From genome sequences to protein function

Peer Bork, Christos Ouzounis and Chris Sander

EMBL, Heidelberg, Germany

A major goal of genome sequencing projects is the complete description of the function of all proteins. For most proteins sequenced in genome projects, an experimentally determined function is not available. Fortunately, evolutionary relationships can be exploited to predict the function of many other proteins from their amino acid sequence. The techniques for such predictions, sequence analysis by computational and database methods, are becoming increasingly sophisticated and are now an essential part of genome analysis.

Current Opinion in Structural Biology 1994, **4**:393–403

Introduction

Today, a sound PhD project in molecular biology often ends with the biochemical characterization and analysis of a cloned and subsequently sequenced gene. In the hope of additional functional insights, as well as interesting structural and evolutionary relationships, the sequence of the gene is usually further characterized by various computer methods. Often, the homologies found are extremely helpful for functional prediction. Today, the analysis of an average gene of about 1 kilobase (kb), including open reading frame (ORF) prediction, database searches, multiple alignment, pattern definition, and profile searches, might take several days or more, depending on the depth of the analysis. Soon genome projects will produce several hundred kilobases of raw sequences a day. Traditional sequence-analysis procedures cannot adequately handle such a high rate of production. The resulting dual challenges facing genome sequence analysis are both quantitative and qualitative: developing more efficient tools and increasing the scope of functional prediction. The work reviewed here represents the first steps on the way to meeting these challenges.

Tip of the iceberg: genomic sequence data today

In spite of the increased production of sequence data, today's databases contain only a small fraction of the genome's complete sequence. Many eukaryotic genome projects, including the Human Genome Initiative, are currently still assembling high-resolution physical maps [1,2], an essential prerequisite of systematic large-scale sequencing. So the real flood of genomic data is still to come.

Current output of genome sequencing projects

The most significant stretches of genomic sequences known to date are sizeable parts of relatively small genomes, such as the nematode *Caenorhabditis elegans*, the yeast *Saccharomyces cerevisiae*, and the bacterium *Escherichia coli* (Tables 1 and 2). In addition, there exist collections of numerous smaller chromosome segments from many different genomes, not yet assembled into long continuous stretches. The largest continuous sequence determined so far is a 2 Megabase (Mb) stretch of chromosome III of *C. elegans* [3], with another 5 Mb expected later in 1994 (R Durbin, personal communication). The next-largest pieces come from the European yeast genome project [4]: the complete sequences of five out of sixteen chromosomes will be available by the end of 1994 (Table 1), giving a total of 2.3 Mb, or 15% of the yeast genome. Bacterial genomes are next, with several large *E. coli* segments [5–7,8•,9,10] resulting in continuous pieces of up to 500 kb (Table 2). The total *E. coli* sequence data already covers more than 60% of the genome (P Rice, personal communication).

In less than 5 years, the genome projects of invertebrates such as the nematode [11], plants such as cress [12], vertebrates such as the buffer fish [13•], and mammals such as human [1], or mouse [2, 14] will probably generate hundreds of Mb of sequence data. Remarkably, the plans for genomic sequencing efforts appear to be on, or ahead of, schedule (Table 2) and are likely to meet the projected completion dates, provided the planned increases in funding are forthcoming. Completion of the *E. coli* and yeast sequencing efforts is expected by 1996. The *C. elegans* genome will be complete by the end of this decade (Table 2). In the human genome project, the initial estimates for completing sequencing were revised downward, from 2020 to 2010 and, most recently, to 2005 [1]. The methodologies for analyzing these data and, in particular, the

Abbreviations

3D—three-dimensional; **EST**—expressed sequence tag; **kb**—kilobase; **Mb**—megabase; **ORF**—open reading frame.

© Current Biology Ltd ISSN 0959-440X

Table 1. Status of sequencing projects of model genomes.

Species	Size (Mb)	Sequenced (Mb)[a]	Sequenced (% of total)[a]	Year of completion[b]
Organelles/Viruses				
Mitochondria (various)	0.01–0.19		100%	1981
Chloroplasts (various)	0.12–0.16		100%	1986
Vaccinia virus	0.19		100%	1991
Prokaryotes				
Mycoplasmas	0.6–1.4	0.3	30%	(1996)
Mycobacterium leprae	2.8	0.1	4%	
Bacillus subtilis	4.2	2	40%	
Escherichia coli	4.7	3	60%	(1996)
Eukaryotes				
Yeast (*Saccharomyces cerevisiae*)	15	5	30%	(1998)
Chromosome III	0.31		100%	1992
Chromosome XI	0.66		100%	1994
Chromosome II	0.84		100%	1994
Cress (*Arabidopsis thaliana*)	100	1	1%	
Nematode (*Caenorhabditis elegans*)	90	3	3%	(1998)
Fruit fly (*Drosophila melanogaster*)	170	3	2%	
Mouse (*Mus musculus*)	3000	9	0.3%	
Human (*Homo sapiens*)	3600	19	0.6%	(2005)

[a]Note that the values are taken from the EMBL nucleotide sequence database; redundancies (only partly excluded) would lead to a decrease, while expressed sequence tags (not included) and data not yet released would lead to an increase in the figures given here. [b]Estimated years of completion are in parentheses to indicate the approximate nature of the projections.

50 000–100 000 human genes will have to follow at the same pace!

The fast track to protein sequences: the EST approach

An attractive alternative to continuous sequencing is the random sequencing of reverse transcribed messenger RNA (cDNA) fragments, often called 'expressed sequence tags' (ESTs) [15]. This approach is 'quick and dirty' in that it is initially limited to single sequencing runs (no verification, no extension). However, it is ideal for very rapid identification of gene products as a first step toward elucidation of gene function, a major goal of genome projects. ESTs correspond to protein fragments of about 100 amino acids and can be used to obtain an expression profile for a particular organism or a tissue, to identify exons or to provide a glimpse into the molecular repertoire of various organisms. Already, a sizeable fraction of the ESTs sequenced have sequence similarity to proteins of known function. In 1991 only a few hundred ESTs from human brain were known [15]; these increased to more than 6000 human ESTs by 1993 [16••], and probably up to 100 000 sequences from several organisms will be available by the end of 1994. The rate of data production is so high that ESTs corresponding to most of the highly expressed human genes will probably approach saturation in 1995, many of them in private commercial efforts. Publicly available ESTs from various organisms are now in a specialized database, called dbEST [17].

