Bioinformatics: from genome data to biological knowledge

Miguel A Andrade* and Chris Sander†

Recently, molecular biologists have sequenced about a dozen bacterial genomes and the first eukaryotic genome. We can now obtain answers to detailed questions about the complete set of genes of an organism. Bioinformatics methods are increasingly used for attaching biological knowledge to long lists of genes, assigning genes to biological pathways, comparing the gene sets of different species, identifying specificity factors, and describing sets of highly conserved proteins common to all domains of life. Substantial progress has recently been made in the availability of primary and added-value databases, in the development of algorithms and of network information services for genome analysis. The pharmaceutical industry has greatly benefited from the accumulation of sequence data through the identification of targets and candidates for the development of drugs, vaccines, diagnostic markers and therapeutic proteins.

Addresses
EMBL-EBI, Genome Campus, Cambridge, CB10 1SD, UK
*e-mail: andrade@embl-ebi.ac.uk
†e-mail: sander@embl-ebi.ac.uk

Current Opinion in Biotechnology 1997, 8:675–683

http://biomednet.com/elecref/0958166900800675

Abbreviations

BLAST	basic local alignment search tool
EST	expressed sequence tag
ORF	open reading frame

Introduction: information transfer and evolution

Life involves the storage, handling and transformation of information. The basic information necessary to construct and manage a living organism is contained in its genome. This information is transformed into physical reality by molecular processes in living cells. Parts of the genome are translated into proteins. Other parts of the genome regulate the expression of these proteins. The translated proteins fold into highly specific three-dimensional structures. The proteins are targeted to their precise cellular location where they perform their function. And so on.

Many details of these processes can be unraveled in the molecular biologist's laboratory. Biological experimentation, however, is generally laborious, expensive, slow and rarely comprehensive. Fortunately, as a result of evolution and of physicochemical constraints, biological organization is redundant: elementary molecular processes are multiply adapted and reused in different contexts and in different species. This redundancy can be wonderfully exploited by using evolutionary comparison and system analysis to transfer biochemical, genetic and cell-biological knowledge from one set of biological molecules to another, both within one species and between species. The art of evolutionary information transfer is in the skill set of bioinformatics.

Bioinformatics is a science of recent creation that uses biological data and knowledge stored in computer databases, complemented by computational methods, to derive new biological knowledge. It is a theoretical biology firmly grounded in comprehensive and detailed experimental facts. Currently, bioinformatics is making a key contribution to the organization and analysis of the massive amount of biological data from genome sequencing projects and, increasingly, from other areas of 'high-throughput', 'massively parallel', robotized and miniaturized methods of biological experimentation.

The classical progression of the pharmaceutical discovery process goes from drug target (typically a receptor protein or enzyme) to lead compound to drug. With the arrival of multiple genome sequences, bioinformatics is already making practical contributions in target identification and in the design of combinatorial libraries based on detailed knowledge of one or more protein structures.

This review focuses on recent bioinformatics methods applied to protein sequences, functions, structures and pathways. Genome analysis can be performed at different levels of complexity: individual sequences, complete genomes, and sets of genomes. We will review applications at each level, from protein sequence analysis to comparative genomics.

Prediction of protein function

The richest information in genome sequences is in the set of genes coding for functional proteins. As only a minority of proteins have their function known by direct experiment, information transfer from one protein to another, by evolutionary analogy, often yields the first indications about the function of a newly sequenced gene. The information transfer relies on the notion (shown experimentally to be true in numerous cases) that proteins similar in sequence are also similar in function. The transfer of functional information from one protein to another can, however, lead to substantial errors unless a number of technical difficulties are properly addressed (e.g., protein domain organization, hierarchial definition of protein functionality).

In practice, given a new protein sequence, the following questions arise. Does the protein belong to a known protein family about which functional information is available? If so, what is the correct alignment and is a 3D structure available to build a model by homology?

How closely is the protein related to the members of the family and what does this imply about the similarity in function? Which region of the protein matches the family, and does the match occur in a region known to relate to the functional properties of the protein? Within that region, are the known or apparent functional residues conserved? While well-trained experts in sequence analysis are often able to address these questions using standard database search software, considerable effort is going into algorithm and software development in these areas. Here, we can only cover selected and recent achievements.

The ability to assess membership in a particular protein family depends critically on the alignment search used. Most searches are based on pairwise sequence-sequence comparison. The most widely used method, the basic local alignment search tool (BLAST) algorithm [1], relates similar sequence fragments. This algorithm has been improved by two recent adaptations (WU-BLAST2 [2]; PSI-BLAST [3•]). PSI-BLAST extends the search capability by comparing the query protein's protein family, as found by the sequence-sequence search, with the sequence database in a second and further passes (this appears to be similar to MOST [4], a sequence-sequence comparison method also based on protein family generation from BLAST results). In general, a set of aligned sequences can be organized (i.e., grouped and aligned) into an emerging family in a variety of ways to define a 'profile' or family 'model'. Such profiles aim at capturing the key functionally constrained features of the family. As a result, profile-sequence comparisons are a more powerful search tool than mere sequence-sequence comparisons [5,6].

The next level of database searches involves profile-profile comparison, with increased power for the detection of remotely related family members; however, the availability of well-curated and comprehensive datasets of protein family profiles is still limited (see Table 1). In rare cases, discovery of similarities in 3D structure, without apparent sequence similarity, can lead to the unification

Table 1

Selected databases and information services.

