to the complex A-importin. Recently, we combined the first three methods with another method that automatically recognises keywords in SWISS-PROT annotations [180] to annotate the localisation of all eukaryotic proteins of known structure [97, 112]. The vast majority of all annotations resulted from homology transfer or lexical analysis (inner circle of top pie chart). When applying the same methods to the entire proteome of *Caenorhabditis elegans*, this picture changed completely: about 87% of all proteins could only be handled by ab initio methods. Interestingly, 43% of all eukaryotic proteins of known structure appear to be extracellular (lower pie).

solely on composition. Other methods have tried to integrate rules based on amino acid composition with databases of known signal sequences; e.g., PSORT II is a knowledge-based expert system that integrates the two kinds of information [108]. In particular, PSORT II uses other original prediction methods such as SignalP [109], ChloroP [110] and NNPSL (Neural Netwoks Predicting Subcellular Localization) [94] as input. Consequently, we may expect that PSORT II would improve if these original methods were improved. Drawid and Gerstein have proposed a Bayesian system based on a diverse range of 30 different features [111]. They applied their method to predicting localisation of the full *Saccharomyces cerevisiae* proteome and providing estimates of the fraction of all yeast proteins found in different compartments. We have recently combined homology transfer with motif-based and ab initio predictions to annotate all eukaryotic proteins of known structure (fig. 2). We learned that combining evolutionary and structural information yielded the most accurate predictions and that prediction methods appeared far less accurate when presented with fragments of the native protein sequence [112].

Post-translational modifications

Basic concept

While more than 325 structural and regulatory post-translational modifications in proteins are known today [113], prediction methods are currently constrained to a few of the most relevant of these. These tools typically employ highly conserved sequence motifs, more complex sequence patterns, or structural properties such as solvent accessibility. Prominent post-translational modifications targeted for prediction include: N-terminal signal peptide cleavage sites [91, 93, 103, 114–118], proteolytic cleavage and, more specifically, proteasome cleavage sites [116, 118–124], phosphorylation sites [125, 126], lipid modification [127] and *N*- and *O*-glycosylations [128, 129].

Archiving known sequence motifs and predicting modifications

PhosphoBase [126] includes information on more than 400 phosphorylated proteins, their phosphorylation sites and the specific kinase of action. These data were used to develop an ab initio method that predicts phosphorylation sites (NetPhos) [125]; predictions for serine, threonine and tyrosine residues reach 69–96% sensitivity. The method uses information about sequence and structure. Given the difficulty in predicting structure around the phosphorylation site and the considerable variation of consensus sequences for kinase substrate specificity, the prediction of phosphorylation remains a difficult task. A similar neural network approach based on charged residues within glycosylation sites together with sequence context and surface accessibility is used to identify *O*-glycosylation modifications at about 80% accuracy [129]. The limited substrate specificity for both *N*-glycosylation and *O*-glycosylation currently limits progress [130]. Predictions of lipid modifications are currently restricted to glycosylphosphatidylinositol (GPI) anchors [127]. C-terminal motifs (omega site) and physical properties of GPI anchors enabled accurate predictions for the effects of mutations on known anchors. N-terminal motifs apparently allow for accurate predictions of *N*-myristoyltransferase (NMT) substrate sites [131]. Finally, a comprehensive study of proteasome digestion data yields a method that accurately predicts major histocompatibility complex (MHC) class I ligand boundaries after proteasomal degradation: 65% of the cleavage sites and 85% of the non-cleavage sites appeared to be predicted correctly [120].

Functional type

Basic concept

Monica Riley introduced the most widely used schema for classes of cellular function to annotate *Escherichia coli* [132]. TIGR [1] and many other genome centres have adopted this schema with minor modifications. Transferring annotations of cellular function by homology has for long been almost the only field in which methods were developed. In fact, many researchers exclusively consider such methods when referring to the prediction of protein function. Recently, however, groups have begun developing methods that predict functional classes in the absence of experimental annotations.

Functional classes can be predicted from sequence

An interesting hybrid system uses inductive logic programming to predict functional classes with and without homology to experimentally annotated proteins [133]. While it is not clear how successful the system is in ab initio prediction, on average the levels of accuracy published appear promising. Genes located in a close neighbourhood on the genome may have some functional commonalities. While such neighbourhood relations sometimes enable predicting aspects such as classes of cellular function, the average signal is very weak; that is, most often neighbours are not related in function [134–136]. The most recent breakthrough in the field of predicting protein function came through a collaboration of the groups from Søren Brunak (CBS Copenhagen) and Alfonso Valencia (CNB Madrid). Their ends are to predict cellular function from sequence alone. Their means are complex, elaborate and hierarchical systems of neural networks

[137]. A first group of networks is used to identify 'sequence features' (such as protein length or amino acid composition) that optimally separate between any two types of functional classes. These basic predictions are then combined into a final prediction step, again through neural networks. The authors applied their method to annotating functional classes for all human proteins. For example, the prion protein is predicted to belong to the 'transport and binding category' and to 'not have enzymatic activity'. This appears compatible with the observation that prion binds and transports copper, while no catalytic activity has ever been observed [138]. Recently, the Brunak group has applied its new concepts to identifying novel enzymes in archae [139] and to predicting the functional type of all human proteins according to the GO classification [140]. The most impressive news from these groundbreaking methods is that aspects of function can be predicted without homology, that is, for completely uncharacterised proteins.