Rising to the challenge: large scale sequence analysis

Coping with the analysis of rising amounts of sequence data has become a major scientific enterprise over the last few years, driven primarily by an abundance of information and its frequently incomplete digestion. Significant recent progress has been made in four main areas: first, development of databases; second, advances in computer networking; third, improvements in information access software; and fourth, improvements in algorithms and methods of sequence analysis by computer. We review each of these in turn.

Grow and link: development of databases

The usefulness of sequence databases, such as EMBL/GenBank (nucleic acids) and Swissprot/PIR (proteins), is currently limited by incomplete integration into a coherent whole, and by incomplete links to other biological databases, such as GDB (genome maps) and PDB (three-dimensional [3D] structures). A number of efforts are under way to provide more inter-database links (technical term: interoperability of databases), to add carefully annotated specialized databases, and to add value to existing databases by deriving additional information from them.

For example, numerous cross-pointers have been added to the Swissprot protein sequence database [18]. Given a protein sequence entry in this database,

Table 2. Principal sources of protein sequences.

Long genomic stretches	kb	ORFs[a]	References[b]
Nematode chromosome III	2000	n.d.	[3]
Yeast chromosome II	840	387	Feldmann *et al.*
Yeast chromosome XI	666	331	Dujon *et al.*
Escherichia coli[c]	500	n.d.	[8•]
Yeast chromosome III	320	170	[4,49•]
Cytomegalovirus	229	190	
Vaccinia virus	191	74	
Liverwort mitochondrion	187	74	
Tobacco chloroplast	156	109	
Bacillus subtilis	97	92	[54]
Fruit fly homeobox loci	80	n.d.	
Bacteriophage λ	49	63	
Mycobacterium leprae	37	12	[55]
Contig/EST collections	**kb**	**pieces**	
Mycoplasma contigs	350	650	[56]; Gillevet *et al.*
Caenorhabditis elegans (nematode) EST	n.d.	4699	[17,57,58]
Human brain EST	n.d.	6000	[16••]
Arabidopsis thaliana (cress) EST	n.d.	4512	[17]
Oryza sativa (rice) EST	n.d.	4231	[17]
Databases[d]	**kb or kaa[a]**	**seqs[a]**	
Swissprot[e](kaa)	11 500	33 300	
EMBL[e]without ESTs (kb)	155 000	155 000	
dbESTs [kb]	~10000	31 800	
PDB [kaa]	444	2200	

[a]Abbreviations: kaa, 1000 amino acids; kb, kilobases; n.d., not determined, or not available to the authors; ORFS, number of open reading frames (predicted proteins); seqs, number of sequence entries. [b]Selected references only are given; unnumbered references are personal communications. [c]There are four neighbouring, slightly overlapping *E. coli* segments in the EMBL nucleotide sequence database — the longest one covers about 176 kb. [d]Redundancy within and among sequence databases slightly reduces the numbers of bases/genes given; PDB = Protein Data Bank of three-dimensional structures. [e]Similar numbers for PIR (proteins) and GenBank (DNA).

cross-references give information, if available and applicable, on the nucleic acid sequence of its gene, the genomic map location of the gene, the enzyme commission functional classification, the three-dimensional structure, known homologs, known characteristic sequence patterns (PROSITE [19]), and so on. An example of added-value databases is the database of sequence–structure alignments (HSSP) that provides multiple sequence alignments, position-specific information on the degree of conservation, and sequence search profiles [20]. Examples of specialized databases, designed to address particular end-user requirements and carefully maintained by curators, are the databases of ESTs [17], *E. coli* databases [21,22], the worm (nematode *C. elegans*) database (see above), and FLYBASE (*Drosophila*) [23]. There are many more, too numerous to list. The urgent need for cross-references and interoperability is illustrated by the fact that a directory of molecular biology databases is a database in itself, e.g. LiMB [24].

Dial-up information: wide-area computer networks

Databases in themselves are of limited use unless they are easily accessible. Computer networks and information-retrieval software provide such access. Recently, both the physical and logical infrastructure of national and international computer networks and network software have improved considerably. As a result, use of wide-area information resources has become an important activity for sequence analysis experts. These resources allow the search, identification, and subsequent transfer of large amounts of publicly available material, including software and data in various forms. Information typically travels on Internet, the precursor of the planned 'information superhighways'. National and international laboratories provide various servers for automated sequence analysis [25]. Database updates can be performed using file transfer protocols (ftp). Resource browsers such as XMosaic™, developed at the US National Center for Supercomputer Applications (NCSA) in Illinois, are valuable software tools for navigating networks in search of data and information.

As more and more institutions offer data and services, the identification of relevant information resources becomes increasingly difficult. What is needed are client–server systems and information service brokers. In such systems, the user specifies the tasks using local (client) software and sends out the request to a remote system (server) which then distributes the required actions over the appropriate and available resources (for example, the 'Hassle' prototype [26] being tested on EMBnet, the European Molecular Biology Network). The key point is that copies of all data, software and information are not, and cannot, be kept locally without considerable effort, and therefore convenient access to network information services is an essential element for efficient sequence analysis work.

Access to integrated information: software tools

An interesting development in the direction of intelligent retrieval systems for genome data is AceDB (**A** ***C. e****legans* **d**ata**b**ase, by Richard Durbin and Jean Thierry-Mieg), an integrated system developed for the nematode sequencing project, now also in use by other genome projects. AceDB incorporates many programs that handle and analyze raw DNA and other sequence data, as well as map data, and contains a convenient graphical user interface with hyperlinks for data browsing. Other systems are under development, each with slightly different orientation, at the US National Center for Biotechnology Information (NCBI), the Genome Database Center at John Hopkins Medical School (GDB), EMBL, the German Cancer Research Center (DKFZ) and elsewhere. An excellent example of future client–server access to information across international boundaries is the network version of the Entrez software distributed by the NCBI that gives access to Medline literature citations relevant to sequence databases. If integrated access to diverse information becomes generally available, the efficiency of sequence analysis work will be much higher. To make effective use of all the information, genome projects will have to develop and apply techniques for information services in parallel with sequencing technology.