Database	Content	World Wide Web URL	Reference or site author
DNA and protein sequences			
EMBL	Nucleotide sequences	www.embl-ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html	[41]
GenBank	Nucleotide and protein sequences	www.ncbi.nlm.nih.gov/Web/Genbank/index.html	[42]
SwissProt	Protein sequences	expasy.hcuge.ch/sprot/sprot-top.html	[43]
PIR	Protein sequences	www-nbrf.georgetown.edu/pir/	[44]
Genome sequencing projects	Genome completion list	www.mcs.anl.gov/home/gaasterl/genomes.html	T Gaasterland
Gene identification			
GeneMark	Gene identification	amber.biology.gatech.edu/~william/genemark.html	[45]
Grail	Gene identification	compbio.ornl.gov/Grail-1.3/	[46]
Gene finder	Gene identification	dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html	V Solovyez
Frame	Frame-shift detection	www.sander.embl-ebi.ac.uk/frame/	[47]
Functional genome analysis			
GeneQuiz	Automatic protein function annotation	www.sander.embl-ebi.ac.uk/genequiz	[11]
Magpie	Genome analysis	www.mcs.anl.gov/home/gaasterl/magpie.html	[48]
Pedant	Automatic analysis of proteins	pedant.mips.biochem.mpg.de/frishman/pedant.html	[49]
Complete Genomes at NCBI	Genome analysis and comparison	www.ncbi.nlm.nih.gov/Complete_Genomes	[32•]
Protein families and sequence patterns			
Prosite	Protein sites and patterns	expasy.hcuge.ch/sprot/prosite.html	[50]
Blocks	Protein patterns	www.blocks.fhcrc.org	[51]
Pfam	Protein families and profiles	www.sanger.ac.uk/Software/Pfam/	[52]
Database links			
SRS	Biological database browser	srs.embl-ebi.ac.uk:5000/srs5/	[12]

There are many more World Wide Web sites relevant to bioinformatics. The selection here relates to the discussion in the text and, in some cases, new developments.

of functional families into functional superfamilies, with an attendant increase in database search power [7•,8•]. Finally, given a reliable multiple-sequence alignment from any method, the analysis of specificity determinants (residues characteristically conserved within a subfamily but varying between subfamilies) provides a detailed handle on the verification of functional hypotheses, regarding, for example, enzyme mechanisms and substrate specificity [9,10•].

Apart from the technical problems in correctly assigning a new protein to a protein family, sequence analysis is increasingly faced with the problems of scale. On a given Monday, a new set of several thousand protein genes may become available as the sequence of a new genome is released to the public. Processing these one by one, applying the expert's intuition to interpretation of program output, is too slow. The flood of new sequence data can only be handled by automation. Following the appearance of the GeneQuiz software system [11] for large scale sequence analysis in 1994, several groups have developed systems with varying degrees of automation (see Table 1).

GeneQuiz [11] aims at the derivation of functional information for sets of protein sequences. The software system performs a number of database searches and analyses using databases updated daily, makes deductions from the output of these searches using a simple expert system, and provides the end user with a summary report about the query protein's function (and structure) as well as hyperlinks to underlying information, such as sequence alignments. As a complement, both the SRS [12] (EMBL European Bioinformatics Institute) and Entrez (US National Center for Biotechnology Information) systems are very useful for browsing directly linked or neighboring information items. For completely sequenced genomes (Table 2), access to protein functional information derived using GeneQuiz is available

Table 1 (contd.)

Selected databases and information services.

Database	Content	World Wide Web URL	Reference or site author
3D Structure			
PDB	Protein structures	www.pdb.bnl.gov	[53]
FSSP	Protein folds	croma.embl-ebi.ac.uk/dali/fssp/	[54]
SCOP	Protein folds	scop.mrc-lmb.cam.ac.uk/scop/	[55]
CATH	Protein folds	www.biochem.ecl.ac.uk/bsm/cath/	CA Orengo *et al.*
DSSP	Protein secondary structure	www.sander.embl-ebi.ac.uk/dssp/	[56]
PDBselect	Representative protein structures	www.sander.embl-heidelberg.de/pdbsel	[57]
HSSP	Protein alignments	www.sander.embl-ebi.ac.uk/hssp/	[58]
PDBfinder	Links PDB, DSSP and HSSP	www.sander.embl-heidelberg.de/pdbfinder	[59]
Swiss-Model	Models by homology	expasy.hcuge.ch/cgi-bin/swmodel-search-de	[60]
Whatif	3D structure models	swift.embl-heidelberg.de/servers	[61]
DALI	Comparison of protein structures	croma.embl-ebi.ac.uk/dali/	[55]
Journal abstracts			
MEDLINE	Biochemical literature	www.nlm.nih.gov/PubMed	US National Library of Medicine
Evaluated MEDLINE	MEDLINE browser	biomednet.com/gateways/db/medline	BioMedNet
Protein domains and modules			
ProDom	Protein domains	protein.toulouse.infra.fr/	[62]
Modules	Extracellular protein modules	www.bork.embl-heidelberg.de/Modules/	[63]
Metabolic pathways			
EcoCyc HinCyc	Bacterial genes and metabolism	www.ai.sri.com/ecocyc/	[20]
WIT	Metabolic reconstructions	www.cme.msu.edu/WIT/	R Overbeek *et al.*
KEGG	Metabolic pathways	www.genome.ad.jp/kegg/kegg2.html	[19•]
GenProtEC	*Escherichia coli*	www.mbl.edu/html/ecoli.html	[64]
Human Genetics			
OMIM	Human genes and genetic disorders	ww3.ncbi.nlm.nih.gov/Omim/	VA McCusick
GBD	Human genome database	gdbwww.gdb.org/	[65]