Protein-protein interactions

Basic concept

Every protein has a biological function, yet most of the biological functions are carried out by groups of proteins interacting in complex networks. Interactions between proteins can be physical (i.e., by chemically binding each other or by binding together to a third substrate), or they can be functional (e.g., by controlling each others' expression or by participating in the same biochemical pathway). To fully understand the molecular mechanism that underlies a certain biological function (or malfunction), we need to decipher the meticulous networks of protein interactions that underlie these mechanisms. Therefore, an extensive research effort is invested in both experimental and computational methods that unravel protein-protein interactions [14, 29, 141–157]. Particularly, many methods and databases attempt to draw complete maps of interactions for entire proteomes. Once it is known with which other proteins a newly discovered protein interacts, it will be easier to predict its function. Furthermore, it is hoped that these interaction maps will surrender the secrets of biological processes and enhance the understanding of the underlying molecular mechanisms. A complete picture of all the proteins that are involved in a certain biological process would also break new ground in drug development by identifying new targets for drugs.

Databases and data-mining techniques compile existing information

A vast amount of information about protein-protein interactions already exists in the literature. However, this information is scattered across millions of text pages of scientific publications. A few different enterprises are aimed at extracting this information from the literature [14, 150–152, 158–162]. The DIP database [162] is an example of a database that is dedicated to protein interactions. The curators of DIP manually survey the literature to find experimentally determined interactions. They also employ automatic techniques to obtain data from other databases. Other approaches to this problem use natural-language-processing algorithms as well as other computational methods to automatically extract interaction information from scientific papers [13, 14, 158–160, 163–165]. SUISEKI, a system for information extraction on interactions [159], is reported to successfully extract 70–80% of the interactions in a large corpus of scientific abstracts.

Computational approaches predict protein-protein interactions

Many groups attempt to develop computational methods that predict protein-protein interactions in silico [14, 134] (fig. 3). Although not all proteins that come from neighbouring genes on the genome interact with one another, gene location occasionally reveals true protein-protein interactions [166, 167]. Another approach screens genomes for sequences that appear as two different chains in one genome and are fused to create a single protein-chain in another genome that is evolutionarily younger [168, 169]. The assumption is that evolution fused these two proteins into a single one because they interact with one another. Another comparative method searches for pairs of proteins that always occur together in all known genomes; that is, there is no genome in which only one of the two proteins occurs [168, 170]. These types of protein pairs are very likely to interact. Using these two methods, Eisenberg et al. [168] proposed thousands of protein pairs that may interact. However, there are no confirmed statistics regarding the reliability of the predictions of these methods. The assumption that interacting proteins co-evolve gave rise to other prediction methods. One specific implementation uses the observation that interacting proteins sometimes have phylogenetic trees that are mirror images of each other [171–173]. Alfonso Valencia and his group introduced an approach based on the assumption that interacting proteins evolve together; hence, the mutations that occur in two interacting proteins along evolution should be correlated. First, they demonstrated that such correlated mutations could distinguish between correct and incorrect docking solutions [174]. Then they developed a method that predicts protein-protein interaction partners by analysing the correlation between the mutations in different proteins across different species [175]. Preliminary results indicate that the predictions of these methods have a low false-negative rate. Sprinzak and Margalit [176]