Progress report: methods of sequence analysis

Common sequence analysis practiced by the casual user, using standard programs, tends to miss a significant fraction of the functional information in protein sequences. It is therefore important to see what can be achieved with new and sophisticated methods. The list of basic procedures (selection in Table 3) is gradually increasing as new algorithms are invented and old ones improved (reviewed in [27]). We review here some of the most interesting programs that have had a practical impact on the field.

Identification of open reading frames

The accuracy of ORF prediction has been improved [28,29], relative to widely used programs such as Genmark [30], Genefinder [31], Grail [32] and the Staden package [33]. Frameshift detection and detection of other errors are an important technical issue, affecting the quality of derived amino acid sequences [34,35]. Routine use will indicate which of these new methods have noticeable practical advantages.

Analysis of amino acid composition bias

Once a putative protein sequence is available, a number of analysis methods can be applied. The best known, and most powerful, are sequence database alignment searches. However, an assessment of the significance of particular 'hits' (match of query with target sequence) depends strongly on the structural and functional class of the protein coded by that sequence (globular, filamentous, transmembrane, etc.). In practice, the problem is twofold. First, standard measures of sequence similarity require different cutoffs depending on the amino acid composition of the sequences being compared. Second, larger proteins have an inhomogeneous amino acid composition, i.e. a distinctly different composition in different regions (e.g. hydrophobic or charged stretches or 'domains'). Thus, before doing alignment database searches, it is useful to first determine the composition bias, i.e. deviation from the typical composition of globular proteins or deviation from the average composition in the protein sequence database. An approximate classification of different types of composition bias is as follows:

Coiled-coil arrangements. Whereas the detection of such regions using the program of Lupas *et al.* [36] appears to be fairly accurate, no reliable distinction between two-stranded and three-stranded, or between parallel and antiparallel, coiled-coil regions is possible yet.

Transmembrane regions. There are many programs for the prediction of transmembrane segments and signal peptides. Most are based on the simple notion of detecting runs of hydrophobic residues. We are not aware of any recent substantial improvement in accuracy.

Other low-entropy (or low-complexity) regions. A particular amino acid composition may be the result of functional selective pressure, e.g. a run of positively charged residues involved in non-specific protein–nucleic acid interaction. Recently, progress has been made in identifying heavily biased regions such as small repeats, long charged clusters, or regions rich in one, or a few, particular amino acids generally atypical of globular proteins (Wootton, this issue, pp 413–421). Complementing the work of Karlin and associates [37], the programs 'Seg' [38•] and 'Xnu' [39•] fill an urgent need in this area. These methods identify and mask composition-biased regions in the query sequence. The surrounding sequence and the biased regions can then be processed separately.

Table 3. From genome sequences to protein function and structure.

Steps	Problem	References[a]
Contig assembly	Detection of DNA overlaps	
Error correction	Identification of frameshifts, etc.	[34,35]
Open reading frame prediction	Identification of putative genes	[28,29]
Masking	Exclusion of regions with amino acid composition bias	[38•,39•]
Coiled-coil detection	Recognition of coiled-coil areas	[36]
Hydrophobicity analysis	Detection of transmembrane and signal segments	
Database homology search	Detection of sequence similarities	[41•,43••]
Multiple alignment and tree construction	Definition and analysis of protein families; definition of profiles	[41•,44,59]
Database profile or pattern search	Detection of distant relationships	
Self-alignment	Detection of internal repeats	
Secondary structure prediction	Prediction of helices, strands, loops, and surface/interior	[45]
Three-dimensional modelling	Construction of detailed atomic models based on homology	

[a]Only references to some recent developments are included.

Derivation of amino acid comparison matrices

All alignment search methods use scoring matrices that assign similarity values for any pair of the 20 amino acids. A new scoring matrix, 'Blosum', was derived from multiple-sequence alignments in a database of conserved regions ('Blocks' [40]). Tests indicate that use of the Blosum matrix leads to improved performance in homology detection [41•] and its use is therefore increasing.

Database searches

A search for sequence similarities in a protein sequence database is the most useful and most widely used method of functional prediction. The limitations arise from the fact that derived amino acid sequences enter the databases only after some delay (due to processing) or not at all (due to errors in nucleic acid sequences, or in their interpretation). It is therefore useful, albeit costly, to search six-frame translations of nucleic acid sequences. One tool in the 'Blast' series [42] of fast search programs, TBlastN, does this very effectively. On occasion, this mode of search reveals homologies with very recently sequenced genes or with adjacent sequences in two different reading frames, usually evidence of a frameshift, possibly as a result of a sequencing error. The reverse operation, searching the protein sequence database with the six-frame translations of a newly sequenced gene using BlastX [43••], is extremely useful for characterizing protein coding regions in raw nucleic acid sequence data, thus complementing ORF prediction programs.

If several putative homologs are detected in a database scan, multiple-sequence alignment can reveal common functional motifs. A new method for the automatic detection of motifs in a set of sequences [44] now offers an alternative to standard programs. The use of profiles derived from multiple-sequence alignments also lead to improved secondary structure prediction [45].

Despite the progress in basic analysis techniques, the interpretation of apparently significant sequence similarities in functional terms is still an underdeveloped area. More sophisticated methods for the analysis of a set of sequences in a family are needed for a more accurate homology-based prediction of protein function.

The key goal: prediction of protein function

When all the analytical procedures (Table 3) have been applied to a newly determined protein sequence of unknown function, one faces the difficult task of assigning a putative function based on evidence from sequence similarities, pattern detection, and so on. This part of the analysis process cannot easily be embodied in an algorithm and therefore is the least automated. Typically, an attempt is made to interpret sequence similarity by transferring the functional information about one protein to the homologous relative. The problem is that prediction of function by analogy can be very precise and complete, or it can be very fuzzy and incomplete, depending on the data at hand and on the precise nature of the similarity.