There are many more World Wide Web sites relevant to bioinformatics. The selection here relates to the discussion in the text and, in some cases, new developments.

at www.sander.embl-ebi.ac.uk/genequiz/ . Generally accepted protein functional annotation is best accessed using Swissprot or Expasy (see Table 1).

Analysis of complete genomes

After the public release of the first bacterial genome sequence by Fleischmann *et al.* in 1995 [13], other bacteria and archaebacteria and the *Saccharomyces cerevisiae* yeast have been completely sequenced (see Table 2). Many more genomes than those publicly available today will either be released soon, are almost finished, or have been sequenced by companies with an uncertain or unknown release policy. Analysis of the complete set of genes and regulatory signals using tools of bioinformatics (including 3D modeling of protein structures and computation of the parameters of their interaction with small molecules) is currently proceeding in a number of research groups and answers to many interesting questions are expected in the near future. Here, we highlight only a few aspects of this rapidly expanding field.

The characterization and extraction of the translated parts of a genome sequence, the open reading frames (ORFs), is one of the first steps in genome sequence analysis. Existing methods [14] use either transcriptional and other regulatory signals to suggest boundaries of the ORFs [15•,16•] or exploit sequence similarity to sequences in databases of nucleotide fragments known to be expressed as proteins, called expressed sequence tags (ESTs) [17•]. The error rate of ORF assignment (also called gene identification) is non-negligible, both in terms of missing proteins as well as ORFs not corresponding to expressed proteins entering the databases [18•]. For eukaryotic genomes, gene identification has to detect correct exon/intron boundaries in order to derive correct amino acid sequences. New or improved statistical methods provide a good starting point, but clearly fall short of 100% accuracy. The comparison of genomic sequence with datasets of cDNA sequences (full length or fragmentary) remains, therefore, essential for the correct identification of protein genes.

The availability of the complete set of proteins of one organism opens up an entirely new set of questions. For example, with certain completeness assumptions, one can now analyse functions that are expected from basic biochemical knowledge or known to be present, but for which no protein has been assigned. Schemes for the organization of linked networks of protein functions, in metabolic pathways [19•,20] or in regulatory cascades, are useful tools for the assignment of protein genes to biological pathways (this step of genome analysis should also be automated). Given a complete pathway description, apparently missing functions lead either to the hunt for undetected homologues or a reinterpretation of the pathway's enzymatic steps with the possible involvement of alternative enzymes [21•].

Comparative genomics

Once the technology, organization and determination was in place to finish the first cellular genome sequence, there was no stopping. Rapidly, detailed comparison of gene sets of different prokaryotic species has become possible, as well as the first such comparisons between prokaryotes, archaea and at least one small eukaryote (the yeast *S. cerevisiae*). The first such comparisons focused on the functional distribution of particular gene sets [21•,22•,23]. The spatial distribution of genes also has been studied, for example, it has been observed that gene order is not necessarily conserved in closely related

Table 2

Complete genomes and automatic functional assignment.

Organism	Size (Mb)	ORFs	Sequencing reference	3D (%)	f (%)	s (%)	GeneQuiz Reference
Bacteria							
Haemophilus influenzae Rd	1.8	1680	[13]	15	68	85	[74]
Mycoplasma genitalium	0.6	468	[66]	15	70	87	[75]
Mycoplasma pneumoniae	0.8	677	[67]	10	59	93	-
Synechocystis sp	3.6	3168	[68,69]	12	56	74	-
Escherichia coli	4.7	4285	[70]	14	72	87	-
Helicobacter pylori	1.7	1590	[71]	11	54	84	-
Bacillus subtilis	4.2	~4000	In press	-	-	-	-
Archaebacteria							
Methanococcus jannaschii	1.7	1735	[72]	8	45	74	[76]
Archaeoglobus fulgidus	2.2	-	TIGR, in preparation	-	-	-	-
Methanobacterium thermoautotrophicum	1.8	1871	Genome Therapeutics Corporation in preparation	-	-	-	-
Eucarya							
Saccharomyces cerevisiae	12.5	6284	[73]	11	60	77	[77]

3D, sequences for which a model could be built by similarity to a sequence of known 3D structure; f, sequences for which function is known or can be inferred from similar sequences; s, sequences that have similar sequences in the database.

species [24•]. Fundamental mathematical methods to describe gene rearrangement between genomes are being developed [25], but in practice it is already useful to use gene position for functional prediction without sequence similarity [26].

Comparative analysis of metabolic pathways in organisms with completely sequenced genomes helps in improving assignment of functionality based on sequence information (for example, see [27]) and on occasion has led to the possible discovery of new pathways or modification of the existing ones (e.g., the nitrogen assimilation cycle of *H. influenzae*) [20,21•,28]. Comparing the genomes of a pathogen and a non-pathogen can lead to the identification of pathogenicity genes [29•]. Proteins specific to a pathogen that have no (close) homologues in non-pathogenic organisms are good drug targets as they may be responsible for pathogenicity.

