Predicting functions from protein sequences—where are the bottlenecks?

Peer Bork[1] & Eugene V. Koonin[2]

The exponential growth of sequence data does not necessarily lead to an increase in knowledge about the functions of genes and their products. Prediction of function using comparative sequence analysis is extremely powerful but, if not performed appropriately, may also lead to the creation and propagation of assignment errors. While current homology detection methods can cope with the data flow, the identification, verification and annotation of functional features need to be drastically improved.

With the rapid growth of sequence-related and other databases, there is increasing concern about the impact of this information explosion[1,2]. Is the burgeoning diversity of information an advantage or will it play havoc with genome analysis and ultimately lead to an error catastrophe? The eventual success of genome projects will depend on our ability to handle information in a manner that enhances the capability for function prediction rather than pollutes the analysis with noise. Only a minority of sequenced genes have been studied in direct experiments. In the foreseeable future, the gap between the number of sequences available and the extent of functional characterization of gene products is expected to broaden even further. It is apparent that computer procedures for the prediction of functional features from sequence are much faster and cheaper than 'wet' experiments and, by default, are applied to each gene that is sequenced. This puts tremendous pressure on computational approaches to ascribe as much functional information as possible to each gene. It appears, however, that within the typical framework of current sequencing projects, optimization of the computer analysis and functional annotation has not yet been achieved. A testimony to this is the repeated discovery of new, functionally relevant features in sequences that already have been subjected to standard computer procedures.

Given the database growth and the accompanying increase in noise and redundancy, we believe that there are currently two major bottlenecks that need to be overcome *en route* to efficient functional predictions from protein sequences. First, there is the lack of a widely accepted, robust and continuously updated suite of sequence analysis methods integrated into a coherent and efficient prediction system. Second, there is considerable 'noise' in the presentation of experimental information, leading to insufficient or erroneous functional assignment in sequence databases.

Here we review some computer-based approaches that allow utilization of more functional and structural information than the current standard schemes, and discuss some of the difficulties in handling and interpreting functional information.

Effects of database growth

From a purely statistical standpoint, the chances of detecting significant similarity between a new sequence and the ones already available in databases decrease with the expansion of the search space, in this case, database growth[3,4]. Fortunately, at least three major factors counter this adverse statistical effect. First, the sequence space is not infinite: new sequences fill it and inevitably increase the chance of finding homologues. Second, complete genome sequences of phylogenetically distant species bring a qualitative improvement to the representation of conserved gene families[5]. With numerous genome sequences of unicellular organisms already available and the majority of human genes represented in the Expressed Sequence Tags (EST) database[6], it is becoming increasingly likely that a family to which a new protein belongs is already represented in the databases[7]. Third, the development of new, more sensitive methods for information filtering and database searching, as well as improved strategies for their application, result in the delineation of previously undetectable, subtle relationships between sequences.

The net effect is that, for a given sequence, the likelihood of detecting a homologue in the databases steadily increases with time. To illustrate this, we followed the kinetics of homology identification and function prediction for an unbiased data set, namely the proteins encoded by the genes on yeast chromosome III. After the initial characterization of this eukaryotic chromosome in 1992 (ref. 8), and early efforts to push the limits of computer-aided predictions[9–11], there has been a continuous linear increase in the fraction of proteins that have identifiable homologues and predictable functions (Fig. 1). Although due in part to a decrease in the number of open reading frames (ORFs) identified as likely genes, this trend demonstrates the increasing utility of computer analysis despite database growth. Of the current set of 25 predicted genes without homology or functional assignment, 15 are smaller than 150 amino acids and may not be expressed at all[12]. With homologues now detectable for 85% of the proteins encoded by genes on yeast chromosome III and at least some functional features identified for 70%, we may soon approach an upper limit for computer-aided predictions. The depth of the functional characterization attainable for many proteins, however, will continue to increase, which is not reflected in the above numbers.

Reducing the noise in sequence searches

In-depth analysis of protein sequences often results in functional predictions not attained in the original studies. This can be illustrated by the results obtained with genes mutated in human diseases (disease genes). Identification of such genes typically involves the time- and labour-consuming process of positional cloning[13]. It is therefore critical that as much functionally relevant information as possible is extracted from the protein sequence encoded by a disease gene once it becomes available. It is not uncommon, however, that rapid computer re-analysis produces unexpected insight into the evolutionary relationships and

[1]*EMBL, Meyerhofstr. 1, 69012 Heidelberg and Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Germany.* [2]*NCBI, NIH, Bethesda, USA. Correspondence should be addressed to P.B. email: bork@embl-heidelberg.de*

Table 1• Selected examples of computer-aided discoveries with positionally cloned disease genes