The simplest cases are those in which there is strong sequence similarity to, for example, an enzyme of known function. The new protein is then predicted to have the same function as its homolog. But even in simple cases caution is needed. Sequence variation can have diverse consequences. Even highly similar proteins might have completely different functions. Striking examples are the eye lens crystallins, structural proteins which apparently evolved recently from metabolic enzymes [46]. Furthermore, isoenzymes fine-tune the metabolism by varying only slightly in substrate affinity; the substrate specificity of clearly homologous proteins can be changed completely, as observed in the different families of sugar kinases [47]. An unambiguously homologous protein may be the equivalent gene in another species, indicating direct lineage (ortholog) or the homology may merely imply descent after gene duplication (paralog). In the latter case, one has to be particularly cautious with functional predictions.

In other cases, structural or functional similarity may occur for only a small part of a larger given protein (a domain), e.g. from the presence of a zinc finger motif. Although this may suggest that the protein binds DNA or RNA, little else can be concluded about the function of the protein of which this domain is a part. Finally, only a few proteins, mainly enzymes, are sufficiently well characterized so that molecular details of catalysis as well as higher levels of regulation and physiological roles can be described.

These problem cases illustrate the need for more precise knowledge about functional variation in evolution. Which functional changes occur as a result of certain types and certain amounts of sequence change? More biochemical and genetic characterizations of sets of homologous proteins (or of engineered mutants of natural proteins) are needed before more precise and quantitative rules for function prediction by homology can be formulated.

Yield: how much functional information from homology ?

Considering the limitations described above, homology- and analogy-based predictions have proved to be extremely powerful in exploring genomic information. When comparing the output of several large-scale sequencing projects (Table 2), an identification of partial function has been possible for 40–65% of all predicted proteins (Fig. 1), with this figure showing an increasing tendency [16••,48,49•]. Remarkably, the number of tentative functional identifications by homology exceeds by far the number of functional determinations by direct experiment (Table 4). The first complete chromosome sequences have led to the rather sudden realization that we already know a considerable fraction of all protein functions. In yeast chromosome III, the most carefully analyzed eukaryotic chromosome to date, the rate of tentative functional assignment by homology exceeds 50% (Table 4).

Extrapolating into the future, we can expect a rapid rise in the probability of homolog detection when comparing a new sequence with all known sequences (Green, this issue, pp 404–412). This is particularly true if the sequencing of human ESTs approaches saturation in 1995, as has been predicted (M Adams, personal communication). However, the rate of functional prediction by homology will not rise nearly as fast. This is simply because on the one hand new sequences enter the databases at a rapid rate, while on the other hand most of these new sequences come without primary, i.e. experimental, functional information. So the gap between the total number of known protein families and those with identified function (Table 4) will increase considerably in the near future, before it ultimately decreases near the saturation limit.

These observations have two implications for the allocation of resources in genome projects. To maximize functional information, it is probably advisable to complement sequencing effort with two key activities. One focus is a concerted program for the experimental determination of new protein functions, in order to increase the store of primary functional information on which all derived functional identifications depend. The other focus is a further improvement in the reliability of homology detection by sequence data analysis, in order to make maximum use of the experimental information. Both activities are essential and interdependent. On the one hand, even a single experimentally determined function can be immediately carried over, at least in part, to all of its sequence relatives and represents a sizeable net gain in information, if it is of a new type. On the other hand, even a single percentage

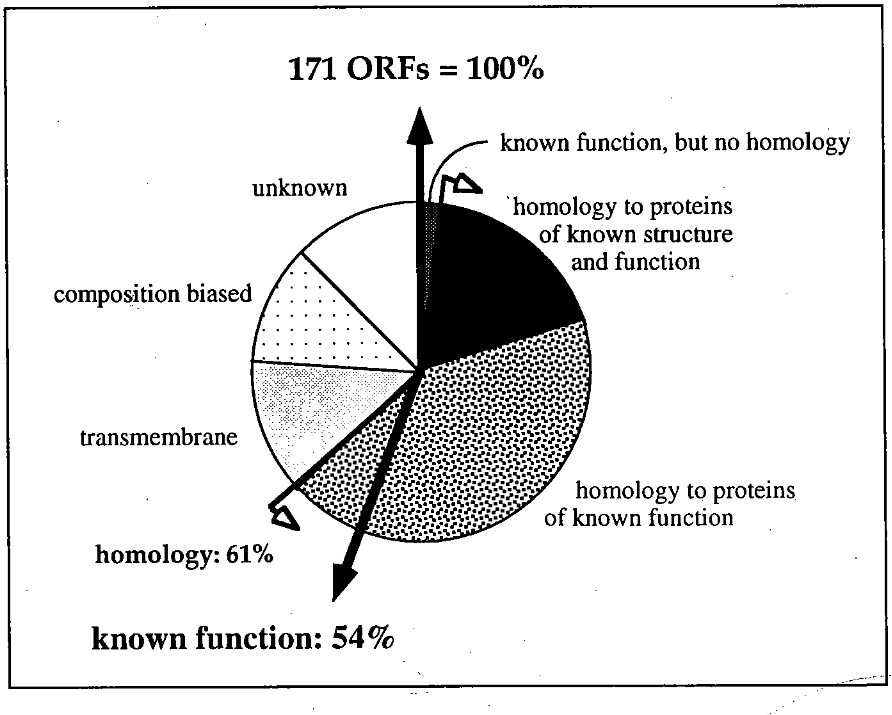

Fig. 1. Information content in the proteins of yeast chromosome III [49•]. There is a 10% gap between all protein families ('homology') and those of known function, termed here the function–homology gap (see text and Table 4). Numerically: 10% = 61% (homology)−54% (known function)+3% (known function, no homology).

Table 4. Identification of protein function by experiment and by homology, showing an increase in the function–homology gap using *Saccharomyces cerevisiae* chromosome III as an example.