A fascinating set of scientific questions arise from the availability of sequences from families that have members in all three biological kingdoms. Such protein families may represent ancient proteins that rarely accept mutations in their sequences because of strong functional constraints; they also may have been an essential part of a simpler organism in early evolution and for this reason are often called 'important' [21•,30•–32•]. Families of widespread hypothetical proteins (proteins without assigned function, typically derived from large-scale sequencing projects) are very good targets for biochemical or genetic experimentation, as well as for determination of 3D structures.

Another set of proteins of particular experimental interest are those from model organisms with truly functional homology to proteins in more complex organisms. Very good examples are yeast proteins with similarity to human disease-related proteins (e.g., the MSH2 yeast gene homolog to the gene related to cystic fibrosis) [33,34]. On a technical point, the availability of complete genome sequences facilitates the more accurate identification of true homologues based on the distribution of similar sequences in each genome; however, the difficulties in distinguishing by sequence similarity alone orthology sequences (real functional homologs) from paralog (proteins with similar sequence and close but different functionality) persists in principle [35].

Transcript and expression profiles

Genome sequence provides a static picture, even when functional information is attached to all genes. The time dimension is being added, however, as powerful methods for the analysis of expression patterns (hybridization on devices containing a multiplicity of oligonucleotides each one of known sequence leading to specificity hybridization patterns, called DNA 'chips') begin to provide complete information about the role of genes in different cellular states [36•]. In parallel, new techniques in mass-spectrometry will soon provide proteome fingerprints (expression patterns of large sets of proteins under a variety of cellular conditions) from very small samples, using fragmentation of peptides that leads to partial sequence information, followed by lookup in sequence databases [37•]. These new high-throughput techniques of transcript (mRNA as ESTs) and protein profiling will soon provide massive amounts of data, complementing static genome sequences. Database groups, such as at the US National Center for Biotechnology Information, the DNA Database of Japan, and EMBL's European Bioinformatics Institute, will soon be faced with the challenge of integrating genome sequence data with other data emanating from large-scale molecular biology.

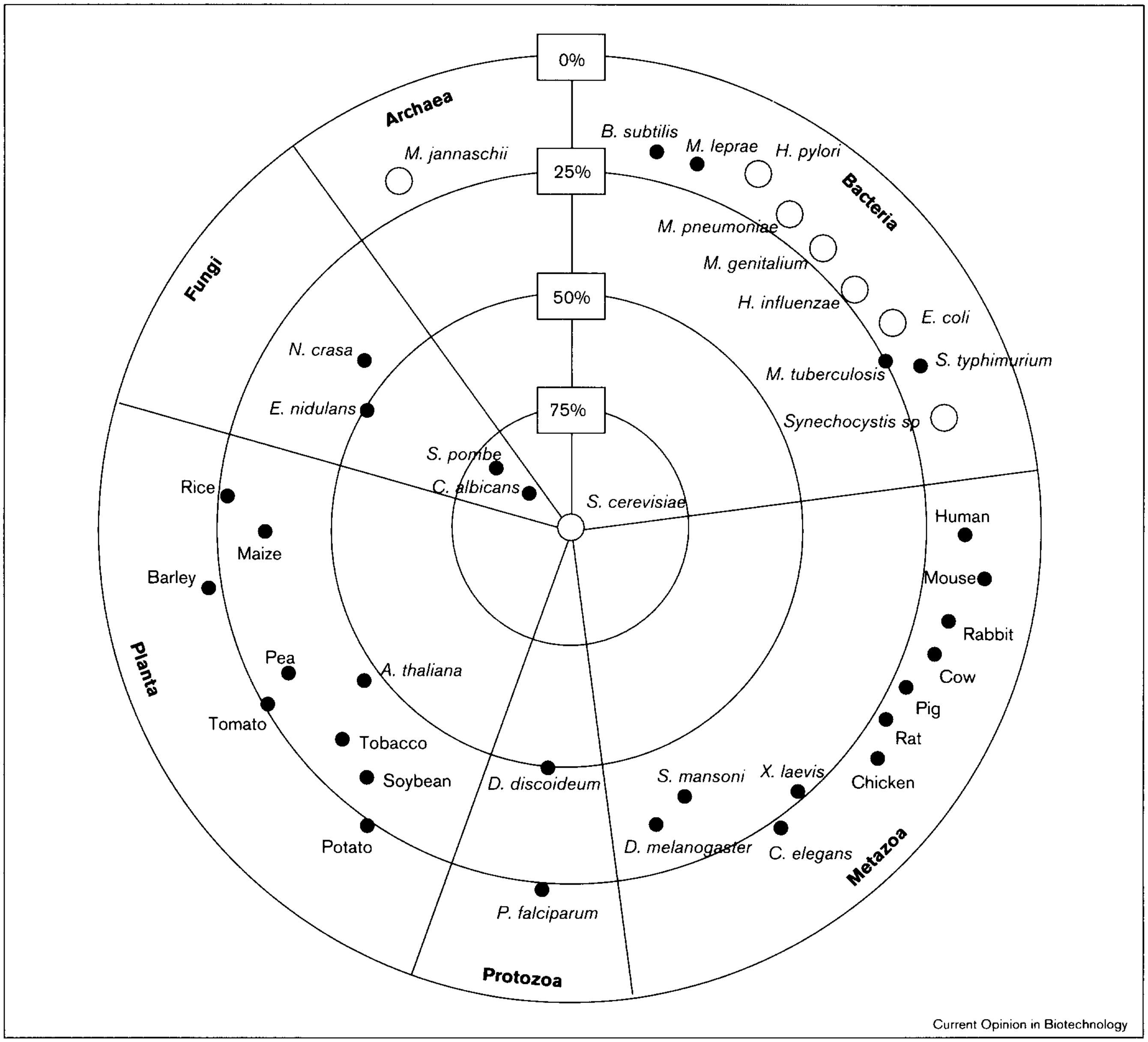
Conclusion and future directions

What concrete improvements in bioinformatics methods do we expect in the near future? Gene identification from nucleic acid sequence will become increasingly important as the nematode genome sequence nears completion (expected in 1998; R Durbin, private communication) and human genome sequencing ramps up to volume production. While algorithms for gene identification are being further refined, we foresee that EST cDNA fragments will saturate the set of all genes, making gene identification a matter of matching EST sequences against the genomic sequence from which they are derived and reducing the practical importance of the use of intrinsic information (derived only from the sequence itself without reference to other known sequences) in the statistical methods for ORF identification.

Improvements are also around the corner in protein sequence alignment algorithms, primarily in terms of better methods for matching sequence profiles, either against single sequences or against other profiles. Prediction of protein function from sequence, also called automated functional assignment, needs to be further automated and an improved set of rules derived for the expert modules of large-scale sequence analysis systems. More accurate hypotheses about the function of a protein can be put forward if full use is made of phylogenetic or subfamily analysis and of pathway mapping. In addition, development of a quantitative concept of similarity of function would remove a serious, conceptual and practical, bottleneck.

For the underlying databases of sequences and functions, there are at least two areas in need of improvement. One is catching up with the process of annotation (attachment of functional information to protein sequence data), that is, removing existing backlogs; this can, in part, be achieved through partial automation of the annotation process. Also, functional annotation by genome sequencing centers, which is now very heterogeneous (S Hoersch, MA Andrade, NP Brown, C Leroy, C Sander, unpublished data and M Ashburner, private communication), can be made more accurate and more formally correct. Second, in protein sequence datasets (e.g., Swissprot) pointers to the origin of functional information are needed where that information was derived by similarity arguments. The need arises from the possible inaccuracy of functional annotation propagated through a series of sequence similarity relations. Ideally, one should be able to trace all functional information to the original experiment, if necessary through a series of hyperlinks.

Figure 1



Current Opinion in Biotechnology

The world according to yeast. Emerging functional map of genomes, viewed by similarity with yeast, the first completely sequenced eukaryote. Each point represents the genome of an organism and is placed at a radial distance from the center corresponding to its functional similarity to the yeast genome. Functional similarity is calculated as the percentage of genes an organism shares with yeast, as detected by comparison of protein sequences. For example, 50% of the currently known genes in the slime mold *Dictyostelium discoideum* (bottom center) have at least one homologous or paralogous gene in yeast. The completely sequenced bacterial genomes (open circles) share about 20–25% of their genes with yeast; the human genome just under 20% (estimated). Functional similarity by gene content is a new measure of evolutionary similarity that requires knowledge of complete sets of genes. As the genomes for most organisms shown here are not yet fully sequenced, this multi-genome functional map is a first estimate. By the time the human genome sequence is finished (estimated by the year 2002), complete sets of genes of many genomes will have been probed for functional similarity, by experimental and computational tools, resulting in an increasingly complete and accurate functional map of all genomes. Significance of sequence similarity, interpreted as evidence of homology or paralogy, was assessed using GeneQuiz [11]. The placement of points in the azimuthal direction is the artist's approximation of grouping according to taxa. For incomplete genomes the percentages refer to the current set of genes of that organism in the Swissprot database.

Bioinformatics has still not solved the protein folding problem (calculating protein structure from amino acid sequence alone), in spite of tenacious efforts; however, interesting new methods for matching the properties of a set of structures to those of a set of multiple sequence alignments look encouraging—at least in cases for which all sequences contain a minimal functional signature. In any event, help is on the way in that a greater number of larger and more complex new protein 3D structures are solved by X-ray crystallography and NMR spectroscopy, albeit at a pace that lags behind genome sequencing. With the sequences of many genomes and more and more protein structures, the day is not far off when almost every new protein has a homologue of known structure such that a reasonably accurate 3D model can be built on the basis of sequence homology to a protein of known 3D structure. That day will signal what we call "the slow death of the (natural) protein folding problem". Meanwhile, structural biologists are faced with the challenge of determining at least one representative of all structural types of proteins, including membrane proteins, and to improve the rate of structure determination, perhaps through the kind of incremental engineering improvements that led to the decision to sequence human DNA on a large scale.

Complete genome sequences are recent arrivals. As such, they require new techniques and offer new kinds of knowledge. Comparative genomics needs new and refined measures of differences and similarities between genomes. It offers a magnifying glass under which we can observe the intricate reasons for the identity of a given organism. In time, a conceptual map of all genomes will emerge. A first, crude attempt at such a map is shown in Figure 1. New ways of comparing genomes will need to be invented.