Gene (protein)	Phenotype/ Disease	Original observations on protein sequence	Novel findings by in-depth computer analysis and implications	Subsequent experimental support
LEP (leptin)	Hereditary obesity	None[23]	Structural similarity to helical cytokines identified by threading; likely cytokine activity and features[24]	Leptin structure is typical of helical cytokines[25]; leptin receptor is homologous and functionally analogous to cytokine receptors[26]
DMD (dystrophin)	Muscular dystrophy	Spectrin repeats[40]	WW, and ZZ signalling domains[41,42]	3D structure of WW and ligand[43]
HD (huntingtin)	Huntington disease	None[44]	HEAT repeats covering a considerable fraction of the protein and indicating structural features[45]	Conservation of repeats in homologues
BRCA1 (BRCA1)	Hereditary breast cancer	N-terminal RING-finger domain[46]	C-terminal BRCT domain conserved in many DNA repair-dependent checkpoint proteins; likely role in cell cycle checkpoints[47–49]	BRCT domain coincides with transcription activation domain[50], BRCA1 involved in repair and cell cycle [51]
BRCA2 (BRCA2)	Hereditary breast cancer	Coiled-coil domains[52]	Previously unknown repeats covering almost a third of this large protein[53]	Verification of the repeats by sequencing homologues[54]
CHM (CHM)	Choroideremia (hereditary blindness)	Homologue of guanine nucleotide dissociation inhibitor Rab-GDI[55]	FAD (NAD)-binding domain[56]	3D structure confirmed presence of a dinucleotide-binding domain[57]
FRDA1 (frataxin)	Freidrich ataxia	Highly conserved eukaryotic homologues[58]	Bacterial homologues (CyaY); on the basis of phylogenetic analysis, a mitochondrial function suggested for frataxin[59]	Mitochondrial localization of frataxin demonstrated[60]
CLCN1 (CLCN1)	Myotonia (Thomsen disease)	Chloride channel[61]	CBS domain[62]	NA
TAZ (tafazzin)	Barth syndrome	None[63]	Acyltransferase domain; possible role in membrane biogenesis[64]	NA
MLH1 (MLH1)	Hereditary non-polyposis colon cancer	Homologue of bacterial and yeast DNA repair protein MutL[65]	ATPase domain conserved in topoisomerase type II (3D structure available), HS90, and His kinases; predicted ATPase activity[14,66]	NA
WRN (WRN)	Werner syndrome (premature aging)	DNA helicase domain (RecQ homologue)[67]	N-terminal exonuclease domain (structure known for homologous domain in bacterial PolA) and putative C-terminal RNA-binding domain conserved in BLM and RNA-ase D; predicted exonuclease activity[14,68,69]	NA
BLM (BLM)	Bloom syndrome	DNA helicase domain (RecQ homologue)[70]	C-terminal nucleic acid-binding domain conserved in WRN and RNAase D[69]	NA
WAS (WAS)	Wiskott-Aldrich syndrome	WH1 (ref. 71)	WH1 in homer with ligand-binding implications[72]	NA
SCA2 (ataxin-2)	Spinocerebral ataxia-2	Polyglutamine stretch expanded in ataxia[73]	Novel conserved domain shared with spliceosomal Sm proteins and bacterial global transcription regulators; possible role in splicing[74]	NA

NA, no data available

likely functions of disease gene products (see Table 1 for selected examples, ref. 14 for a more detailed list). These predictions have the potential to greatly facilitate and accelerate functional characterization of disease genes, and ultimately, the development of therapeutics. Thus, it is instructive to explore the main reasons why these relationships have been missed in the original studies but have been revealed subsequently.

Generally, the problem of obtaining the best results from a database search can be formulated in terms of signal-to-noise ratio[15]. One way to compensate for noise is to increase the sensitivity of the search method. The recently developed PSI-BLAST method (Position-Specific Iterative BLAST; ref. 16) is the latest major step in this direction. It combines the advanced version of the popular BLAST algorithm, which has been modified to incorporate gapped alignments, with profile analysis. PSI-BLAST is fast and highly sensitive and, when combined with appropriate filtering of low complexity regions (see below), also highly selective. Nevertheless, there is still room for improvement as the signal-to-noise ratio can be increased further using protein-family–dependent strategies such as concentrating on motifs, local conserved regions (that is, lowering the signal, but to a lesser extent than the noise) or including global alignments to increasing the signal (but to a greater extent than the noise, due to variable regions; ref. 15).

To reduce the noise, it is advisable first to pre-process the query sequence to account for the different types of compositionally-biased regions. This is achieved by filtering low-complexity regions using the SEG program[17], which is currently implemented as a default parameter in the BLAST programs. Approximately 15% of protein-sequence databases appear to be represented by these low complexity regions, which are thought to correspond to non-globular domains; most of these are found in eukaryotic proteins[18]. Low-complexity sequences tend to produce database hits with artifactually low *P*-values (estimated probability of a random match) simply because there are many sequences with a similarly reduced residue alphabet in the database. For the detection of particular biases such as coiled-coil regions, accurate methods have been developed[19] that also should be included in a pre-processing step. In addition, programs exist for delineating other low-complexity regions such as transmembrane segments[20], although one should always be aware of the prediction accuracy and the choice of parameters that need to be adapted to the specific problem (Table 2). These *caveats* notwithstanding, integration of different pre-processing methods into the search scheme considerably increases the signal-to-noise ratio.

Noise reduction is particularly important when dealing with multidomain proteins. Usually each of the domains not only is a defined structural unit with its own evolutionary history but also carries a distinct subfunction contributing to a more complex overall function of the protein. In a standard database search, the multidomain architecture of proteins causes several problems that frequently hamper extraction of functional information from sequences. First, the domain with the strongest signal will always score highest so that lower-scoring domains in different locations in the query sequence may be overlooked. Second, hits that involve single domains are often misleading with regard to functional characterization. For example, identification of a src-homology domain (SH3) of a src-like protein kinase within a query sequence has often caused mis-annotation of the query as a protein kinase. Third, searching with the entire sequence of a multidomain protein is much less sensitive than searching with segments that are located between known domains. Thus, scanning databases of known domains such as PROSITE, BLOCKS, PRINTS, PFAM or SMART (refs 21,22) is an important complement to standard database searches.