	Yeast chromosome III (as of 1/1993)	Yeast chromosome III (as of 1/1994)	Anticipated increase or decrease[a]	Speculative estimate for yeast (1995/6)
Function[b]				
Function by experiment	11%	14%	+	17%
Function only by homology	31%	40%	++	50%
No function yet	58%	46%	––	33%
Homology[c]				
Member of sequence family	38%	61%	+++	80%
No family yet	62%	39%	–––	20%
Function–homology gap[d]	**1%**	**10%**		**15%**

[a] Increase: slow (+), intermediate (++) and rapid (+++); same for decrease (––), (–––). [b] Known full or partial function is given, identified by direct experiment or by detection of homology to a protein of known function (data from [4,49•,50]). [c] Experimental function represents an estimate, as there is some ambiguity in the definition of function; e.g. 'temperature sensitive lethal' was not counted as known function, while 'DNA repair protein' was counted. [c] Homology represents a significant sequence similarity to at least one other protein in the sequence databases, thus defining a protein family of two or more sequences, with or without known function.
[d] 'Gap' represents the fraction of all known protein sequences which have at least one homolog (belong to a family), but for which the function is not yet known, and was estimated using the numbers in the table and the number of proteins with known function that have no known homolog; the reason for the increasing gap is the determination of numerous new protein sequences without known function.

point improvement in functional prediction (currently at about 50%) translates to a large absolute number of identified protein functions, with immediate savings in experimental effort.

Molecular detail: implied three-dimensional structures

Detection of significant sequence similarity to a protein of known 3D structure immediately implies prediction of the 3D structure of the new protein by homology. The prediction leads to detailed arguments about the mechanism of protein action and about the role of particular residues. Often, very conserved residues distributed along the protein chain are in spatial proximity in 3D, explaining previously puzzling conservation patterns and suggesting detailed experiments. It is therefore interesting that as many as 19% of all yeast chromosome III sequences are significantly homologous to a known 3D structure (Fig. 1). For these proteins, approximate atomic models can be built, inspected using 3D graphics, and used as a basis for planning experiments. However, coverage is incomplete for certain types of protein structures, e.g. membrane proteins of which only a handful of 3D structures are known.

As the rate of 3D structure determination of proteins now exceeds one structure a day, the rate of 3D prediction by homology will probably reach the 20–25% level in the not so distant future. It already exceeds 24% for the Swissprot database (R Schneider, personal communication). This number comes from a systematic all-against-all alignment of the sequences in the Protein Data Bank (known structures) with the 32 000 sequences in Swissprot [20], assessing homology by the application of a threshold for structural similarity. It is remarkable that we already have the opportunity to identify the approximate tertiary structure of one in four newly sequenced proteins.

Feedback: checking predictions by experiment

The experimental verification of predictions by homology is very slow. The higher the belief in the validity of computer-based sequence analysis, the lower the incentive to perform control experiments. However, there is a small but non-negligible number of cases where prediction entices experiment, for a variety of reasons. One case is the surprising occurrence of a certain functional type in another organism.

An example of this is the discovery of a certain type of polymerase in yeast. The homology and predicted functional analogy between yeast ORF YCR14C and mammalian DNA polymerase-β [48,50] was experimentally verified by *in vitro* tests on the expressed protein [51], now named yeast DNA polymerase IV (*pol IV* gene) — an example of a successful prediction based on multiple alignment and sequence pattern analysis [50,52]. This now opens the way to biochemical and genetic studies in yeast, not just mammals, to understand the presumed role of the short gap-filling activity of β-polymerases in DNA repair.

Improvements: what to expect

The main gain in the level of function prediction over the last 2 years came primarily from three sources. First, protein-sequence databases have improved in quantity (more sequences, more frequent updates) as well as in quality (better annotation and more cross-pointers). Second, multiple-sequence alignment methods have improved, especially profile and pattern definition for protein families and their detection in new sequences. Third, diverse information resources are more conveniently accessible over the Internet, lowering the threshold for careful analysis, and controls and cross-checks for borderline cases. Full use of these improvements typically requires expertise.

```
MBHA_MYXXA    34 VALDIKSDDGGKTLKGTMT  myxobacterial hemagglutinin
MBHA_MYXXA   101 VAVSIKSNDGGKTLTGTTT  (Myxococcus xanthus)
MBHA_MYXXA   168 INVDAKSNDGGKTLSGTMT
MBHA_MYXXA   235 VALNVASSDGGKTLAGTMI
NANH_BACFR     1 DVGLSRSTDGGKTWEKMRL  sialidase (EC 3.2.1.18)
NANH_BACFR    80 QLVLAKSTDDGKTWSAPIN  (Bacteroides fragilis)
NANH_BACFR   140 NAGIMYSKDGGKNWKMHNY
NANH_BACFR   247 NTTIKISLDGGVTWSPEHQ
NANH_CLOSE   426 DTGIKRSTDGGVTWDEGKI  sialidase precursor (EC 3.2.1.18)
NANH_CLOSE   559 FLSLIYSDDDGQTWSDPID  (Clostridium septicum)
NANH_CLOSE   623 SSAVIYSDDNGATWNIGET
NANH_CLOSE   696 RVRIATSFDGGATWEDDVV
NANH_SALTY    66 DTAAARSTDGGKTWNKKIA  sialidase (EC 3.2.1.18)
NANH_SALTY   140 DLVLYKSTDDGVTFSKVET  (Salmonella typhimurium)
NANH_SALTY   205 NTSFIYSTD-GITWSLPSG
NANH_SALTY   251 LRRSFETKDFGKTWTEFPP
TCNA_TRYCR    19 DTVAKYSVDDGETWETQIA  sialidase (EC 3.2.1.18) SA85-1.2 -
TCNA_TRYCR   130 TPEVTKSTAGGKITASIKW  major surface antigen
TCNA_TRYCR   159 FSKIFYSEDDGKTWKFGKG  (Trypanosoma cruzi)
TCNA_TRYCR   205 RRLVYESSDMEKPWVEAVG
PEP1_YEAST   411 KGVTKISVDNGLTWTMLKV  PEP1 on yeast chromosome II
PEP1_YEAST   483 DQRTFISRDGGLTWKLAFD  (Saccharomyces cerevisiae)
PEP1_YEAST   529 QSEFYYSLDQGKTWTEYQL
PEP1_YEAST   581 TTNFIYAIDFSTAFNDKTC
YCZ0_YEAST    63 LSEIFISDSQGLKFSPIPF  YCR100C on yeast chromosome III
YCZ0_YEAST   120 GGETKISVDNGLTWSNLKV  (Saccharomyces cerevisiae)
YCZ0_YEAST   192 DRKTFISRDGGLTWRVAHN
YCZ0_YEAST   238 QSKLYFSLDQGRTWNQYEL
motif:               hhhS D G TW          (h - hydrophobic)
```

Fig. 2. Sialidase sequence motif in yeast proteins [49•]. A typical example of short and weak motifs that are not detectable by standard homology searches. The motifs are, however, detectable by profile and pattern searches. The significance of the sialidase motif is supported by its multiple occurrence with an average spacing of about 50 amino acids. As a result, a sialidase activity is predicted for the yeast proteins. As a 3D structure of one bacterial sialidase is now known (Fig. 3), structural, as well as functional, information can be inferred for the yeast proteins.