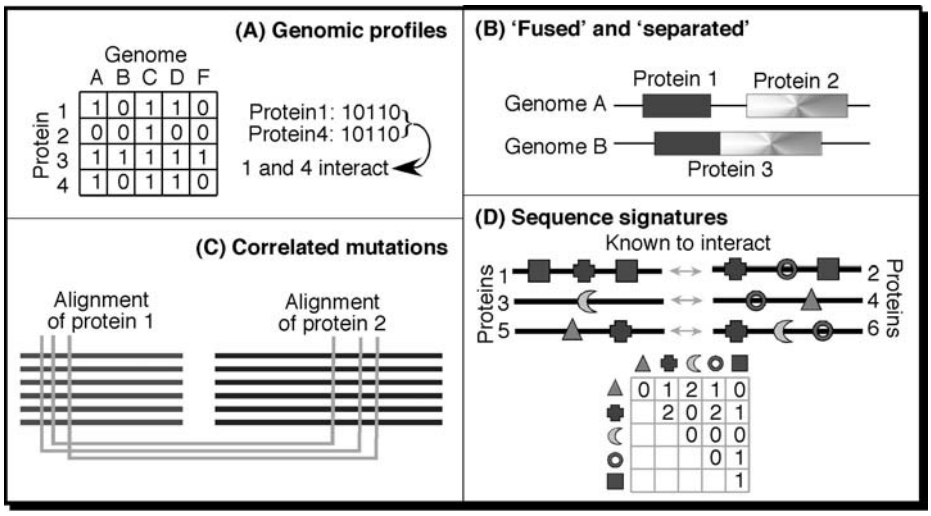

Figure 3. Methods predicting protein-protein interactions. (*A*) Genomic profiles [168, 170]: entire genomes are searched for the presence of each protein; the table represents the presence (1) or absence (0) of a certain protein in a given genome. If two proteins have an identical pattern, that is, neither of them appears in any of the genomes without the other, the method assumes that they interact. (*B*) Rosetta stone [168, 169]: if two separate proteins 1 and 2 are in genome A, and the same two are merged as one single protein in genome B, the method assumes that proteins 1 and 2 interact with one another. (*C*) Correlated mutations [174]: if a particular pair of interacting residues is important to maintain the interaction between protein 1 and 2, we might expect to find examples of mutations that are correlated between 1 and 2; for example, a positive acid in protein 1 salt-bridging a negative acid in 2 might be mutated to a negative acid. This could be compensated by a reverse mutation of the negative acid into a positive acid in protein 2. A refined version of this basic idea is implemented in methods that predict protein-protein interaction sites and interaction pairs based on correlated mutations. (*D*) Sequence signatures [176, 177]: sequence motifs are marked in pairs of proteins that interact. The likelihood of interaction between every pair of motifs is used to predict interactions between the proteins carrying these motifs.

predicted protein-protein interactions based on a very simple concept: assume we experimentally know that proteins P1 and P2 interact and that both contain the particular motifs or domains M1 and M2. If we find the same motifs in proteins P1′ and P2′, we might suspect that P1′ and P2′ also interact. The method can be improved by adding filters that take into account entire networks by skewing the probability for the prediction of the interaction between P1′ and P2′ according to how often this combination is observed in an organism [177]. Aloy and Russell use alignments and 3D structures of known interactions to predict possible binding partners [178]. Given a 3D structure of a complex, they assess the likelihood that homologues are involved in similar interactions. An extension of this concept has been proposed by Skolnick et al., who applied algorithms developed to detect more distant sequence relations (threading) to identify binding partners given an experimentally known 3D complex [179]. However, the major restriction of all these methods is that each of them is applicable only to a limited set of proteins. Another shortcoming of most methods is that they merely indicate whether a pair of proteins is in interaction, but they do not identify the interaction sites, a crucial piece of information for molecular research.

Conclusions

Homology transfer of function: use with extreme caution!

No matter what definition of 'the function' or 'the fold' of a protein you may have, function clearly involves fewer residues directly. Therefore, random mutations are more likely to influence function than structure. In other words, when proteins diverge they will, on average, lose their function before they lose their fold. This makes it more difficult for computational biology to predict function than to predict structure. Despite this problem, the most accurate means of predicting function for particular proteins undoubtedly is the expert-controlled transfer of experimental annotations through homology [25]. However, in the context of automatic, non-expert and/or proteome-wide searches, homology transfer becomes problematic. On the one hand, we need very high levels of sequence similarity to reliably infer aspects of function through homology (fig. 1A). On the other hand, the likelihood of finding close homologues is exponentially smaller than that of finding more divergent homologues (fig. 1D). Thus, we find relatively few homologues of known function that allow transfer at very high accuracy. Many estimates for the functional coverage of entire genomes appear optimistically high by accepting very high error rates. An additional complication is that computational biology has to establish thresholds for accuracy of trans-

ferring function by homology for any aspect of function. Due to the lack of experimental data in general, and of machine-readable data in particular, such analyses have just begun over the last few years. What we can learn from the large-scale analyses for the few aspects is extreme caution in transferring function by homology!

Ab initio prediction of function: first successes scored

For some aspects of function, such as subcellular localisation, ab initio prediction methods have been pursued for a while. However, for most aspects of function the transfer by homology has long been the only means. Thus, overall the field of predicting function in silico is still at its infancy. Nevertheless, a few very promising methods have been proposed recently. Methods that predict subcellular localisation are becoming increasingly accurate; methods that predict post-translational modifications are becoming increasingly useful and comprehensive, although the vast majority of post-translational modifications experimentally observed has not been covered yet. The first breakthroughs have been made in predicting protein-protein interactions and cellular function from sequence. In combination, all those novel methods may aid the advance of molecular biology considerably. Given that the appetite of molecular and medical biologists for functional annotations grows with the exponential increase in the number of known proteomes, the recent advances in computational biology are falling on fertile ground.

Acknowledgements. Particular thanks to Arthur Lesk (MRC, Cambridge, England) for essential comments and for making his master review available to us before publication. Our work was supported by the grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institutes of Health (NIH), and the grant DBI-0131168 from the National Science Foundation (NSF). Last but not least, thanks to the GeneOntology team around Michael Ashburner (Cambridge, England) for their gargantuan effort, to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (University of California, San Diego) and their crews for maintaining excellent databases, and to all experimentalists who enable computational biology by making their data publicly available.