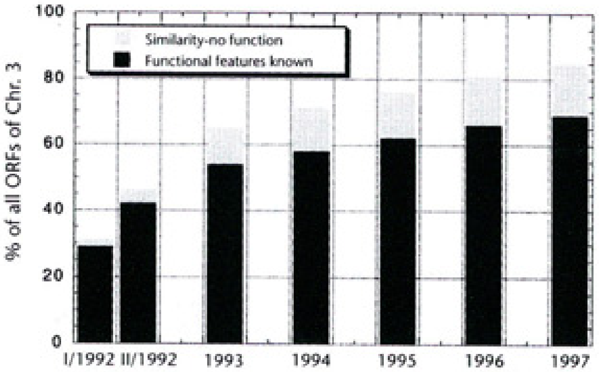

Fig. 1 Increase in knowledge of ORFs encoded by genes on yeast chromosome III, as recorded at the end of each year. Note that the number of ORFs (originally 182 ; ref. 8) is steadily decreasing and might fall below 150 because 16 of the 25 remaining ORFs without homologues are small and have unusual amino-acid compositions. On the other hand, an ORF with a homologue in another species (85% of the total) is very likely to encode an expressed protein. The term 'function' is loosely defined and we have included even weak functional features such as the presence of an SH3 domain or a defined ATP-binding motif, which hint at cellular or molecular roles. There are currently only three proteins encoded by genes on chromosome III which have a described function but lack protein homology.

Multidomain architecture is most common in eukaryotic proteins—and the majority of disease gene products contain multiple domains[14,22]. Indeed, computational re-analysis has revealed numerous undetected domains in disease genes, which suggest additional functional and structural features (Table 1).

In the absence of recognizable sequence similarity, 'threading' approaches—that is, fold assignments by means of checking for sequence compatibility with known three-dimensional structures—might reveal additional structural insights, as has been successfully demonstrated for leptin, the obesity gene product[23]. The prediction of a helical cytokine-like fold[24] in leptin has been confirmed both by direct structure determination[25] and by numerous functional studies[26]. However, the accuracy of threading methods is limited[27,28], and the sensitivity of the most powerful modern methods for sequence similarity analysis approaches that of threading. For example, highly accurate fold assignments for more than 35% of all *Mycobacterium genitalium* proteins (or some of their domains) are currently possible using iterative homology searches (Huynen, M.A. *et al.*, submitted). This fraction includes the 8% which previously has been claimed to be identifiable only by using threading techniques[29].

Effects of noise on functional predictions

Functional information is hard to quantify and a critical point in an analysis is determining how many functional features are shared between the protein in question and sequence(s) similar to it in the database. In the best case, they are orthologues[30,31], corresponding genes in two species that perform the same molecular function. Even this, however, is often difficult to discern in a situation where the database hits are paralogues (other members of a multigene family which have distinct, but perhaps related, functions). With many complete genome sequences available, databases of orthologues are expected to become indispensable for functional annotation[5]. However, more often than not, it is clear that the cellular role of the protein in question differs from that of the detected homologue(s) and there is currently no automatic means to establish how much functional information can be legitimately transferred by analogy from the homologue to the query[1].

As the problem of orthology *versus* paralogy has only recently been introduced into genome analysis, many erroneous func-

Fig.2 Sequence alignment of a selected set of SAM domains with p73-like proteins. Conserved hydrophobic positions are shown in bold; residues that are conserved in at least 75% of the sequences are highlighted. The SAM consensus[33] corresponds to the consensus of the alignment. Secondary structure predictions ('a' represents 'alpha helix') have been taken from ref. 33. Position in the sequence and database accession numbers are given in the third and last column, respectively.

```
Ste11        yeast       17  EKTNDLPFVQLFLEEIGCT-QYLDSFIQCNLVTEEEIKYLDKDILIA-LGVNKIGDRLKILRKSKSFQR  P23567
DGKd         human     1100  HLWG-TEEVAAWLEHLSLC-EYKDIFTRHDIRG-SELLHLERRDLKD-LGVTKVGHMKRILCGIKELSR  D73409
51CP         human       60  SGLG-EAGMSAWLRAIGLE-RYEEGLVHNGWDDLEFLSDITEEDLEE-AGVQDPAHKRLLLDTLQ-LSK  L24444
Bicaudal-c   fly        800  MQLAKHKDIQTLLTSLGLE-HYIKIFVLNEIDL-EVFTTLTEENLME-LGIAAFGARKKLLTAIHTLLA  U15928
Byr2         f.yeast      2  EYYT-SKEVAEWLKSIGLE-KYIEQFSQNNIEG-RHLNHLTLPLLKD-LGIENTAKGKQFLKQRDYLRE  P28829
Ste4         f.yeast      9  WNWN-NEAVCNWIEQLGFP--HKEAFEDYHILG-KDIDLLSSNDLRD-MGIESVGHRIDILSAIQSMKK  P36622
Bob1         yeast      226  KSWS-PEEVTDYFSLVGFDQSTCNKFKEHQVSG-KILLELELEHLKE-LEINSFGIRFQIFKEIRNIKS  P38041
RhoGAP       rat          4  TQIE-AKEACDWLRVTGFP-QYAQLYEDLLFPI-DIALVKREHDFLDRDAIEALCRRLNTLNKCAVMKL  D31962
Lmep         mouse      936  LDFPCLDSPQAWLSAIGLE-CYQDNFSKFGLSTFSDVAQLSLEDLPG-LGITLAGHQKKLLHNIQLLQQ  L77867
SAM  consensus                 h     h  WL  hGh   Yh  F    h     h  h   DL   h I     R  hh  h  h
Sqp53        sqid       452  EPTE--NTIAQWLTKLGLQ-AYIDNFQQKGLHNMFQLDEFTLEDLQS-MRIG-TGHRNKIWKSLLDYRR  U43595
Ket          rat        495  PPYPTDCSIVSFLARLGCS-SCLDYFTTQGLTTIYQIEHYSMDDLAS-LKIP-EQFRHAIWKGILDHRQ  Y10258
p73          human      484  PPYHADPSLVSFLTGLGCP-NCIEYFTSQGLQSIYHLQNLTIEDLGA-LKIP-EQYRMTIWRGLQDLKQ  Y11416
2D prediction                       aaaaaaaaaaa  aaaaaaaaaa    aaaaaaaaaaaaa       aaaaaaaaaaaaa
```