An example of a result obtained with more sophisticated methods in the context of large-scale genome analysis is the identification of sialidases in yeast chromosomes III and II. During the analysis of chromosome III [49•,50] the putative ORF YCR100 initially did not match any database protein apart from PEP1, a functionally uncharacterized protein from yeast chromosome II. However, the conservation profile between both proteins revealed four short, but conserved, internal repeats. Using this information in pattern search methods [52], the investigators detected subtle similarities to conserved repeats in bacterial and protozoan sialidases (Fig. 2). Very recently, the three-dimensional structure of one of these sialidases was determined [53] (Fig. 3), and indeed, it contains internally repeated β-sheets forming a superbarrel or propeller fold, fully consistent with the sequence repeat; the most conserved residues (Fig. 2) are located in equivalent positions in the respective sheets (Fig. 3). Thus, based on a short signature motif (too short and weak to be detected by conventional homology search programs) a rather precise functional and structural prediction was made. This example emphasizes the need to incorporate such methods into standard analysis procedures and illustrates the potential gain.

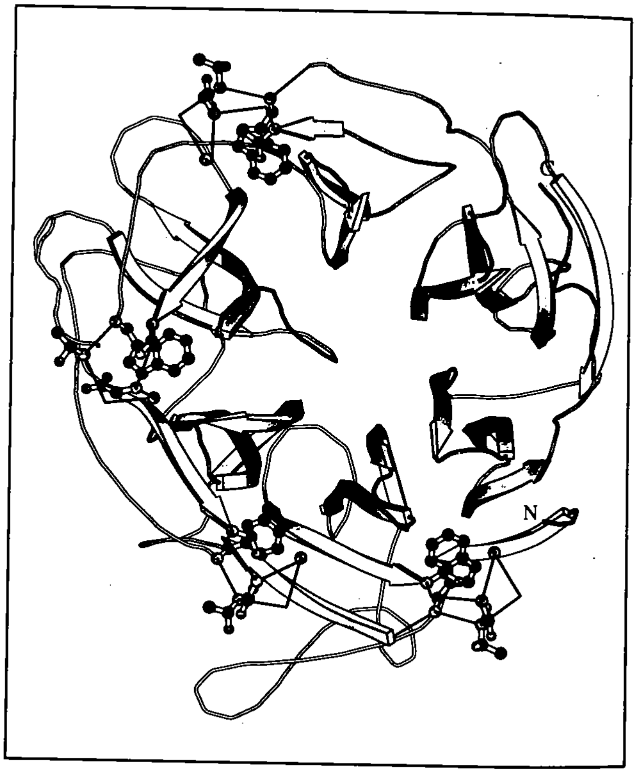

Fig. 3. Ribbon plot of the sialidase from *Salmonella typhimurium* [53]. It is an example of the so-called β-propeller fold in which six four-stranded β-sheets form a superbarrel. The four sequence motifs are located in equivalent positions in four of the six β-sheets. In the remaining two sheets, the corresponding motifs are not detectable (Fig. 2).

In summary, for the immediate future the most pressing and promising needs of genome sequence analysis are manifold. First, further refinement of pattern and profile searches. Second, automation of the analysis process, especially for sequence families. Third, improved data support by direct access to specialized sequence and bibliographic databases. Fourth, earlier public accessibility of data from major sequencing projects. Fifth, training of analyzers in advanced methods. Sixth, the development of a more refined classification of protein function, and finally, a better understanding of the effect of sequence changes on protein function.

Conclusion

In a few years' time, the complete sequences of several entire genomes will become known, resulting in a series of historical achievements: *E. coli*, mycoplasma(s), yeast, nematode, are proceeding apace (Table 1). Genomes from organisms of all taxonomic ranks will follow, including *'Homo ignorans'*. The analysis of these data will require a high degree of automation in sequence analysis without sacrificing the sensitivity of present methods for the detection of distant sequence similarities. Where sequence analysis has no answers, experimental technologies will be essential for the genetic and biochemical characterization and the physical identification of completely new types of proteins. In addition, many experiments will be stimulated by the detailed predictions of sequence analysis. The database — or is it knowledge base? — of all proteins will gradually be completed, including 3D structures and diverse functional knowledge.

Imagine that at some time almost all proteins of a particular organism will be known — sequence, 3D structure and function — and stored in the World-Prot database. The obvious question that then arises is whether the molecular repertoire of an organism is sufficient to characterize its physiological and evolutionary behavior. The obvious answer is that it is not, and that biological experiments and theories at higher, less microscopic, levels are needed to complement the atomic information available from genome-sequencing projects.

Will a graduate student today, trained in sequencing and in sequence analysis, end his career like a zoologist in the beginning of this century, hunting the last unclassified butterflies in Madagascar? In other words, will experimental and computational sequence analysis be transformed from a skilful scientific endeavor to an activity of lesser scientific interest? Or, will it provide the ultimate answers to the grand questions of biological science, about structure and function, development and evolution? The truth lies somewhere in between. The functional classification of all proteins will be an excellent intermediate goal for that graduate student, but also an excellent point of departure for addressing the real questions of human health, the environment, and the future evolution of life.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:
- • of special interest
- •• of outstanding interest.

1. Collins F, Galas D: **A new Five-Year Plan for the US Human Genome Project.** *Science* 1993, 262:43–46.

2. Copeland NG, Jenkins NA, Gilbert DJ, Eppig JT, Maltais LJ, Miller JC, Dietrich WF, Weaver A, Lincoln SE, Steen RG, *et al.*: **A Genetic Linkage Map of the Mouse: Current Applications and Future Prospects.** *Science* 1993, 262:57–66.

3. Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, Burton J, Connell M, Copsey T, Cooper J, *et al.*: **The *C. elegans* Genome Project: Contiguous Nucleotide Sequence of Over Two Megabases from Chromosome III.** *Nature* 1994, 368:32–38

4. Oliver SG, van der Aart QJM, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, Antoine G, Anwar R, Ballesta JPG, Benit P, *et al.*: **The Complete DNA Sequence of Yeast Chromosome III.** *Nature* 1992, 356:38–46.

5. Wang MX, Church GM: **A Whole Genome Approach to *in vivo* DNA-Protein Interactions in *E. coli*.** *Nature* 1992, 360:606–610.

6. Plunkett G III, Burland V, Daniels DL, Blattner FR: **Analysis of the *Escherichia coli* Genome. III. DNA Sequence of the Region from 87.2 to 89.2 Minutes.** *Nucleic Acids Res* 1993, 21:3391–3398.

7. Burland V, Plunkett G III, Daniels DL, Blattner FR: **DNA Sequence and Analysis of 136 kilobases of the *Escherichia coli* Genome: Organizational Symmetry Around the Origin of Replication.** *Genomics* 1993, 16:551–561.

8. • Blattner FR, Burland V, Plunkett G III, Sofia HJ, Daniels DL: **Analysis of the *Escherichia coli* Genome. IV. DNA Sequence of the Region from 89.2 to 92.9 Minutes.** *Nucleic Acids Res* 1993, 21:5408–5417.

This is the last report in a series of papers that describe long segments of the *E. coli* genome. With this report, the data from this model organism can be assembled to a continuous segment of about 500 kb.

9. Yura T, Mori H, Nagai H, Nagata T, Ishihama A, Fujita N, Isono K, Mizobuchi K, Nakata A: **Systematic Sequencing of the *Escherichia coli* Genome: Analysis of the 0–2.4 Min Region.** *Nucleic Acids Res* 1992, 20:3305–3308.

10. Daniels DL, Plunkett III G, Burland V, Blattner FR: **Analysis of the *Escherichia coli* Genome: DNA Sequence of the Region from 84.5 to 86.5 Minutes.** *Science* 1992, 257:771–778.

11. Sulston J, Du Z, Thomas K, Wilson R, Hillier L, Staden R, Halloran N, Green P, Thierry-Mieg J, Qiu L, *et al.*: **The *C. elegans* Genome Sequence Project: a Beginning.** *Nature* 1992, 356:37–41.

12. Reiter RS, Williams JG, Feldmann KA, Rafalski JA, Tingey SV, Scolnik PA: **Global and Local Genome Mapping in *Arabidopsis thaliana* by Using Recombinant Inbred Lines and Random Amplified Polymorphic DNAs.** *Proc Natl Acad Sci USA* 1992, 89:1477–1481.

13. • Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S: Characterization of the Pufferfish (*Fugu*) Genome as a Compact Model Vertebrate Genome. *Nature* 1993, 366:265–268.

The authors identify the *Fugu* genome as a model vertebrate genome for sequencing projects, due to its small size (only four times larger than *C. elegans*). Evidence is presented that this particular genome is 90% unique, and devoid of repetitive sequences.

14. Chapman VM, Copeland NG, Costantini FD, Dove WF, Nadeau JH, Reeves RH, Rossant J, Smithies O, Woychik RP: A Plan for the Mouse Genome Project. *Mammalian Genome* 1993, 4:293–300.

15. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al.*: Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science* 1991, 25:1651–1656.

16. •• Adams MD, Kerlavage AR, Fields C, Venter JC: 3,400 New Expressed Sequence Tags Identify Diversity of Transcripts in Human Brain. *Nature Genet* 1993, 4:256–267.

The authors provide a very interesting analysis of the ever-increasing EST data. A profile of the transcriptional activity of human brain is presented, based on sequence similarity searches. Cytoskeletal and other structural proteins seem to be the most abundant.

17. Boguski MS, Lowe TMJ, Tolstoshev CM: dbEST Database for Expressed Sequence Tags. *Nature Genet* 1993, 4:332–333.

18. Bairoch A, Boeckmann B: The SWISS-PROT Protein Sequence Data Bank, Recent Developments. *Nucleic Acids Res* 1993, 21:3093–3096.

19. Bairoch A: The PROSITE Dictionary of Sites and Patterns in Proteins, its Current Status. *Nucleic Acids Res* 1993, 21:3097–3103.

20. Sander C, Schneider R: The HSSP Data Base of Protein Sequence-Structure Alignments. *Nucleic Acids Res* 1993, 21:3105–3109.

21. Kröger M, Wahl R, Rice P: Compilation of DNA Sequences of *Escherichia coli* (Update 1993). *Nucleic Acids Res* 1993, 21:2973–3000.

22. Rudd KE: Maps, Genes, Sequences, and Computers: An *Escherichia coli* Case Study. *Am Soc Med News* 1993, 59:335–441.

23. FLYBASE: The Drosophila Genetic Database, 1993. Available from ftp.bio.indiana.edu network server.

24. Lawton JR, Martinez FA, Burks C: Overview of the LiMB Database. *Nucleic Acids Res* 1989, 17:5885–5899.

25. Henikoff S: Sequence Analysis by Electronic Mail Server. *Trends Biochem Sci* 1993, 18:267–268.

26. Doelz R: Hassle — A Tool to Access Sequence Databases Remotely. *Comput Appl Biosci* 1994, 10:31–34

27. Doolittle RF: Protein Sequence Comparisons: Searching Databases and Aligning Sequences. *Curr Opin Biotechnol* 1994, 5:24–28.

28. Guigo R, Knudson S, Drake N, Smith T: Prediction of Gene Structure. *J Mol Biol* 1992, 226:141–157.

29. Farber R, Lapedes A, Sirotkin K: Determination of Eukaryotic Protein Coding Regions Using Neural Networks and Information Theory. *J Mol Biol* 1992, 226:471–479.

30. Borodovsky M, McIninch J: Recognition of Genes in DNA Sequences with Ambiguities. *Biosystems* 1993, 30:161–171.

31. Green P, Hillier L: Genefinder Software, Unpublished. 1993, Department of Genetics, University of Washington, Missouri.