Returning to practical applications, one of the first direct medical benefits from genome analysis will be in the area of microbial pathogens. Bioinformatics will contribute through the assignment of protein functional information, leading to the identification of drug targets and candidate molecules for vaccines [29]. Ideal targets for novel narrow-spectrum antibiotics are proteins that are essential for the survival of the pathogen and that are sub-species specific, that is, not present in other microbial species.

Once a target has been identified and validated, bioinformatics can contribute to the identification of a lead compound that would interact with the target protein. For example, multiple sequence alignment of the target with its evolutionary analogs is the input to an analysis of residue specificity (which residues are characteristic of particular protein subfamilies); if a reasonable 3D structure model can be produced exploiting sequence similarity to a protein of known structure (homology model building [38•,39,40]), then a refined molecular design on the background of structural knowledge of a series of target proteins of different specificity becomes feasible. Combinatorial chemistry can benefit directly as the diversity needed for a given target can be narrowed down using knowledge of the 3D structure model.

We are still in the era of the bacterial genomes. Most eukaryotic genomes are still to come but already the medical applications of genome projects are invaluable. Beyond the development of new antibiotics, antiviral drugs and vaccines, an entire range of highly specific diagnostic, preventive and therapeutic agents are on the horizon, optimally tailored to the individual's biological profile. Bioinformatics and genome analysis will contribute the coming revolution in pharmaceutical technology.

Acknowledgements

This work has been stimulated by the discussions within our group, which includes NP Brown, S Hoersch, L Holm, A Franchini, C Leroy and C Reich.

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. Pearson WR: **Effective protein sequence comparison.** *Methods Enzymol* 1996, **266**:227-258.

2. Altschul SF, Gish W: **Local alignment statistics.** *Methods Enzymol* 1996, **266**:460-480.

3. • Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs** *Nucleic Acids Res* 1997, **25**:3389-3402.

The latest version of the National Center for Biotechnology Iinformation's basic local alignment search tool programs for sequence comparison includes new ideas based on wide-spread experience with previous versions: the construction of gapped alignments and the ability to generate and apply profiles from significant hits in an iterative fashion. Not identical to WU-BLAST2 by Warren Gish, which also does gapped alignments.

4. Tatusov R, Altschul S, Koonin E: **Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks.** *Proc Natl Acad Sci USA* 1994, **91**:12091-12095.

5. Bork P, Gibson T: **Applying motif and profile searches.** *Methods Enzymol* 1996, **266**:162-184.

6. Livingstone C, Barton G: **Identification of functional residues and secondary structure from protein multiple sequence alignment** *Methods Enzymol* 1996, **266**:497-512.

7. • Holm L, Sander C: **Mapping the protein universe.** *Science* 1996, **273**:595-603.

An all-on-all comparison of 3D protein structures using the Dali method leads to a classification scheme for folds, an overview of the space of all known protein structures and provides tools leading to the unification of apparently unrelated protein families.

8. • Rossjohn J, Buckley J, Hazes B, Murzin A, Read R, Parker M: **Aerolysin and pertussis toxin share a common receptor-binding domain.** *EMBO J* 1997, **16**:3426-3434.

An example of an unexpected discovery of structural similarity leading to the unification of functionally but remotely related protein families and interesting hypotheses about functional apsects.

9. Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins.** *Nat Struct Biol* 1995, **2**:171-178.

10. • Andrade M, Casari G, Sander C, Valencia A: **Classification of protein families and detection of the determinant residues with an improved self-organizing map.** *Biol Cybern* 1997, **76**:441-450.

An algorithm using protein vectors in abstract sequence space. The vectors are classified by a neural network which derives the tree-determinant (subfamily-specific) residues.

11. Casari G, Ouzounis C, Valencia A, Sander C: *GeneQuiz II: Automatic Function Assignment for Genome Sequence Analysis.* Hawaii: World Scientific; 1996.

12. Etzold T, Ulyanov A, Argos P: **SRS: information retrieval system for molecular biology data banks.** *Methods Enzymol* 1996, **266**:114-128.

13. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.** *Science* 1995, **269**:496-512.

14. Burset M, Guigó R: **Evaluation of gene structure prediction programs.** *Genomics* 1996, **34**:353-367.

15. • Bucher P, Fickett J, Hatzigeorgiou A: **Computational analysis of transcriptional regulatory elements: a field in flux.** *Comput Appl Biosci* 1996, 12:361-362.

Summary of an international workshop containing references for the latest advances in the field. Identification of transcriptional elements is very important for the detection of genes in genome sequences.

16. • Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7**:399-406.

Description of the latest developments in the detection of regulatory element in human DNA and model vertebrates.

17. • Gelfand M, Mironov A, Pevzner P: **Gene recognition via spliced sequence alignment.** *Proc Natl Acad Sci USA* 1996, **93**:9061-9066.

Formalization of a commonly used method for gene detection in nucleotide sequences based on finding matching expressed sequence tags (cDNA fragments) and/or proteins using dynamic programming.

18. • Das S, Yu L, Gaitatzes C, Rogers R, Freeman J, Bienkowska J, Adams R, Smith T, Lindelien J: **Biology's new Rosetta stone.** *Nature* 1997, **385**:29-30.

The authors express their concerns about the quality of results from the yeast genome project. They point out problems with some of the annotations derived by sequence similarity rather than experimentation and with putative short open reading frames that may not be expressed.

19. • Kanehisa M: **Toward pathway engineering: a new database of genetic and molecular pathways.** *Sci Technol Japan* 1996, **59**:34-38.

Development of one of the best and most compehensive collections of information on gene function and metabolic pathways.

20. Karp P, Riley M, Paley S, Pellegrini-Toole A, Krummenacker M: **EcoCyc: enyclopedia of *Escherichia coli* genes and metabolism.** *Nucleic Acids Res* 1997, **25**:43-51.

21. • Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: **Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*.** *Curr Biol* 1996, **6**:279-291.

Comparison of these two bacteria resulted in assignment of many genes to metabolic pathways.

22. • Ouzounis C, Casari G, Sander C, Tamames J, Valencia A: **Computational comparisons of model genomes.** *Trends Biotechnol* 1996, **14**:280-285.

Genome comparison needs new methodologies to cope with the aggregate data from multiple sequences. Here, relative gene position is used for functional classification and gene functional class is derived by an automated procedure based on keywords associated to homologs.

23. Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R: **Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*.** *Nucleic Acids Res* 1997, **25**:701-712.

24. • Kolstø A: **Dynamic bacterial genome organization** *Mol Microbiol* 1997, **24**:241-248.

Comparison of the first full bacterial genomes shows an extraordinary complexity and variability of gene order.

25. Blanchette M, Kunisawa T, Sankoff D: **Parametric genome rearrangement.** *Gene* 1996, **172**:GC11-GC7.

26. Tamames J, Casari G, Ouzounis C, Valencia A: **Conserved clusters of functionally related genes in two bacterial genomes.** *J Mol Evol* 1997, **44**:66-73.

27. Danchin A: **Comparison between the Escherichia coli and Bacillus subtilis genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP.** *DNA Res* 1997, **4**:9-18.

28. Karp P, Ouzounis C, Paley S: **HinCyc: a knowledge base of the complete genome and metabolic pathways of H. influenzae.** *Ismb* 1996, **4**:116-124.

29. • Strauss E, Falkow S: **Microbial pathogenesis: genomics and beyond.** *Science* 1997, **276**:707-712.

Review on the use of genomic bacterial information for elucidation of bacterial biochemistry with stress on pathogenicity.

30. • Ouzounis C, Kyrpides N: **The emergence of major cellular processes in evolution** *FEBS Lett* 1996, **390**:119-123.

Analysis of the species distribution of set of protein families leads to a hypothesis about the evolutionary split between the three kingdoms.

31. • Mushegian A, Koonin E: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci USA* 1996, **93**:10268-10273.

An analysis of the conserved proteins between the *H. influenzae* and the *M. genitalium* bacterial genomes. A set of conserved proteins is proposed as representing a minimal cell.

32. • Koonin E, Mushegian A, Galperin M, Walker D: **Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, **25**:619-637.

Through the comparison of the first complete genomes and the kingdom expansion of their proteins the authors conclude that *M. jannaschii* has a very heterogeneous behaviour by protein functional class.

33. Bassett DJ, Boguski M, Hieter P: **Yeast genes and human disease.** *Nature* 1996, **379**:589-590.

34. Mushegian A, Bassett D, Boguski M, Bork P, Koonin E: **Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs.** *Proc Natl Acad Sci USA* 1997, **94**:5831-5836.

35. Koonin E, Mushegian A, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12**:334-336.

36. • Richard G, Fairhead C, Dujon B: **Complete transcriptional map of yeast chromosome XI in different life conditions.** *J Mol Biol* 1997, **268**:303-321.

Transcriptional maps of full genomes provide experimental evidence for gene function. Such techniques are expected to grow in importance. Here is one of the first large scale examples.

37. • Mann M, Talbo G: **Developments in matrix-assisted laser desorption/ionization peptide mass spectrometry.** *Curr Opin Biotechnol* 1996, **7**:11-19.

New advances in mass spectrometry of proteins lead to very sensitive and efficient detection of proteins from delicate biological samples.

38. • Chinea G, Padrón G, Hooft RWW, Sander C, Vriend G: **The use of position specific rotamers in model building by homology.** *Proteins* 1995, **23**:415-421.

New rules for model building 3D protein structures by homology, as incorporated in WHATIF, a comprehensive protein analysis and engineering package.

39. Peitsch M: **ProMod, Swiss-Model: internet-based tools for automated comparative protein modelling.** *Biochem Soc Trans* 1996, **24**:274-279.

40. Sánchez R, Sali A: **Advances in comparative protein-structure modelling.** *Curr Opin Struct Biol* 1997, **7**:206-214.

41. Stoesser G, Sterk P, Tuli MA, Stoehr PJ, Cameron GN: **The EMBL nucleotide sequence database.** *Nucleic Acids Res* 1997, **25**:7-14.

42. Benson DA, Boguski MS, Lipman DJ, Ostell J: **GenBank.** *Nucleic Acids Res* 1997, **25**:1-6.

43. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its new supplement TrEMBL.** *Nucleic Acids Res* 1997, **25**:31-36.

44. George D, Dodson R, Garavelli J, Haft D, Hunt L, Marzec C, Orcutt B, Sidman K, Srinivasarao G, Yeh L *et al.*: **The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database.** *Nucleic Acids Res* 1997, **25**:24-28.