1 Fleischmann R. D., Adams M. D., White O., Clayton R. A., Kirkness E. F., Kerlavage A. R. et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:** 496–512
2 Liu J. and Rost B. (2001) Comparing function and structure between entire proteomes. Protein Sci. **10:** 1970–1979
3 Carter P., Liu J. and Rost B. (2003) PEP: predictions for entire proteomes. Nucleic Acids Res. **31:** 410–413
4 Pruess M., Fleischmann W., Kanapin A., Karavidopoulou Y., Kersey P., Kriventseva E. et al. (2003) The Proteome analysis database: a tool for the in silico analysis of whole proteomes. Nucleic Acids Res. **31:** 414–417
5 Reference removed in proof
6 Koonin E. V. (2001) Computational genomics. Curr. Biol. **11:** R155–R158
7 Andrade M. A. and Bork P. (2000) Automated extraction of information in molecular biology. FEBS Lett. **476:** 12–17
8 Lewis S., Ashburner M. and Reese M. G. (2000) Annotating eukaryote genomes. Curr. Opin. Str. Biol. **10:** 349–354
9 Fleischmann W., Moller S., Gateau A. and Apweiler R. (1999) A novel method for automatic functional annotation of proteins. Bioinformatics **15:** 228–233
10 Luscombe N. M., Laskowski R. A. and Thornton J. M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. **29:** 2860–2874
11 Thornton J. M. (2001) From genome to function. Science **292:** 2095–2097
12 Holm L. and Sander C. (1999) Protein folds and families: sequence and structure alignments. Nucleic Acids Res. **27:** 244–247
13 Valencia A. (2002) Bioinformatics: biology by other means. Bioinformatics **18:** 1551–1552
14 Valencia A. and Pazos F. (2002) Computational methods for the prediction of protein interactions. Curr. Opin. Str. Biol. **12:** 368–373
15 Liu J. and Rost B. (2002) Target space for structural genomics revisited. Bioinformatics **18:** 922–933
16 Teichmann S. A., Chothia C. and Gerstein M. (1999) Advances in structural genomics. Curr. Opin. Str. Biol. **9:** 390–399
17 Vitkup D., Melamud E., Moult J. and Sander C. (2001) Completeness in structural genomics. Nat. Struct. Biol. **8:** 559–566
18 Moult J. and Melamud E. (2000) From fold to function. Curr. Opin. Str. Biol. **10:** 384–389
19 Wolf Y., Brenner S., Bash P. and Koonin E. (1999) Distribution of protein folds in the three superkingdoms of life. Genome Res. **9:** 17–26
20 Gerstein M. and Levitt M. (1997) A structural census of the current population of protein sequences. Proc. Natl. Acad. Sci. USA **94:** 11911–11916
21 Andrade M. A., Brown N. P., Leroy C., Hoersch S., de Daruvar A., Reich C. et al. (1999) Automated genome sequence analysis and annotation. Bioinformatics **15:** 391–412
22 Bork P., Ouzounis C., Sander C., Scharf M., Schneider R. and Sonnhammer E. (1992) What's in a genome? Nature **358:** 287
23 Iliopoulos I., Tsoka S., Andrade M. A., Janssen P., Audit B., Tramontano A. et al. (2001) Genome sequences and great expectations. Genome Biol. **2:** interactions 2000
24 Airozo D., Allard R., Brylawski B., Canese K., Kenton D., Knecht L. et al. (1999) MEDLINE, vol. 1999. National Library of Medicine (NLM)
25 Whisstock J. C. and Lesk A. M. (2003) Prediction of protein function from protein sequence and structure. Quart. Rev. Biophys. in press
26 Casari G., Sander C. and Valencia A. (1995) A method to predict functional residues in proteins. Nat. Struct. Biol. **2:** 171–178
27 del Sol Mesa A., Pazos F. and Valencia A. (2003) Automatic methods for predicting functionally important residues. J. Mol. Biol. **326:** 1289–1302
28 Lichtarge O., Bourne H. R. and Cohen F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. **257:** 342–358
29 Lichtarge O. and Sowa M. E. (2002) Evolutionary predictions of binding surfaces and interactions. Curr. Opin. Str. Biol. **12:** 21–27
30 Pupko T., Bell R. E., Mayrose I., Glaser F. and Ben-Tal N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics **18:** S71–S77
31 Glaser F., Pupko T., Paz I., Bell R. E., Bechor-Shental D., Martz E. et al. (2003) ConSurf: identification of functional re-

gions in proteins by surface-mapping of phylogenetic information. Bioinformatics **19:** 163–164
32 Mizuguchi K., Deane C. M., Blundell T. L., Johnson M. S. and Overington J. P. (1998) JOY: protein sequence-structure representation and analysis. Bioinformatics **14:** 617–623
33 Andersen C. A. F., Palmer A. G., Brunak S. and Rost B. (2002) Continuum secondary structure captures protein flexibility. Structure **10:** 175–184
34 Boeckmann B., Bairoch A., Apweiler R., Blatter M. C., Estreicher A., Gasteiger E. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31:** 365–370
35 Junker V., Contrino S., Fleischmann W., Hermjakob H., Lang F., Magrane M. et al. (2000) The role SWISS-PROT and TrEMBL play in the genome research environment. J. Biotechnol. **78:** 221–234
36 Stoesser G., Baker W., van Den Broek A., Camon E., Garcia-Pastor M., Kanz C. et al. (2001) The EMBL nucleotide sequence database. Nucleic Acids Res. **29:** 17–21
37 Apweiler R. (2000) Protein sequence databases. Adv. Protein Chem. **54:** 31–71
38 Tamames J., Clark D., Herrero J., Dopazo J., Blaschke C., Fernandez J. M. et al. (2002) Bioinformatics methods for the analysis of expression arrays: data clustering and information extraction. J. Biotechnol. **98:** 269–283
39 Sigrist C. J., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M. et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Briefing Bioinf. **3:** 265–274
40 Frishman D., Kaps A. and Mewes H. W. (2002) Online genomics facilities in the new millennium. Pharmacogenomics **3:** 265–271
41 Kriventseva E. V., Biswas M. and Apweiler R. (2001) Clustering and analysis of protein families. Curr. Opin. Str. Biol. **11:** 334–339
42 Liu J. and Rost B. (2003) Domains, motifs and clusters in the protein universe. Curr. Opin. Chem. Biol. **7:** 5–11
43 Xu D., Xu Y. and Uberbacher E. C. (2000) Computational tools for protein modeling. Curr. Protein Pept. Sci. **1:** 1–21
44 Rost B. (2002) Enzyme function less conserved than anticipated. J. Mol. Biol. **318:** 595–608
45 Kallberg Y. and Persson B. (1999) KIND-a non-redundant protein database. Bioinformatics **15:** 260–261
46 Kriventseva E. V., Servant F. and Apweiler R. (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. Nucleic Acids Res. **31:** 388–389
47 Henikoff S., Henikoff J. G. and Pietrokovski S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics **15:** 471–479
48 O'Donovan C., Martin M. J., Glemet E., Codani J. J. and Apweiler R. (1999) Removing redundancy in SWISS-PROT and TrEMBL. Bioinformatics **15:** 258–259
49 Li W., Jaroszewski L. and Godzik A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics **17:** 282–283
50 Mika S. and Rost B. (2003) UniqueProt: creating representative protein sequence sets. Nucleic Acids Res. **31:** 3789–3791
51 Harrison P. M., Bamborough P., Daggett V., Prusiner S. and Cohen F. E. (1997) The prion folding problem. Curr. Opin. Str. Biol. **7:** 53–59
52 Gaasterland T. and Sensen C. W. (1996) Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. Biochimie **78:** 302–310
53 Eisenberg D., Marcotte E. M., Xenarios I. and Yeates T. O. (2000) Protein function in the post-genomic era. Nature **405:** 823–826
54 Ashburner M., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P. et al. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. **25:** 25–29
55 Todd A. E., Orengo C. A. and Thornton J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. **307:** 1113–1143
56 O'Donovan C., Martin M. J., Gattiker A., Gasteiger E., Bairoch A. and Apweiler R. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Briefing Bioinf. **3:** 275–284
57 Wrzeszczynski K. O. and Rost B. (2003) Cataloguing proteins in cell cycle control. In: Cell Cycle Checkpoint Control Protocols, pp. 219–233, Lieberman, H. (ed), Humana Press, Totowa, NJ
58 Devos D. and Valencia A. (2000) Practical limits of function prediction. Proteins **41:** 98–107
59 Nair R. and Rost B. (2002) Sequence conserved for sub-cellular localization. Protein Sci. **11:** 2836–2847
60 Pawlowski K., Jaroszewski L., Rychlewski L. and Godzik A. (2000) Sensitive sequence comparison as protein function predictor. Pac. Symp. Biocomput. **8:** 42–53
61 Shah I. and Hunter L. (1997) Fifth International Conference on Intelligent Systems for Molecular Biology, Halkidiki, Greece
62 Wilson C. A., Kreychman J. and Gerstein M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J. Mol. Biol. **297:** 233–249
63 Reference removed in proof
64 Ouzounis C., Perez-Irratxeta C., Sander C. and Valencia A. (1998) Are binding residues conserved? Pac. Symp. Biocomput. **3:** 399–410
65 Casari G., Andrade M. A., Bork P., Boyle J., Daruvar A., Ouzounis C. et al. (1995) Challenging times for bioinformatics. Nature **376:** 647–648
66 Ouzounis C., Casari G., Sander C., Tamames J. and Valencia A. (1996) Computational comparisons of model genomes. Trends Biotechnol. **14:** 280–285
67 Kyrpides N. C. and Ouzounis C. A. (1999) Whole-genome sequence annotation: 'Going wrong with confidence'. Mol. Microbiol. **32:** 886–887
68 Kyrpides N. C. and Ouzounis C. A. (1998) Errors in genome reviews. Science **281:** 1457
69 Mushegian A. R. (2000) Annotations of biochemically uncharacterized open reading frames (ORFs). Mol. Microbiol. **35:** 697–698
70 Tamames J., Gonzalez-Moreno M., Mingorance J., Valencia A. and Vicente M. (2001) Bringing gene order into bacterial shape. Trends Genet. **17:** 124–126
71 Devos D. and Valencia A. (2001) Intrinsic errors in genome annotation. Trends Genet. **17:** 429–431
72 Galperin M. Y. and Koonin E. V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol. **1:** 55–67
73 Brenner S. E. (1999) Errors in genome annotation. Trends Genet. **15:** 132–133
74 Iyer L. M., Aravind L., Bork P., Hofmann K., Mushegian A. R., Zhulin I. B. et al. (2001) Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. Genome Biol. **2:** RESEARCH0051
75 Webb E. C. (1992) Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, Academic Press, New York
76 Mewes H. W., Frishman D., Guldener U., Mannhaupt G., Mayer K., Mokrejs M. et al. (2002) MIPS: a database for

genomes and protein sequences. Nucleic Acids Res. **30:** 31–34
77 Eisenhaber F. and Bork P. (1998) Wanted: subcellular localization of proteins based on sequence. Trends Cell Biol. **8:** 169–170
78 Tsoka S. and Ouzounis C. A. (2001) Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*. Genome Res. **11:** 1503–1510
79 Rost B. (1999) Twilight zone of protein sequence alignments. Prot. Eng. **12:** 85–94
80 Sander C. and Schneider R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. Proteins **9:** 56–68
81 Nielsen H., Engelbrecht J., von Heijne G. and Brunak S. (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. Proteins **24:** 165–177
82 Altschul S., Madden T., Shaffer A., Zhang J., Zhang Z., Miller W. et al. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402
83 Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H. et al. (2000) The protein data bank. Nucleic Acids Res. **28:** 235–242
84 Przybylski D. and Rost B. (2002) Alignments grow, secondary structure prediction improves. Proteins **46:** 195–205
85 Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S. R. et al. (2002) The Pfam protein families database. Nucleic Acids Res. **30:** 276–280
86 Haft D. H., Selengut J. D. and White O. (2003) The TIGRFAMs database of protein families. Nucleic Acids Res. **31:** 371– 373
87 Mattaj I. W. and Englmeier L. (1998) Nucleocytoplasmic transport: the soluble phase. Annu. Rev. Biochem. **67:** 265– 306
88 Schatz G. and Dobberstein B. (1996) Common principles of protein translocation across membranes. Science **271:** 1519–1526
89 Pelham H. R. and Rothman J. E. (2000) The debate about transport in the Golgi – two sides of the same coin? Cell **102:** 713–719
90 Nakai K. (2000) Protein sorting signals and prediction of subcellular localization. Adv. Protein Chem. **54:** 277–344
91 Nielsen H., Brunak S. and von Heijne G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Prot. Eng. **12:** 3–9
92 Emanuelsson O., von Heijne G. and Schneider G. (2001) Analysis and prediction of mitochondrial targeting peptides. Methods Cell Biol. **65:** 175–187
93 Menne K. M., Hermjakob H. and Apweiler R. (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics **16:** 741–742
94 Reinhardt A. and Hubbard T. (1998) Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. **26:** 2230–2235
95 Gaasterland T. and Oprea M. (2001) Whole-genome analysis: annotations and updates. Curr. Opin. Str. Biol. **11:** 377–381
96 Cokol M., Nair R. and Rost B. (2000) Finding nuclear localisation signals. EMBO Rep. **1:** 411–415
97 Nair R., Carter P. and Rost B. (2003) NLSdb: database of nuclear localization signals. Nucleic Acids Res. **31:** 397–399
98 Nishikawa K. and Ooi T. (1982) Correlation of the amino acid composition of a protein to its structural and biological characteristics. J. Biochem. **91:** 1821–1824
99 Nakashima H. and Nishikawa K. (1992) The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Lett. **303:** 141–146
100 Nakai K. and Kanehisa M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. Proteins **11:** 95–110
101 Nakai K. and Kanehisa M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics **14:** 897–911
102 Horton P. and Nakai K. (1996) Fourth International Conference on Intelligent Systems for Molecular Biology, St. Louis, MO
103 Claros M. G. and Vincens P. (1995) Computational method to predict mitochondrially imported proteins and their transit peptides. Eur. J. Biochem. **241:** 779–786
104 Hua S. and Sun Z. (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics **17:** 721–728
105 Cedano J., Aloy P., Pérez-Pons J. A. and Querol E. (1997) Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. **266:** 594–600
106 Mott R., Schultz J., Bork P. and Ponting C. P. (2002) Predicting protein cellular localization using a domain projection method. Genome Res. **12:** 1168–1174
107 Marcotte E. M., Xenarios I., van Der Bliek A. M. and Eisenberg D. (2000) Localizing proteins in the cell from their phylogenetic profiles. Proc. Natl. Acad. Sci. USA **97:** 12115–12120
108 Nakai K. and Horton P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. **24:** 34–36
109 Nielsen H., Engelbrecht J., Brunak S. and von Heijne G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Prot. Eng. **10:** 1–6
110 Emanuelsson O., Nielsen H. and von Heijne G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. **8:** 978–984
111 Drawid A. and Gerstein M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. J. Mol. Biol. **301:** 1059–1075
112 Nair R. and Rost B. (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. Proteins, in press
113 Garavelli J. S. (2003) The RESID Database of Protein Modifications: 2003 developments. Nucleic Acids Res. **31:** 499–501
114 Ladunga I., Czakó F., Csabai I. and Geszti T. (1991) Improving signal peptide prediction accuracy by simulated neural network. CABIOS **7:** 485–487
115 Schneider G. (1999) How many potentially secreted proteins are contained in a bacterial genome? Gene **237:** 113–121
116 Jagla B. and Schuchhardt J. (2000) Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. Bioinformatics **16:** 245–250
117 Emanuelsson O., Nielsen H., Brunak S. and von Heijne G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. **300:** 1005–1016
118 Nakai K. (2001) Prediction of in vivo fates of proteins in the era of genomics and proteomics. J. Struct. Biol. **134:** 103–116
119 Wrede P., Landt O., Klages S., Fatemi A., Hahn U. and Schneider G. (1998) Peptide design aided by neural networks: biological activity of artificial signal peptidase I cleavage sites. Biochemistry **37:** 3588–3593
120 Kesimir C., Nussbaum A. K., Schild H., Detours V. and Brunak S. (2002) Prediction of proteasome cleavage motifs by neural networks. Prot. Eng. **15:** 287–296
121 Graber J. H., McAllister G. D. and Smith T. F. (2002) Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3′-processing sites. Nucleic Acids Res. **30:** 1851–1858
122 Cai Y. D., Yu H. and Chou K. C. (1998) Artificial neural network method for predicting HIV protease cleavage sites in protein. J. Protein Chem. **17:** 607–615

123 Jarmer H., Larsen T. S., Krogh A., Saxild H. H., Brunak S. and Knudsen S. (2001) Sigma A recognition sites in the *Bacillus subtilis* genome. Microbiology **147:** 2417–2424

124 Nussbaum A. K., Kuttler C., Hadeler K. P., Rammensee H. G. and Schild H. (2001) PAProC: a prediction algorithm for proteasomal cleavages available on the www. Immunogenetics **53:** 87–94

125 Blom N., Gammeltoft S. and Brunak S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. **294:** 1351–1362

126 Kreegipuu A., Blom N. and Brunak S. (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Res. **27:** 237–239

127 Eisenhaber B., Bork P. and Eisenhaber F. (2001) Post-translational GPI lipid anchor modification of proteins in kingdoms of life: analysis of protein sequence data from complete genomes. Prot. Eng. **14:** 17–25

128 Gupta R. and Brunak S. (2002) Prediction of glycosylation across the human proteome and the correlation to protein function. Pac. Symp. Biocomput. 310–322

129 Hansen J., Lund O., Tolstrup N., Gooley A. A., Williams K. L. and Brunak S. (1998) NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. Glycoconjugate J. **15:** 115–130

130 Christlet T. H., Biswas M. and Veluraja K. (1999) A database analysis of potential glycosylating Asn-X-Ser/Thr consensus sequences. Acta Crystallogr. Sect. D Biol. Crystallogr. **55:** 1414–1420

131 Maurer-Stroh S., Eisenhaber B. and Eisenhaber F. (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. J. Mol. Biol. **317:** 541–557

132 Riley M. (1993) Function of the gene products in *Escherichia coli*. Microbiol. Rev. **57:** 862–952

133 Clare A. and King R. D. (2002) Machine learning of functional class from phenotype data. Bioinformatics **18:** 160–166

134 Galperin M. Y. and Koonin E. V. (2000) Who's your neighbor? New computational approaches for functional genomics. Nat. Biotechnol. **18:** 609–613

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

136 Overbeek R., Fonstein M., D'Souza M., Pusch G. D. and Maltsev N. (1999) Use of contiguity on the chromosome to predict functional coupling. In Silico Biol. **1:** 93–108

137 Jensen L. J., Gupta R., Blom N., Devos D., Tamames J., Kesmir C. et al. (2002) Prediction of human protein function from post-translational modifications and localization features. J. Mol. Biol. **319:** 1257–1265

138 Brown D. R. (2002) Copper and prion diseases. Biochem. Soc. Trans. **30:** 742–745

139 Jensen L. J., Skovgaard M. and Brunak S. (2002) Prediction of novel archaeal enzymes from sequence-derived features. Protein Sci. **11:** 2894–2898

140 Jensen L. J., Gupta R., Staerfeldt H. H. and Brunak S. (2003) Prediction of human protein function according to Gene Ontology categories. Bioinformatics **19:** 635–642

141 Uetz P., Giot L., Cagney G., Mansfield T. A., Judson R. S., Knight J. R. et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature **403:** 623–627

142 Xenarios I., Fernandez E., Salwinski L., Duan X. J., Thompson M. J., Marcotte E. M. et al. (2001) DIP: the database of interacting proteins: 2001 update. Nucleic Acids Res. **29:** 239–241

143 Teichmann S. A., Murzin A. G. and Chothia C. (2001) Determination of protein function, evolution and interactions by structural genomics. Curr. Opin. Str. Biol. **11:** 354–363

144 Sheinerman F. B. and Honig B. (2002) On the role of electrostatic interactions in the design of protein-protein interfaces. J. Mol. Biol. **318:** 161–177

145 Marcotte E. M. (2000) Computational genetics: finding protein function by nonhomology methods. Curr. Opin. Str. Biol. **10:** 359–365

146 Mann M., Hendrickson R. C. and Pandey A. (2001) Analysis of proteins and proteomes by mass spectrometry. Annu. Rev. Biochem. **70:** 437–473

147 DeLano W. (2002) Unravelling hot spots in binding interfaces: progress and challenges. Curr. Opin. Str. Biol. **12:** 14–20

148 Michnick S. W. (2001) Exploring protein interactions by interaction-induced folding of proteins from complementary peptide fragments. Curr. Opin. Str. Biol. **11:** 472–477

149 von Mering C., Huynen M., Jaeggi D., Schmidt S., Bork P. and Snel B. (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. **31:** 258–261

150 Ng S. K., Zhang Z., Tan S. H. and Lin K. (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res. **31:** 251–254

151 Bock J. R. and Gough D. A. (2003) Whole-proteome interaction mining. Bioinformatics **19:** 125–134

152 Bader G. D., Betel D. and Hogue C. W. (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res. **31:** 248–250

153 Aloy P. and Russell R. B. (2003) InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics **19:** 161–162

154 Tong A. H., Drees B., Nardelli G., Bader G. D., Brannetti B., Castagnoli L. et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science **295:** 321–324

155 Smith G. R. and Sternberg M. J. (2002) Prediction of protein-protein interactions by docking methods. Curr. Opin. Str. Biol. **12:** 28–35

156 Gavin A. C., Bosche M., Krause R., Grandi P., Marzioch M., Bauer A. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415:** 141–147

157 Ho Y., Gruhler A., Heilbut A., Bader G. D., Moore L., Adams S. L. et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature **415:** 180–183

158 Krauthammer M., Kra P., Iossifov I., Gomez S. M., Hripcsak G., Hatzivassiloglou V. et al. (2002) Of truth and pathways: chasing bits of information through myriads of articles. Bioinformatics **18:** S249–S257

159 Blaschke C. and Valencia A. (2001) The potential use of SUISEKI as a protein interaction discovery tool. Genome Inform. Ser. Workshop Genome Inform. **12:** 123–134

160 Marcotte E. M., Xenarios I. and Eisenberg D. (2001) Mining literature for protein-protein interactions. Bioinformatics **17:** 359–363

161 Gromiha M. M. and Selvaraj S. (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J. Mol. Biol. **310:**: 27–32

162 Xenarios I., Salwinski L., Duan X. J., Higney P., Kim S. M. and Eisenberg D. (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30:** 303–305

163 Blaschke C., Hirschman L. and Valencia A. (2002) Information extraction in molecular biology. Briefing Bioinf. **3:** 154–165

164 Blaschke C., Oliveros J. C. and Valencia A. (2001) Mining functional information associated with expression arrays. Funct. Integr. Genomics **1:** 256–268

165 Friedman C., Kra P., Yu H., Krauthammer M. and Rzhetsky A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics **17:** S74–S82
166 Huynen M., Snel B., Lathe W. and Bork P. (2000) Predicting protein function by genomic context. Genome Res. **4:** 1204–1210
167 Dandekar T., Snel B., Huynen M. and Bork P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. **23:** 324–328
168 Marcotte E. M., Pellegrini M., Ng H. L., Rice D. W., Yeates T. O. and Eisenberg D. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science **285:** 751–753
169 Enright A. J., Ilipoulos I., Kyrpides N. C. and Ouzounis C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature **402:** 86–90
170 Gaasterland T. and Ragan M. A. (1998) Constructing multigenome views of whole microbial genomes. Microb. Comp. Genomics **3:** 177–192
171 Pazos F. and Valencia A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. Prot. Eng. **14:** 609–614
172 Goh C. S., Bogan A. A., Joachimiak M., Walther D. and Cohen F. E. (2000) Co-evolution of proteins with their interaction partners. J. Mol. Biol. **299:** 283–293
173 Goh C. S. and Cohen F. E. (2002) Co-evolutionary analysis reveals insights into protein-protein interactions. J. Mol. Biol. **324:** 177–192
174 Pazos F., Helmer-Citterich M., Ausiello G. and Valencia A. (1997) Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. **271:** 511–523
175 Pazos F. and Valencia A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins **47:** 219–227
176 Sprinzak E. and Margalit H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. J. Mol. Biol. **311:** 681–692
177 Gomez S. M., Lo S. H. and Rzhetsky A. (2001) Probabilistic prediction of unknown metabolic and signal-transduction networks. Genetics **159:** 1291–1298
178 Aloy P. and Russell R. B. (2002) Interrogating protein interaction networks through structural biology. Proc. Natl. Acad. Sci. USA **99:** 5896–5901
179 Lu L., Lu H. and Skolnick J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins **49:** 350–364
180 Nair, R and Rost B. (2002) Inferring sub-cellular localisation through automated lexical analysis. Bioinformatics **18:** S78–S86