tional annotations have already been incorporated and subsequently propagated in sequence databases. A quality index for functional annotation in databases still remains a distant goal and new approaches are required to improve the sensitivity of functional characterization so as to avoid functional over- and under-predictions for a given database match.

Transfer of functional information from sequences in the database to the query is also hampered by the effects of noise in the functional description of proteins. Updates on functional features require an awareness of the scientific literature; an experiment in one species on a previously sequenced gene brings important consequences for homologous genes in other species. This has led to a gap between functional information contained in the sequence databases and the specialized knowledge embodied in the literature. At present, there is no automatic method that can replace literature searches. A recent case in point is p73, a human paralogue of the tumour suppressor p53. It contains a carboxy-terminal extension, for which similarity has only been found in squid 'p53' (ref. 32). However, relevant and potentially important information about this region could be obtained by simply searching the PubMed database for the combination of terms 'p53' and '*Loligo*' (Latin for squid). It has been shown that the squid p53 homologue contains a C-terminal SAM domain, a distinct protein-protein interaction and dimerization domain found primarily in developmental regulators[33]. Furthermore, just weeks after the publication of human p73, the gene encoding rat KET protein, a close p73 homologue was sequenced[34]. With the KET sequence in the database, a PSI-BLAST search using the conservation profile of the two p73 species readily reveals significant similarity between their C-terminal regions and numerous SAM domains (Fig. 2). The SAM domain in p73 may be involved in dimerization and may mediate an interaction with another protein(s) involved in transcription regulation and developmental control of gene expression. A more thorough literature search or the use of 'awareness' tools (see below) would have helped to retrieve, in an automatic manner, potentially important information on p73.

Toward integration of functional and structural features

In summary, the currently available methods for sequence analysis are sophisticated, and while further improvements will certainly ensue, they are already capable of extracting subtle but functionally relevant signals from protein sequences. Whether or not a researcher actually reaps maximal benefit from such analysis, however, depends on the application of an appropriate combination of methods in the correct setting. There is no single, universal recipe for this purpose, but we have attempted to compile a short checklist, which will minimize the risk of missing important functional signals hidden in protein sequences (Table 2).

A comprehensive, precisely defined and standardized classification of biological functions is required for the automation of the prediction of gene functions. The task of constructing such a classification is immensely difficult given that even small proteins—not to mention large, multidomain ones—are certain to have multiple roles in the cell. Classification schemes that have been proposed for prokaryotic gene products[35–37] are useful in comparative genome analysis but grossly oversimplify the problem. We believe that, at present, there is still no proper language for the adequate and uniform description of functions and therefore it is hard to predict when, if at all, a (a)periodic system of biological functions may become a reality.

Generally, the incorporation of known functional information into databases at various levels is a pressing need, requiring the combined efforts of experimentalists, computational biologists and database developers. The challenge is particularly formidable as new types of information, such as tissue- and organ-specific gene expression patterns on a genome scale as well as numerous data on protein interactions and post-translational modifications are rapidly becoming available (for details see refs 38,39).

This continuous flow of information also requires 'update' and 'awareness' tools that filter and incorporate incoming data (sequences, literature, *etcetera*) and new applications (servers, methods). Ideally, these tools will notify researchers upon subscription using a customized query profile (keywords, sequences of interest, problem description) or systematically integrate the filtered information into dynamic databases. The major problem that remains is the quality of the information, as no data-mining tools yet exist that can judge the various experimental protocols described in the respective literature.

There is little doubt that a new generation of bioinformatics approaches will soon integrate sequence and structure analysis methods with awareness tools, information filters and dynamic data processing in preparation for the forthcoming postgenomics aera. However, all these tools will only facilitate but not replace the work of a scientist who defines the questions and interprets the results.

Note added in proof: Elizabeth Greene & Steven Henikoff have constructed an article for the *Nature Genetics* website (http://www.genetics.nature.com/gazing/), which provides direct links to a range of web-based tools for database sequence searches, and homology and structural predictions for query protein sequences.

Table 2 • A checklist for in-depth analysis of protein sequences and prediction of function from sequence[a]

Procedure	Purpose and comment
Identification of and filtering for structural features	
Mask non-globular or highly compositionally-biased regions (reduced residue alphabet)	Reduce noise and avoid spurious hits due to low complexity regions
Mask coiled-coil regions	This is a special type of low-complexity region that causes numerous other coiled-coil regions to match but is not efficiently detected by general methods for complexity analysis
Identify transmembrane regions (including signal sequences and GPI anchors)	Yet another form of composition bias that may result in matches with non-homologous membrane regions; the presence of these domains should be taken into account in all functional predictions.
Identify internal repeats	Reduces the search space for remaining parts and also may lead to the detection of novel repeat types
Predict secondary structure	The best programs predict a protein's structural class; use of multiple alignments significantly improves the accuracy of prediction
Identification of homologues	
Identify known domains in dedicated databases (for example, Pfam, PROSITE, BLOCKS, PRINTS, SMART) prior to a BLAST search	Identification of annotated domains may be more sensitive when these databases are used; removal of known domains also reduces the search space for remaining parts of the protein.
Search complete sequence databases with subsequences of long (>200 a.a.) proteins individually; preferably use subsequences separated by known domains or low-complexity regions	Increase search sensitivity by reducing the search space; exclude domains with numerous homologues (for example, protein kinases), which may obscure even highly significant similarity to other domains of the query
Perform reciprocal searches to verify weak similarity to possible homologues	The alignments of a potentially relevant database hit with its indisputable homologues support (if the conservation pattern is consistent) or reject (if it is different) a weak pairwise similarity.
Perform exhaustive, iterative database searches	Database search methods are non-transitive and non-symmetric; therefore analysis of a protein family should be performed iteratively, starting with different members, until no new homologues are detected
Combine search for pairwise sequence similarity (for example, first BLAST scan) with profile, motif, and pattern searches (explicit example in Psiblast[16]) or by using the various programs available[15]	The information contained in a multiple alignment provides for amplification of weak but potentially important sequence signals and is indispensable for the delineation of protein superfamilies.
Prediction of protein functions	
Carefully consider domain organization and distinct functions of individual domains	Many proteins are multifunctional; assignment of a single function, which is still common in genome projects, results in loss of information and outright errors
Do not take database annotation for granted, especially if only one homologue is detectable or there is inconsistency between different homologues	Databases contain a number of incorrect annotations due to experimental errors as well as functional assignments on the basis of dubious sequence similarity
Do not simply transfer functional information from the best hit	The best hit is frequently hypothetical or poorly annotated; other hits with similar or even lower scores may be more informative; even the best hit may have a different function (see below)
Do cluster analysis of the homologues to identify the appropriate level of precision for functional prediction	It is typical that the general function of a protein can be identified easily but the prediction of substrate specificity is unwarranted; for example, many permeases of different specificity show approximately the same level of similarity to each other
Check sequence context (e.g. likely clashes in the co-occurrence of a signal sequence and a zinc-finger or glycolysation sites in a cytoplasmic protein)	Comparison of different predicted structural and functional features helps avoiding erroneous predictions
Identify similarities to proteins with known 3D structure	Models of highly conserved homologues can be built and might reveal further functional insights

[a] Complementary checklists can be found in ref. 15

Fig.2 Sequence alignment of a selected set of SAM domains with p73-like proteins. Conserved hydrophobic positions are shown in bold; residues that are conserved in at least 75% of the sequences are highlighted. The SAM consensus[33] corresponds to the consensus of the alignment. Secondary structure predictions ('a' represents 'alpha helix') have been taken from ref. 33. Position in the sequence and database accession numbers are given in the third and last column, respectively.

```
Ste11       yeast      17  EKTNDLPFVQLFLEEIGCT-QYLDSFIQCNLVTEEEIKYLDKDILIA-LGVNKIGDRLKILRKSKSFQR  P23567
DGKd        human    1100  HLWG-TEEVAAWLEHLSLC-EYKDIFTRHDIRG-SELLHLERRDLKD-LGVTKVGHMKRILCGIKELSR  D73409
51CP        human      60  SGLG-EAGMSAWLRAIGLE-RYEEGLVHNGWDDLEFLSDITEEDLEE-AGVQDPAHKRLLLDTLQ-LSK  L24444
Bicaudal-c  fly       800  MQLAKHKDIQTLLTSLGLE-HYIKIFVLNEIDL-EVFTTLTEENLME-LGIAAFGARKKLLTAIHTLLA  U15928
Byr2        f.yeast     2  EYYT-SKEVAEWLKSIGLE-KYIEQFSQNNIEG-RHLNHLTLPLLKD-LGIENTAKGKQFLKQRDYLRE  P28829
Ste4        f.yeast     9  WNWN-NEAVCNWIEQLGFP--HKEAFEDYHILG-KDIDLLSSNDLRD-MGIESVGHRIDILSAIQSMKK  P36622
Bob1        yeast     226  KSWS-PEEVTDYFSLVGFDQSTCNKFKEHQVSG-KILLELELEHLKE-LEINSFGIRFQIFKEIRNIKS  P38041
RhoGAP      rat         4  TQIE-AKEACDWLRVTGFP-QYAQLYEDLLFPI-DIALVKREHDFLDRDAIEALCRRLNTLNKCAVMKL  D31962
Lmep        mouse     936  LDFPCLDSPQAWLSAIGLE-CYQDNFSKFGLSTFSDVAQLSLEDLPG-LGITLAGHQKKLLHNIQLLQQ  L77867
SAM   consensus              h     h  WL  hGh   Yh  F    h     h  h   DL   h I     R  hh  h  h
Sqp53       sqid      452  EPTE--NTIAQWLTKLGLQ-AYIDNFQQKGLHNMFQLDEFTLEDLQS-MRIG-TGHRNKIWKSLLDYRR  U43595
Ket         rat       495  PPYPTDCSIVSFLARLGCS-SCLDYFTTQGLTTIYQIEHYSMDDLAS-LKIP-EQFRHAIWKGILDHRQ  Y10258
p73         human     484  PPYHADPSLVSFLTGLGCP-NCIEYFTSQGLQSIYHLQNLTIEDLGA-LKIP-EQYRMTIWRGLQDLKQ  Y11416
2D prediction                      aaaaaaaaaaa  aaaaaaaaaa     aaaaaaaaaaaaa       aaaaaaaaaaaaa
```

tional annotations have already been incorporated and subsequently propagated in sequence databases. A quality index for functional annotation in databases still remains a distant goal and new approaches are required to improve the sensitivity of functional characterization so as to avoid functional over- and under-predictions for a given database match.

Transfer of functional information from sequences in the database to the query is also hampered by the effects of noise in the functional description of proteins. Updates on functional features require an awareness of the scientific literature; an experiment in one species on a previously sequenced gene brings important consequences for homologous genes in other species. This has led to a gap between functional information contained in the sequence databases and the specialized knowledge embodied in the literature. At present, there is no automatic method that can replace literature searches. A recent case in point is p73, a human paralogue of the tumour suppressor p53. It contains a carboxy-terminal extension, for which similarity has only been found in squid 'p53' (ref. 32). However, relevant and potentially important information about this region could be obtained by simply searching the PubMed database for the combination of terms 'p53' and '*Loligo*' (Latin for squid). It has been shown that the squid p53 homologue contains a C-terminal SAM domain, a distinct protein-protein interaction and dimerization domain found primarily in developmental regulators[33]. Furthermore, just weeks after the publication of human p73, the gene encoding rat KET protein, a close p73 homologue was sequenced[34]. With the KET sequence in the database, a PSI-BLAST search using the conservation profile of the two p73 species readily reveals significant similarity between their C-terminal regions and numerous SAM domains (Fig. 2). The SAM domain in p73 may be involved in dimerization and may mediate an interaction with another protein(s) involved in transcription regulation and developmental control of gene expression. A more thorough literature search or the use of 'awareness' tools (see below) would have helped to retrieve, in an automatic manner, potentially important information on p73.

Toward integration of functional and structural features

In summary, the currently available methods for sequence analysis are sophisticated, and while further improvements will certainly ensue, they are already capable of extracting subtle but functionally relevant signals from protein sequences. Whether or not a researcher actually reaps maximal benefit from such analysis, however, depends on the application of an appropriate combination of methods in the correct setting. There is no single, universal recipe for this purpose, but we have attempted to compile a short checklist, which will minimize the risk of missing important functional signals hidden in protein sequences (Table 2).

A comprehensive, precisely defined and standardized classification of biological functions is required for the automation of the prediction of gene functions. The task of constructing such a classification is immensely difficult given that even small proteins—not to mention large, multidomain ones—are certain to have multiple roles in the cell. Classification schemes that have been proposed for prokaryotic gene products[35–37] are useful in comparative genome analysis but grossly oversimplify the problem. We believe that, at present, there is still no proper language for the adequate and uniform description of functions and therefore it is hard to predict when, if at all, a (a)periodic system of biological functions may become a reality.

Generally, the incorporation of known functional information into databases at various levels is a pressing need, requiring the combined efforts of experimentalists, computational biologists and database developers. The challenge is particularly formidable as new types of information, such as tissue- and organ-specific gene expression patterns on a genome scale as well as numerous data on protein interactions and post-translational modifications are rapidly becoming available (for details see refs 38,39).

This continuous flow of information also requires 'update' and 'awareness' tools that filter and incorporate incoming data (sequences, literature, *etcetera*) and new applications (servers, methods). Ideally, these tools will notify researchers upon subscription using a customized query profile (keywords, sequences of interest, problem description) or systematically integrate the filtered information into dynamic databases. The major problem that remains is the quality of the information, as no data-mining tools yet exist that can judge the various experimental protocols described in the respective literature.

There is little doubt that a new generation of bioinformatics approaches will soon integrate sequence and structure analysis methods with awareness tools, information filters and dynamic data processing in preparation for the forthcoming postgenomics aera. However, all these tools will only facilitate but not replace the work of a scientist who defines the questions and interprets the results.

Note added in proof: Elizabeth Greene & Steven Henikoff have constructed an article for the *Nature Genetics* website (http://www.genetics.nature.com/gazing/), which provides direct links to a range of web-based tools for database sequence searches, and homology and structural predictions for query protein sequences.

1. Bork, P. & Bairoch, A. Go hunting in sequence databases but watch out for the traps. *Trends Genet.* **12**, 425–427 (1996).
2. Bhatia, U., Robison, K. & Gilbert, W. Dealing with database explosion: a cautionary note. *Science* **276**, 1724–1725 (1997).
3. Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119–129 (1994).
4. Smith, R.F. Sequence database searching in the era of large-scale genomic sequencing. *Genome Res.* **6**, 653–660 (1996).
5. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
6. Boguski, M.S., Tolstoshev, C.M. & Bassett, D.E. Jr. Gene discovery in dbEST. *Science* **265**, 1993–1994 (1994).
7. Green, P., Lipman, D., Hillier, L., Waterstone, R., States, D. & Claverie, J.-M. Ancient conserved regions in new gene sequences and the protein databases. *Science* **259**, 1711–1716 (1993).
8. Oliver, S.G. *et al.* The complete sequence of yeast chromosome III. *Nature* **357**, 38–46 (1992).
9. Bork, P. *et al.* What's in a genome? *Nature* **358**, 287 (1992).
10. Sharp, P.M. & Lloyd, A.T. Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. *Nucleic Acids Res.* **21**, 179–183 (1993).
11. Koonin, E.V., Bork, P. & Sander, C. Yeast chromosome III: New gene functions. *EMBO J.* **13**, 493–503 (1994).
12. Fickett, J.W. ORF's and genes: how strong a connection? *J. Comput. Biol.* **2**, 117–123 (1995).
13. Collins, F. S. Positional cloning from moves from perditional to traditional. *Nature Genet.* **9**, 347–350 (1995).
14. Mushegian, A.R., Bassett, D.E. Jr, Boguski, M., Bork, P. & Koonin, E.V. Positionally cloned human disease genes: Patterns of evolutionary conservation. *Proc. Natl. Acad. Sci. USA* **94**, 5831–5836 (1997).
15. Bork, P. & Gibson, T.J. Applying motif and profile searches. *Methods Enzymol.* **266**, 162–184 (1996).
16. 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).
17. Wootton, J.C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571 (1996).
18. Wootton, J.C. Sequences with unusual amino acid composition. *Curr. Opin. Struct. Biol.* **4**, 413–421 (1994).
19. Lupas, A. Predicting coiled coil regions in proteins *Curr. Opin. Struct. Biol.* **7**, 388–393 (1997).
20. Rost, B. & O'Donoghue, S. Sysisphus and the prediction of protein structure *Comput. Appl. Biosci.* **13**, 345–356 (1997).
21. Henikoff, S. *et al.* Gene families: the taxonomy of protein paralogues and chimeras. *Science* **278**, 609–613 (1997).
22. Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. SMART, a simple modular architecture research tool: Identification of signalling domains *Proc. Natl. Acad. Sci. USA*, in press.
23. Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–432 (1994).
24. Madej, T., Boguski, M.S. & Bryant, S.H. Threading analysis suggests that the obese gene product may be a helical cytokine. *FEBS Lett.* **373**, 13–18 (1995).
25. Zhang F. *et al.* Crystal structure of the obese protein leptin-E100. *Nature* **387**, 206–209 (1997).
26. Tartaglia, L.A. The leptin receptor. *J. Biol. Chem.* **272**, 6093–6096 (1996).
27. Rost, B., Schneider, R. & Sander, C. Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471–480 (1997).
28. Smith, T.F. *et al.* Current limitations to protein threading approaches. *J. Comput. Biol.* **4**, 217–225 (1997).
29. Fischer, D. & Eisenberg, D. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl. Acad. Sci. USA* **94**, 11929–11934 (1997).
30. Fitch, W.M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113 (1970).
31. Fitch, W.M. Uses for evolutionary trees. *Phil. Trans. R. Soc. Lond. B.* **349**, 93–102 (1995).
32. Kaghad, M. *et al.* Monoallelically expressed gene related to p53 at 1p36, a region frequently deleted in neuroblastoma and other human cancers. *Cell* **90**, 809–819 (1997).
33. Schultz, J., Ponting, C.P., Hofmann, K. & Bork, P. SAM as a protein interaction domain involved in developmental regulation. *Prot. Sci.* **6**, 249–253 (1997).
34. Schmale, H. & Bamberger, C. A novel protein with strong homology to the tumor surpressor p53. *Oncogene* **15**, 1363–1367 (1997).
35. Riley, M. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57**, 862–952 (1993).
36. Bork, P. *et al.* Exploring the *Mycoplasma capricolum* genome: a minimal cell reveals its physiology. *Mol. Microbiol.* **16**, 955–967 (1995).
37. Tatusov, R.L. *et al.* Metabolism and evolution of *Haemophilus influenzae* deduced from a whole genome comparison to *Escherichia coli*. *Curr. Biol.* **6**, 279–291 (1996).
38. Hieter, P. & Boguski, M. functional genomics: It's all how you read it. *Science* **278**, 601–602 (1997).
39. Zhang, L. *et al.* Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272 (1997).
40. Koenig, M., Monaco, A.P. & Kunkel, L.M. The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein. *Cell* **53**, 219–226 (1988).
41. Bork, P. & Sudol, M. The WW domain: a signalling site in dystrophin? *Trends Biochem. Sci.* **19**, 531–533 (1994).
42. Ponting, C.P., Blake, D.J., Davies, K.E., Kendrick-Jones, J. & Winder, S.J. ZZ and TAZ: new putative zinc fingers in dystrophin and other proteins. *Trends Biochem. Sci.* **21**, 11–13 (1996).
43. Macias, M.J. *et al.* Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature* **382**, 646–649 (1996).
44. Huntington's disease collaborative research group. A novel gene containing a trinucleotide repeat that is expanded and unstable in huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
45. Andrade, M. & Bork, P. HEAT repeats in Huntington's disease protein. *Nature Genet.* **11**, 115–116 (1995).
46. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71 (1994).
47. Koonin, E.V., Altschul, S. & Bork, P. BRCA1 protein products: functional motifs. *Nature Genet.* **13**, 266–268 (1996).
48. Bork, P. *et al.* A superfamily of conserved domains in DNA damage responsive cell cycle checkpoint proteins. *FASEB J.* **11**, 68–76 (1997).
49. Callebaut, I. & Mornon, J.P. From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair. *FEBS Lett.* **400**, 25–30 (1997).
50. Monteiro, A.N.A., August, A. & Hanafusa, H. Evidence for a transcriptional activation function of BRCA1 C-terminal region. *Proc. Natl. Acad. Sci. USA* **93**, 13595–13599 (1996).
51. Scully, R. *et al.* Dynamic changes of BRCA1 subnuclear location and phosphorylation state are initiated by DNA damage. *Cell* **90**, 425–435 (1997).
52. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792 (1995).
53. Bork, P., Blomberg, N. & Nilges, M. Internal repeats in the BRCA2 protein sequence. *Nature Genet.* **13**, 22–23 (1996).
54. Bignell, G., Micklem, G., Stratton, M.R., Ashworth, A. & Wooster, R. The BRC repeats are conserved in mammalian BRCA2 proteins. *Hum. Mol. Genet* . **6**, 53–58 (1997).
55. Cremers, F.P. *et al.* Cloning of a gene that is rearranged in patients with choroideraemia. *Nature* **347**, 674–677 (1990).
56. Koonin, E.V. Human choroideraemia protein contains an FAD-binding domain. *Nature Genet.* **12**, 237–239 (1996).
57. Wu, S.K., Zeng, K., Wilson, I.A. & Balch, W.E. Structural insights into the function of the Rab GDI superfamily. *Trends Biochem. Sci.* **21**, 472–476 (1996).
58. Campuzano, V. *et al.* Freidrich's ataxia: autosomal recessive disease caused by an intronic GAA repeat expansion. *Science* **271**, 1423–1427 (1996).
59. Gibson, T., Koonin, E.V., Musco, G. Pastore, A. & Bork, P. Freidrich's ataxia protein: phylogenetic evidence for mitochondrial dysfunction. *Trends Neurosci.* **19**, 465–468 (1996).
60. Koenig, M. & Mandel, J.-L. Deciphering the cause of Freidrich's ataxia. *Curr. Opin. Neurobiol.* **7**, 689–694 (1997).
61. Koch, M.C. *et al.* The skeletal muscle chloride channel in dominant and recessive myotonia. *Science* **257**, 797–800 (1992).
62. Bateman, A. The structure of a domain common to archebacteria and the homocystinuria disease protein. *Trends Biochem. Sci.* **22**, 12–13 (1997).
63. Bione, S. *et al.* A novel X-linked gene, G4.5 is responsible for Barth syndrome. *Nature Genet.* **12**, 385–389 (1996).
64. Neuwald, A.F. Barth syndrome might be due to acyltransferase deficiency. *Curr. Biol.* **7**, 465–466 (1997).
65. Kolodner, R. *et al.* Biochemistry and genetics of eukaryotic mismatch repair. *Genes Dev.* **10**, 1433–1442 (1996).
66. Bergerat, A. *et al.* An atypical topoisomerase II from Archea with implications for meiotic recombination. *Nature* **386**, 414–417 (1997).
67. Yu, C.E. *et al.* Positional cloning of the Werner's syndrome gene. *Science* **272**, 258–262 (1996).
68. Mian, I.S. Comparative sequence analysis of ribonucleases HII, II, II PH and D. *Nucleic Acids Res.* **25**,3187–3195 (1997).
69. Morozov, V., Mushegian, A.R., Koonin, E.V. & Bork, P. A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases. *Trends Biochem. Sci.* **22**, 417–418 (1997).
70. Ellis, N.A. *et al.* The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* **83**, 655–666 (1995).
71. Symons, M. *et al.* Wiskott-Aldrich syndrome protein, a novel effector for the GTPase CDC42Hs, is implicated in actin polymerization. *Cell* **84**, 723–734 (1996).
72. Ponting, C.P. & Phillips, C. Identification of homer as a homologue of the Wiskott-Aldrich syndrome protein suggests a receptor-binding function for WH1 domains. *J. Mol. Med.* **75**, 769–771 (1997).
73. Imbert, G. *et al.* Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nature Genet.* **14**, 285–291 (1996).
74. Neuwald, A.F. & Koonin, E.V. Ataxin-2, global regulators of bacterial gene expression, and spliceosomal snRNP proteins share a conserved domain. *J. Mol. Med.* **76**, 3–5 (1998).