45. Borodovsky M, McIninch JD: **GENEMARK: parallel gene recognition for both DNA strands.** *Comput Chem* 1993, **17**:123-133.

46. Uberbacher E, Xu Y, Mural R: **Discovering and understanding genes in human DNA sequence using GRAIL.** *Methods Enzymol* 1996, **266**:259-281.

47. Brown NP, Sander C, Bork P: **A web tool for detection of genomic errors.** *Comput Appl Biosci* 1997, in press.

48. Gaasterland T, Sensen CW: **Fully automated genome analysis that reflects user needs and preferences – a detailed introduction to the MAGPIE system architecture.** *Biochimie* 1996, **78**:302-310.

49. Frishman D, Hans-Werner M: **PEDANtic genome analysis.** *Trends Genet* 1997, **13**:415-416.

50. Bairoch A, Bucher P, Hofmann K: **The PROSITE database, its status in 1997.** *Nucleic Acids Res* 1997, **25**:217-221.

51. Henikoff JG, Henikoff S: **Blocks database and its applications.** *Methods Enzymol* 1996, **266**:88-104.

52. Sonnhammer E, Eddy S, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments.** Proteins 1997, **28**:405-420.

53. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein DataBank: a computer based archival file for macromolecular structures.** *J Mol Biol* 1977, *112*:535-542.

54. Holm L, Sander C: **Dali/FSSP classification of three-dimensional protein folds.** *Nucleic Acids Res* 1997, **25**:231-234.

55. Hubbard T, Murzin A, Brenner S, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res* 1997, **25**:236-239.

56. Kabsch W, Sander C: **Dictionary of protein secondary structure:pattern recognition of hydrogen bonded and geometrical features.** *Biopolymers* 1983, **22**:2577-2637.

57. Hobohm U, Sander C: **Enlarged representative set of protein structures.** *Protein Sci* 1994, **3**:522-524.

58. Schneider R, de Daruvar A, Sander C: **The HSSP database of protein structure-sequence alignments.** *Nucleic Acids Res* 1997, **25**:226-230.

59. Hooft RWW, Scharf M, Sander C, Vriend G: **The PDB finder database: a summary of PDB, DSSP and HSSP information with added value.** *CABIOS* 1996, **12**:525-529.

60. Peitsch M: **ProMod and Swiss-Model: Internet-based tools forautomated comparative protein modelling.** *Biochem Soc Trans* 1996, **24**:274-279.

61. Chinea G, Padrsn G, Hooft RWW, Sander C, Vriend G: **The use of position specific rotamers in model building by homology.** *Proteins* 1995, **23**:415-421.

62. Gouzy J, Corpet F, Kahn D: **Graphical interface for ProDom domainfamilies.** *Trends Biochem Sci* 1996, **21**:493.

63. Bork P, Bairoch A: **Extracellular protein modules: a proposed nomenclature.** *Trends Biochem Sci* 1995, **20**:poster C02.

64. Riley M: **Genes and proteins of Escherichia coli K-12 (GenProtEC).** *Nucleic Acids Res* 1997, **25**:51-52.

65. Fasman K, Letovsky S, Li P, Cottingham R, Kingsbury D: **The GDB Human Genome Database Anno 1997.** *Nucleic Acids Res* 1997, **25**:72-81.

66. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischman RD, Bult CJ, Kerlavage AR, Sutton G, Kelly JM *et al.*: **The minimal gene complement of Mycoplasma genitalium.** *Science* 1995, **270**:397-403.

67. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li B, Herrmann R: **Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae.** *Nucleic Acids Res* 1996, **24**:4420-4449.

68. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S *et al.*: **Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.** *DNA Res* 1996, **3(suppl)**:185-209.

69. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S *et al.*: **Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.** *DNA Res* 1996, **3**:109-136.

70. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al.*: **The complete genome sequence of Escherichia coli K-12.** *Science* 1997, **277**:1453-1462.

71. Tomb J, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA: **The complete genome sequence of the gastric pathogen Helicobacter pylori.** *Nature* 1997, **388**:539-547.

72. Bult CJ, White OW, Olsen GJ, Zhou LZ, Fleischmann RD, Granger G, Sutton GG, Blake JA, Fitzgerald LM, Clayton RA *et al.*: **Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii.** *Science* 1996, **273**:1058-1073.

73. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: **Life with 6000 genes** *Science* 1996, **274**:546-567.

74. Casari G, Andrade MA, Bork P, Boyle J, Daruvar A, Ouzounis C, Schneider R, Tamames J, Valencia A, Sander C: **Challenging times for bioinformatics.** *Nature* 1995, **376**:647-648.

75. Ouzounis C, Casari G, Valencia A, Sander C: **Novelties from the complete genome of Mycoplasma genitalium.** *Mol Microbiol* 1996, **20**:897-899.

76. Andrade MA, Casari G, Daruvar A, Sander C, Schneider R, Tamames J, Valencia A, Ouzounis C: **Sequence analysis of the Methanococcus janaschii genome and the prediction of protein function.** *Comp Appl Biosci* 1997, **13**:481-483.

77. Casari G, de Daruvar A, Sander C, Schneider R: **Bioinformatics and the discovery of gene function.** *Trends Genet* 1996, **12**:244-245.