32. Uberbacher E, Mural R: Locating Protein-Coding Regions in Human DNA Sequences by a Multiple Sensor-Neural Network Approach. *Proc Natl Acad Sci USA* 1991, 88:11261–11265.

33. Staden R: Finding Protein Coding Regions in Genomic Sequences. *Methods Enzymol* 1990, 183:163–180.

34. Posfai J, Roberts RJ: Finding Errors in DNA Sequences. *Proc Natl Acad Sci USA* 1992, 89:4698–4702.

35. Claverie J-M: Detecting Frameshifts by Amino Acid Comparison. *J Mol Biol* 1993, 234:1040–1057.

36. Lupas A, Dyke Mv, Stock J: Predicting Coiled Coils from Protein Sequences. *Science* 1991, 252:1162–1164.

37. Karlin S, Brendel V: Chance and Statistical Significance in Protein and DNA Sequence Analysis. *Science* 1992, 257:39–49.

38. • Wootton JC, Federhen S: Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. *Comput Chem* 1993, 17:149–163.

An elaborate analysis of low-complexity regions in protein sequences and their detection is described. The approach extends beyond single sequences, and provides means for the generation of small databases.

39. • Claverie J-M, States D: Information Enhancement Methods for Large Scale Sequence Analysis. *Comput Chem* 1993, 17:191–201.

The authors describe an algorithm and its implementation for detection and masking of compositionally biased regions in protein sequences.

40. Henikoff S, Henikoff JG: Automated Assembly of Protein Blocks for Database Searching. *Nucleic Acids Res* 1991, 19:6565–6572.

41. • Henikoff S, Henikoff JG: Performance Evaluation of Amino Acid Substitution Matrices. *Proteins* 1993, 17:49–61.

Evidence is provided for the superior performance of the BLOSUM family of substitution matrices for database searching and sequence comparison.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic Local Alignment Search Tool. *J Mol Biol* 1990, 215:403–410.

43. •• Gish W, States DJ: Identification of Protein Coding Regions by Database Similarity Search. *Nature Genet* 1993, 3:266–272.

An extension of the BLAST suite of programs that allows the search of protein databases using a nucleotide as a query. Apart from its use for the identification of protein coding regions, this approach has other applications, such as the detection of frameshifts and the quality control of DNA sequencing.

44. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science* 1993, 262:208–214.

45. Rost B, Sander C: Prediction of Protein Secondary Structure at Better Than 70% Accuracy. *J Mol Biol* 1993, 232:584–599.

46. Piatigorsky J, Wistow GJ: Enzyme/Crystallins: Gene Sharing as an Evolutionary Strategy. *Cell* 1989, 57:197–199.

47. Bork P, Sander C, Valencia A: Convergent Evolution of Similar Enzymatic Function on Different Protein Folds: The Hexokinase, Ribokinase, and Galactokinase Families of Sugar Kinases. *Protein Sci* 1993, 2:31–40.

48. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E: What's in a Genome? *Nature* 1992, 358:287–287.

49. • Koonin EV, Bork P, Sander C: Yeast Chromosome III: New Gene Functions. *EMBO J* 1994, 13:493–503.

The power of functional prediction by homology was demonstrated for yeast chromosome III. The various methods used set the percentage of functional assignments for yeast chromosome III proteins at 60%.

50. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E: **Comprehensive Sequence Analysis of the 182 Predicted Open Reading Frames of Yeast Chromosome III.** *Protein Sci* 1992, 1:1677–1690.

51. Prasad R, Widen SG, Singhal RK, Watkins J, Prakash L, Wilson SH: **Yeast Open Reading Frame YCR14C Encodes a DNA β-Polymerase-Like Enzyme.** *Nucleic Acids Res* 1993, 21:5301–5307.

52. Rohde K, Bork P: **A Fast, Sensitive Pattern-Matching Approach for Protein Sequences.** *Comput Appl Biosci* 1993, 9:183–189.

53. Crennel SJ, Garman EF, Laver WG, Vimr ER, Taylor GL: **Crystal Structure of a Bacterial Sialidase (from *S. typhimurium* LT2) Shows the Same Fold as an Influenza Virus Neuraminidase.** *Proc Natl Acad Sci USA* 1993, 90:9852–9856.

54. Glaser P, Kunst F, Arnaud M, Coudart M-P, Gonzales W, Hullo M-F, Ionescu M, Lubochinsky B, Marcelino L, Moszer I, *et al.*: ***Bacillus subtilis* Genome Project: Cloning and Sequencing of the 97 Region from 325 to 333.** *Mol Microbiol* 1993, 10:371–384.

55. Honoré N, Bergh S, Chanteau S, Doucet-Populaire F, Eiglmeier K, Garnier T, Georges C, Launois P, Limpaiboon T, Newton S *et al.*: **Nucleotide Sequence of the First Cosmid from the *Mycobacterium leprae* Genome Project: Structure and Function of the Rif-Str Regions.** *Mol Microbiol* 1993, 7:207–214.

56. Peterson SN, Hu P-C, Bott KF, Hutchison CAI: **A Survey of the *Mycoplasma genitalium* Genome by Using Random Sequencing.** *J Bacteriol* 1993, 175:7918–7930.

57. Waterston R, Martin C, Craxton M, Huynh C, Coulson A, Hillier L, Durbin R, Green P, Shownkeen R, Halloran N, *et al*: **A Survey of Expressed Genes in *Caenorhabditis elegans*.** *Nature Genet* 1992, 1:114–123.

58. McCombie WR, Adams MD, Kelley JM, FitzGerald MG, Utterbeck TR, Khan M, Dubnick M, Kerlavage AR, Venter JC, Fields C: ***Caenorhabditis elegans* Expressed Sequence Tags Identify Gene Families and Potential Disease Gene Homologues.** *Nature Genet* 1992, 1:124–131.

59. Higgins DG, Bleasby AJ, Fuchs R: **CLUSTAL V: Improved Software for Multiple Sequence Alignment.** *Comput Appl Biosci* 1992, 8:189–191.

P Bork, C Ouzounis and C Sander, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany.