Article No. mb982144 J. Mol. Biol. (1998) 283, 707–725

JMB

REVIEW

Predicting Function: From Genes to Genomes and Back

Peer Bork*, Thomas Dandekar, Yolande Diaz-Lazcoz
Frank Eisenhaber, Martijn Huynen and Yanping Yuan

European Molecular Biology Laboratory, Meyerhofstr. 1 PF 10.2209, D-69117 Heidelberg, Germany and Max-Delbrück-Centrum für Molekulare Medizin Robert-Rössle-Str. 10 D-13122 Berlin-Buch, Germany

**Corresponding author*

Predicting function from sequence using computational tools is a highly complicated procedure that is generally done for each gene individually. This review focuses on the added value that is provided by completely sequenced genomes in function prediction. Various levels of sequence annotation and function prediction are discussed, ranging from genomic sequence to that of complex cellular processes. Protein function is currently best described in the context of molecular interactions. In the near future it will be possible to predict protein function in the context of higher order processes such as the regulation of gene expression, metabolic pathways and signalling cascades. The analysis of such higher levels of function description uses, besides the information from completely sequenced genomes, also the additional information from proteomics and expression data. The final goal will be to elucidate the mapping between genotype and phenotype.

Keywords: genomes; computational tools; function prediction; comparative genome analysis; proteomics

Genomes and function prediction

Prediction of protein function using computational tools becomes more and more important as the gap between the increasing amount of sequences and the experimental characterization of the respective proteins widens (Bork & Koonin, 1998; Smith, 1998). With the availability of complete genomes we face a new quality in the prediction process (Table 1) as context information can be utilized when analysing particular sequences. This review focuses on the added value of genomic information on the many steps of function prediction from genomic sequence. The first reports on completely sequenced genomes give an excellent overview of the evolving state of the art in the analyses of particular genomes (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995, 1998; Himmelreich *et al.*, 1996; Goffeau *et al.*, 1996; Kaneko *et al.*, 1996 Blattner *et al.*, 1997; Tomb *et al.*, 1997; Kunst *et al.*, 1997; Bult *et al.*, 1996; Smith *et al.*, 1997; Klenk *et al.*, 1997). In addition, there are numerous reviews that touch on the extraction of functional features from sequence (e.g. Bork *et al.*, 1994; Andrade *et al.*, 1997; Koonin & Galparin, 1997; Bork & Koonin, 1998), but very few reviews have been published that systematically summarize the additional information for function prediction that is provided by the presence of entirely sequenced genomes (original papers e.g. by Mushegian & Koonin, 1996a,b; Himmelreich *et al.*, 1997; Koonin *et al.*, 1997; Tatusov *et al.*, 1996, 1997; Huynen & Bork, 1998; Huynen *et al.*, 1997, 1998a; Dandekar *et al.*, 1998b).

What is function?

"Function" is a very loosely defined term that only makes sense in context. Most current efforts aim at predicting protein function, but there are other types of function, e.g. RNA function or organelle function, that also need to be explored. Even to describe "protein function" requires a broad range of attributes and features (Figure 1). Molecular features such as enzymatic activity, interaction

Present address: Y. Diaz-Lazcoz, Laboratoire Genome et Informatique; Batiment BUFFON, Universite de Versailles-Saint Quentin, 45, avenue des Etats-Unis, 78035 Versailles Cedex, France.

E-mail address of the corresponding author: Bork@EMBL-Heidelberg.de, Dandekar@EMBL-Heidelberg.de, Yolande.Diaz@genetique.uvsq.fr Eisenhaber@EMBL-Heidelberg.de, Huynen@EMBL-Heidelberg.de, Yuan@EMBL-Heidelberg.de

Table 1. Added features from complete genome analysis for function prediction

Genome specific patterns in the DNA and their usage in genome annotation		
Feature:		Genome-specific (poly)nucleotide frequencies, codon usage
Usage	→	Identification of genes
	→	Identification of recent horizontal gene transfers into the genome
Feature:		Genome-specific signal sequences like regulatory regions, promotors
Usage	→	Gene identification, identification of the mode of regulation of genes, regulatory regions in mRNA, specification of the boundaries of genes
	→	Operon identification
Usage of the complete set of genes in a genome and comparative genome analysis		
Feature:		The finding of orthologs by comparative genome analysis
Usage	→	Narrowing down the function of a gene
	→	Identification of (conserved) regulatory signals neighbouring the orthologues
Feature:		Conserved genome organization
Usage	→	Genes in a conserved clusters have related functions, show physical interaction
Feature:		Differential genome analysis
Usage	→	Identification of the functions that are absent from a genome
	→	If an orthologous gene is absent, but the function is present, missing genes point either to a wrong annotation or a non-orthologous gene transfer
	→	Identification of the functions that are specific to a genome, and might be responsible for the species' specific phenotype, delineation of the mapping between genotype and phenotype

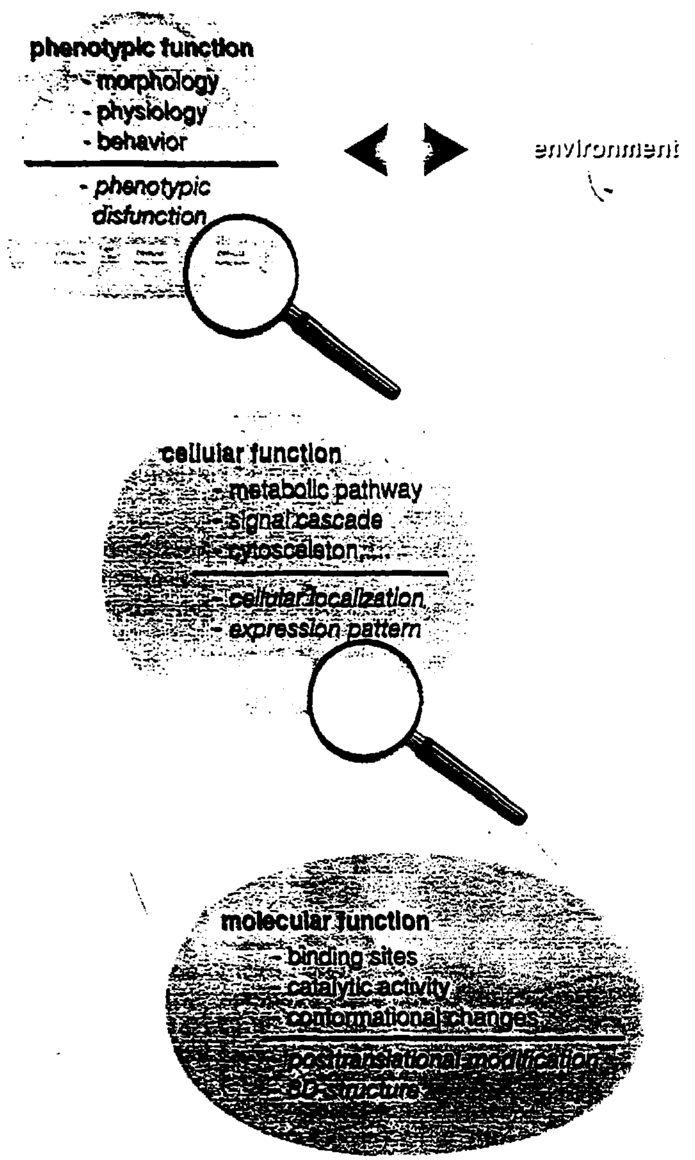


Figure 1. Characterization of protein function. Whereas nucleic acids fulfil the tasks of storage, transfer and processing of genetic information contained in the genome of living organisms, the proteins (gene products) form a complex (single- or multi-) cellular machinery for the realization of this genetic program (resulting in the phenotype) in dependency and in response to changing environment conditions. Therefore, protein function requires also a multilevel, hierarchical description comparable with the notions of primary, secondary, tertiary, and quaternary protein structure. Here, we propose a possible framework for functional characterization and, for each hierarchical level, both functional features and attributes are described. It should be noted that, for most proteins, a quantitative functional characterization is still a matter of the future and, today, a qualitative description of function for at least some hierarchical levels can be considered an achievement for many proteins. (1) Each protein has molecular (elementary) functions; e.g. it can have specific binding sites for substrates, low-molecular effectors, nucleic acids or other proteins. Given the set of allowed allosteric conformational changes, of possible interactions with other molecules, and of kinetic properties, etc., the protein can, for example, catalyse a metabolic reaction, it may transmit a signal to other proteins or DNA or be able to fit into cytoskeletal macromolecular associates. Structural properties of a protein are attributes for the execution of function, therefore, 3D structural information greatly facilitates understanding of function. Also, possible posttranslational modifications such as glycosylation, propeptide cleavage, or protein splicing are an important preposition for a protein to fulfil its molecular function. (2) A set of many co-operating proteins is responsible for a physiological (cellular) function (metabolic pathway, signal transduction cascade, structural associate etc.). The cellular function of a protein is always context-dependent and is characterized by taxon, organ, tissue, etc. Subcellular localization is an essential attribute for this level. For proper functioning, the protein has to be translocated to the correct intra- or extracellular compartments in a soluble form or to be attached to a membrane. All types of regulation of protein activity are another attribute. For example, the amount of protein molecules is often controlled *via* gene expression which might be limited to certain types of cells or tissues or to specific periods in the cell cycle or the individual ontogenese (expression pattern). (3) Finally, the totality of the physiological subsystems and their interplay with various environmental stimuli determines phenotype properties (phenotypic

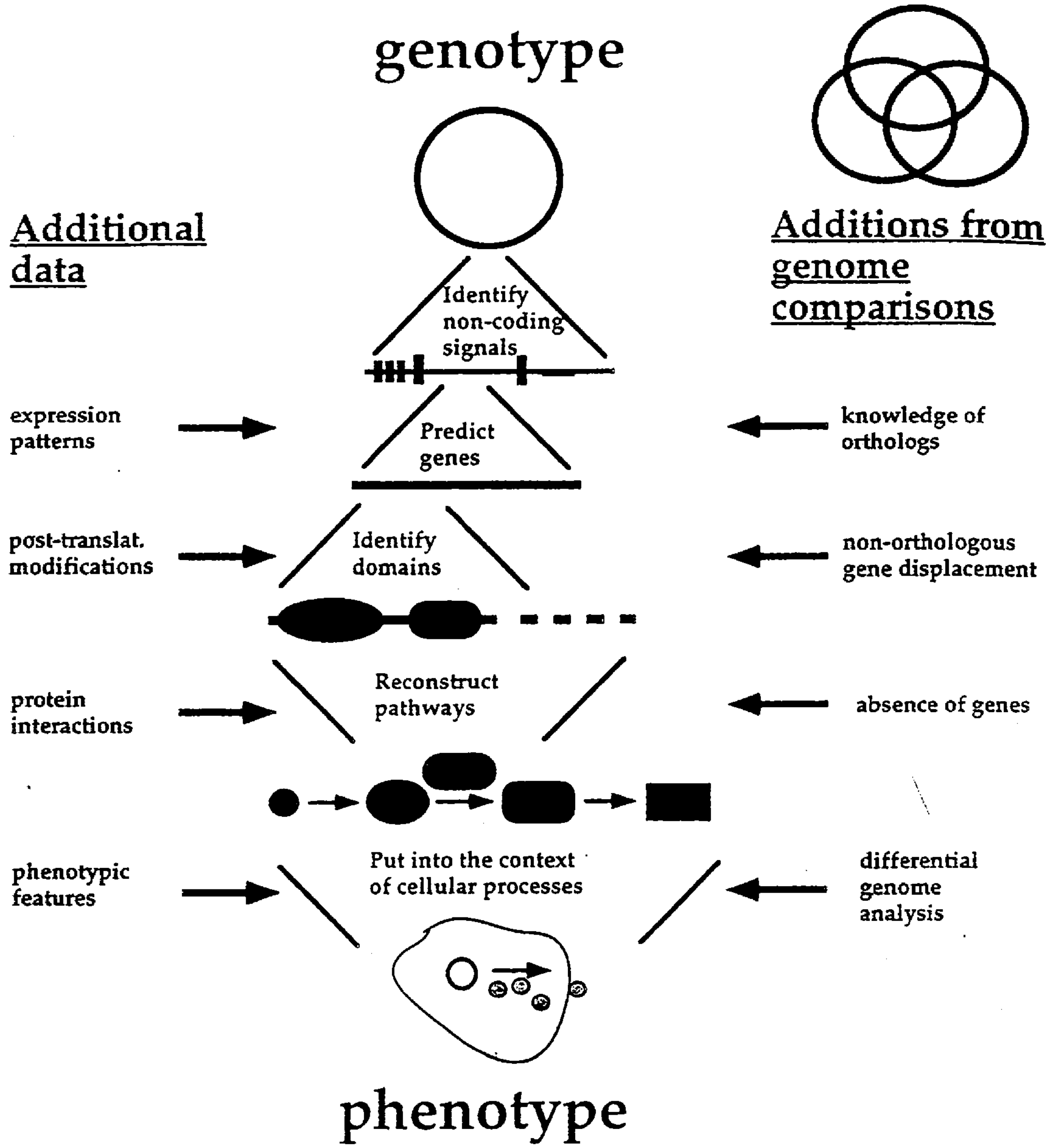


Figure 2. Function prediction scheme: zooming in and out. Whether gene, contig or genome: current methods concentrate on gene prediction and the annotation of individual genes that are then put into context. Due to our limited understanding of the genome, this is only possible by accessing complementary experimental information generated among others by proteomics research. Nevertheless, exploitation of genome information provides additional hints. This review follows the individual steps.

functions should be analysed in the biological context of the organism considered. Unfortunately, the functional knowledge of proteins reflected in their annotation (Figure 1) is frequently incomplete, sometimes erroneous or inconsistent, and often only cellular or even phenotypic functions are listed. For example, the human glia maturation factor (P17774) is described as growth factor (by definition extracellular!) but an in-depth sequence analysis revealed ADP-domains characteristic for cytoskeletal proteins (intracellular!).

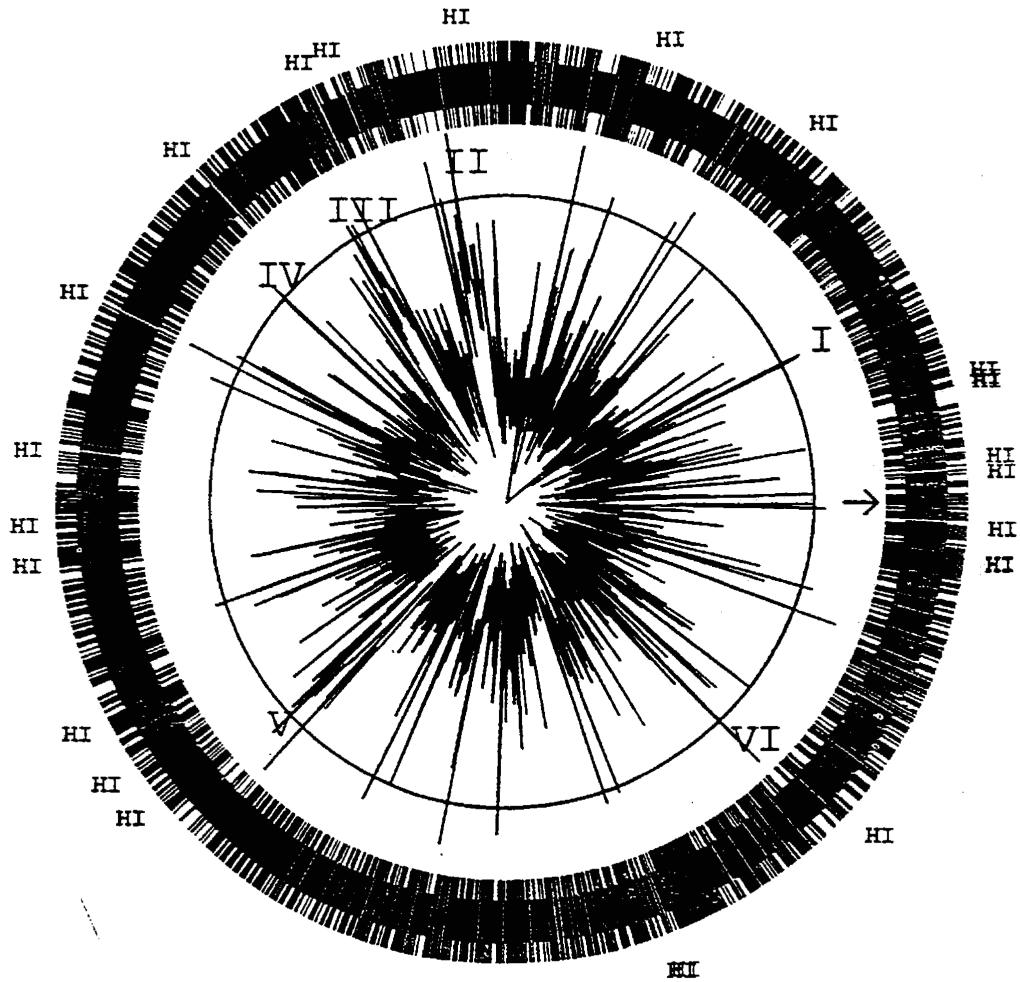
Sequence and annotation quality in molecular databases

Function transfer by analogy requires knowledge about the quality of sequence data and functional annotation. Concerns have been raised about an accumulation (Bork & Bairoch, 1996) and even an explosion (Bhatia *et al.*, 1997) of errors in sequence databases.

In genome projects, two to tenfold sequence coverage is usually sampled. This is critical as automated raw data acquisition (single read) is less than 99% accurate even when using optimized sequencers and software (Ewing *et al.*, 1998). Most of the ESTs (expressed sequence tags, i.e. single or sometimes double gel reads from cDNA) stored in current databases have much lower quality and require special caution as they are often also contaminated by cloning vectors or DNA of other sources including non-coding regions. More reasonable accuracy (at least over 99.85%) in all regions can be achieved only by systematic multiple coverage (Richterich, 1998). Nevertheless this

duction cascades, etc.) and phenotypic functions. However, functions are currently annotated in the form of plain text incorporating a large variety of vocabulary for the in-depth description of particular phenomena. Thus, they are not easily retrievable with keyword search engines such as SRS (Etzold *et al.*, 1996).

Computer-readable hierarchical systems of function description as envisioned in Figure 1 might be helpful, but controlled vocabularies such as in FLYBASE (ftp.ebi.ac.uk/pub/databases/edgp/misc/ashburner/fly_function_tree), the keywords in



Helicobacter pylori

———	Genes with a homolog in the *E. coli* genome or a homolog in a non pathogenic *E. coli* plasmid
———	Genes (putatively) involved in interaction with the host
HI	Genes with a homolog in *H. influenzae* but not in *E.coli*

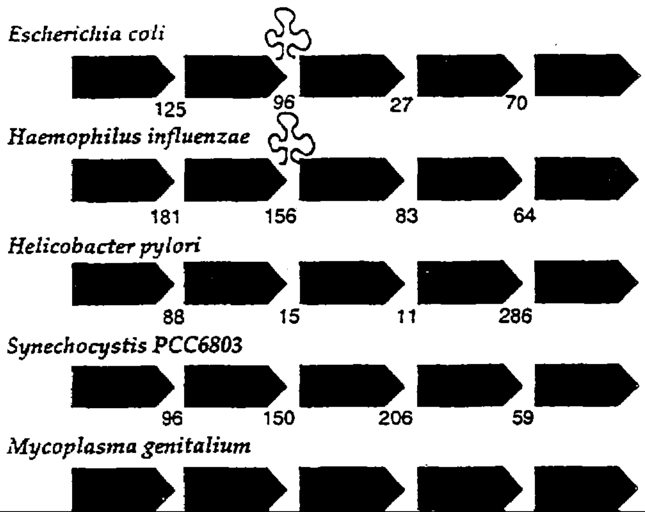


Figure 4. Variability of regulatory regions. Structure of the *str* operon and surrounding genes in eight Bacteria and two Archaea. The organisation of the *str* operon in *Archaeoglobus fulgidus* is essentially the same as in *M. jannaschii*, while the structure of the *str* operon in *Mycoplasma pneumoniae* is the same as that in *M. genitalium*. The arrows indicate the direction of transcription, the numbers under the arrows the lengths of the intergenic regions. Genes in green are orthologous to the genes in the same location in *E. coli*. Genes in light-red are shared at that location

1997). However, an analysis of the accuracy of all available packages for the prediction of coding

not exhaustively annotated (which is a huge amount of work). BLOCKS (Henikoff *et al.*, 1998) is

specific phenotypes, as can be shown in the comparison of the pathogenic *H. influenzae* with the closely related but relatively benign *E. coli*. A large fraction (70%) of the genes in *H. influenzae* for which there are no homologues in *E. coli* and for which some functional annotation is possible can indeed be considered host interaction factors (Huynen *et al.*, 1997). Also in the pathogen *H. pylori* the fraction of genes that is not shared with *E. coli* is relatively enriched in host interaction factors (Figure 3). Taking differential genome analysis one step further one can show how gene content correlates with phenotype in multiple genome comparisons (Huynen *et al.*, 1998a). Although the correlations between gene content and phenotype cannot be used to predict the function of specific genes, they can serve as a filter to select genes that are probably responsible for specific functions. Or, in other words, to search for "genes for a function" rather than to search for "functions for a gene".

Incorporating proteomics data

Proteomics focuses on the protein products of the genome and their interactions rather than on DNA sequences (Humphery-Smith & Blackstock, 1997). It is thus complementary to the genomic and nucleic acid information (Kahn, 1995) exploiting novel tools such as 2D large scale analysis (Vietor & Huber, 1997) and powerful mass-spectrometry applications (Yates, 1998).

Protein identification and gene expression

Protein reading frames and expression behaviour in particular are not easy to predict from the genome sequence and profit from incorporation of additional experimental data (Figure 2, left). Co-expression as well as tissue- and organ-specific expression patterns at genomic scale are intensively studied (Hieter & Boguski, 1997; Zhang *et al.*, 1997) and recent techniques collect data on a genomic level.

Expressed sequence tag (EST) databases are available which contain information on gene expression that should correlate with the amount of redundancy, and on the tissue distribution of mRNA which can yield complex expression patterns (Boguski *et al.*, 1994; Zweiger & Scott, 1997). However, retrieval of this information is hampered by the high sequence error rate, by different splicing variants and by the often missing 5′ region necessary to determine the exact CDS start. Another caveat is that the EST approach has difficulties measuring genes with low expression.

Serial analysis of gene-expression (SAGE, Velculescu *et al.*, 1995) is a more rapid method to obtain partial sequence information from a very large set of expressed genes, e.g. differences in gene expression profiles in normal and cancer cells are identified by hundreds of differentially expressed transcripts, many of them growth factors (Zhang *et al.*, 1997). DNA chip-based gene-expression screening procedures are currently the fastest approach. Polymorphism with single base resolution is detected within minutes in the entire human mitochondrial genome (16.6-kilobases) by applying 135,000 probes simultaneously (array generated by light-directed chemical synthesis) and a two-colour fluorescent labelling scheme (Chee *et al.*, 1996). Systematic PCR of the entire yeast genome allows fluorescent readout of mRNA levels in different yeast environmental conditions such as changing glucose concentrations (DeRisi *et al.*, 1997; http://cmgm.stanford.edu/pbrown/explore). The correlation between mRNA and protein expression level is, however, debatable. Anderson & Seilhamer (1997) give a correlation coefficient 0.48 for expression levels in human liver measured either by two-dimensional electrophoresis (protein abundances) or by transcript image methodology (mRNA abundance measured by cDNA sequencing and cDNA clone count).

Direct determination of the major expressed proteins may thus be an independent and attractive alternative. The huge amount of work involved in this can today be substantially reduced by applying 2D gels and mass spectrometry and comparing experimental data to the annotated and predicted genome sequence. Link *et al.* (1997) identify the major part of the proteins and protein complexes from *H.influenzae* (300 out of 400 spots) after liquid chromatography (LC) and separation of the protein cleavage products of each 2D gel spot in a first mass spectrograph (MS) and further analysis in a second (LC/MS/MS approach). Several proteins not annotated in the genome sequence were identified by this approach.

Posttranslational modifications

After translation many proteins are further processed. This includes chemical modification of amino acids. Over 200 amino acid modification types are classified (Krishna & Wold, 1997), many more are expected (Annan & Carr, 1997). Such modifications are not apparent from the genome sequence, however, they are often critical for protein function. Two-dimensional gel electrophoresis coupled to mass spectrometry and modern software allows not only peptide mass fingerprinting for low quantities (Küster & Mann, 1998) but also specific detection of amino acid modifications on a large scale (Dongre *et al.*, 1997). For this, a database has to cover many of the reading frames likely to be encountered in the protein mixture analysed by the mass spectrometer. The EXPASY server (www.expasy.ch/www/tools.html) comprehensively links 2D gel experiments (e.g. separation from pH 4.0 to above 8.0 in the first dimension and from M_r 8-200 kDa in the second) to computer analysis tools. Nevertheless, determination of e.g. sugar modifications both by experiment and by software (e.g. EXPASY suite above) has limited accuracy, even including the kind of carbohydrate attached.

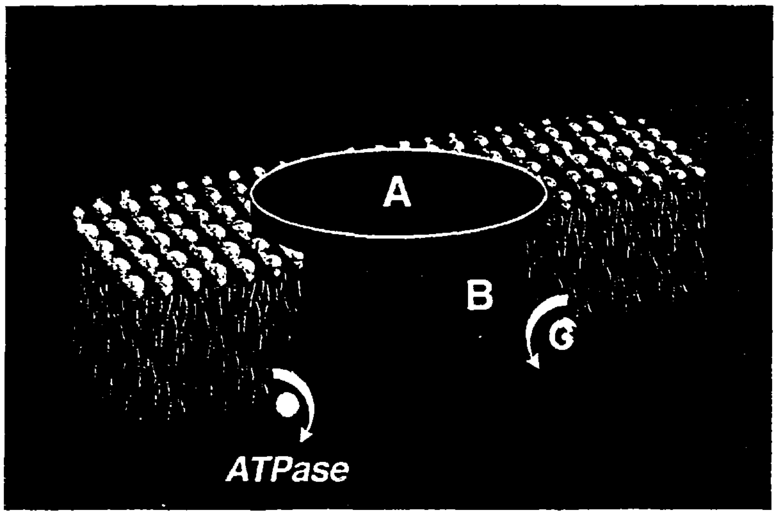


Figure 5. Protein interactions. Shown are ABC transporter proteins in the membrane of Gram-negative bacteria. Encoding genes are found as conserved gene clusters in the same sequential order in the complete genome sequences of *E. coli, H. influenzae* and *H. pylori.* They are an example of protein interactions predicted by triple comparison of complete genomes and additional confirmation by standard methods (see the text). A number of other protein interactions can also be suggested by comparative analysis of complete genomes.

Predicting function in higher order processes

Having predicted or determined functions for as many genes as possible and having assigned their interactions as well as their expression levels, it is a challenging task to put all the information into the context of cellular processes (Figure 1). A variety of databases and tools are emerging to support this procedure.

Information on tissue distribution

On the molecular level the processing machinery for metabolites differs in diverse tissues including absence of enzymes, receptors and structural proteins. On higher levels such as organ function, clinical impairment, drug metabolism or susceptibility to infections, tissue and phenotypic specific expression differences are key features of differentiation and help to find substances of therapeutical value. Data for humans are provided e.g. by TIGR (www.tigr.org/tdb/hgi/hgi.html), by NCBI (www.ncbi.nlm.nih.gov/UniGene/index.html and www.ncbi.nih.gov/dbEST/index.html), by SANBI-South African National Bioinformatics Institute (www.sanbi.ac.za/Dbases.html) and by the MRC human genetics unit (glengoyne.hgu.mrc.ac.uk). Such data should be used critically as low expression transcripts important for regulation such as tyrosine kinases may escape detection by EST sequencing or even Northern blots, and hence are misrepresented in databases. Techniques are still being improved (e.g. DNA chips (Brown, 1994)) and many data are not yet on the Web or are even completely inaccessible (e.g. in companies).

Analysis of protein interactions

Prediction and analysis of protein interaction uses both experimental (e.g. antibody precipitation, Maniatis *et al.*, 1989) and theoretical approaches. Two hybrid screening systems (Tsukamoto *et al.*, 1997) allow large scale screening, e.g. for 20 residue peptide sequences that correctly recognize (so called "aptamers") and inhibit cyclin-dependent kinase 2 (Colas *et al.*, 1997). Automation (with a considerable error rate though) and matching the data gathered with context and information such as common pathways is possible (Brent & Finley, 1997). Logical connections of protein interactions (e.g. with ras protein) can be revealed by a careful choice of reporter plasmids (Xu *et al.*, 1997).

A new way to identify protein interactions, comparative analysis between genomes, has revealed that the conservation of gene order between genomes with less than 50% protein identity is limited to those genes that code for proteins that physically interact with each other (Dandekar *et al.*, 1998b). Protein candidates for physical interaction that are identified by the conservation of their gene order can further be analysed by the methods mentioned above. An example are the ABC transporters which were experimentally shown to consist of physically interacting proteins (Eym *et al.*, 1996) and are found in conserved gene clusters in different genomes (Figure 5). The conservation of gene order can of course be used for the prediction of functional features of hypothetical proteins (interaction with a neighbour and, if this one is characterized, even participation in a pathway).

Reconstruction of pathways

The prediction of reactions and pathways (example: Figure 6) of the respective organisms integrates all the data above (including errors at different levels!) into its phenotypic context and yields a more complete picture of the biochemical and adaptive capabilities of the sequenced organism (Overbeek *et al.*, 1996). Mispredictions, wrong annotations and higher level errors (substrate specificity etc.) have to be minimized by context information and additional experimental data. Problems specific to pathway predictions arise, such as non-orthologous displacements (enzymatic

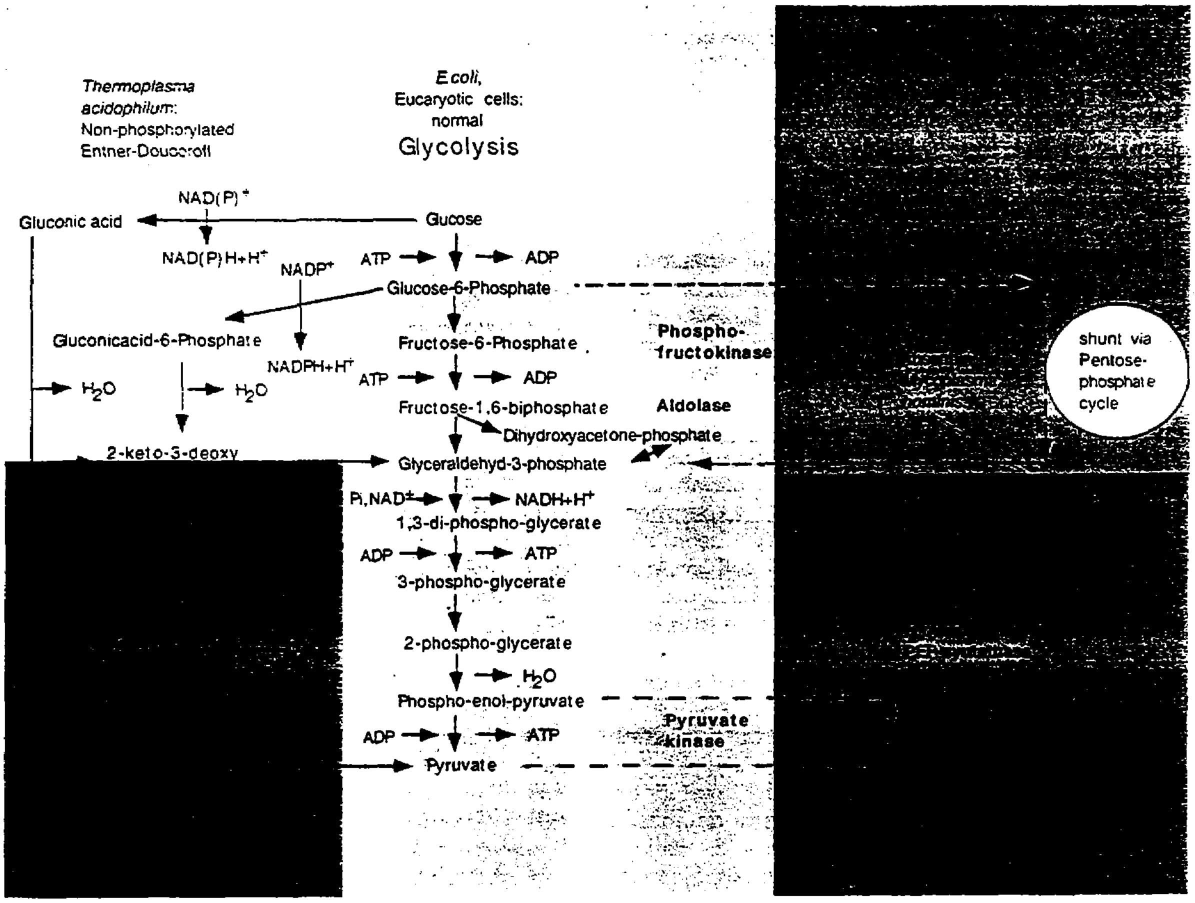


Figure 6. Prediction of metabolic pathways and pathway alignment. The glycolytic pathway (centre) and alternative other routes (sides) predicted from the genome and observed in several microorganisms are shown and compared (pathway alignment) to illustrate the often underestimated variability of metabolic pathways. Key enzymes discussed are shown in bold. In the glycolytic pathway (centre) two molecules of triose are derived from one hexose (as dihydroxyacetone phosphate can also be converted into glyceraldehyde-3-phosphate), the energy yield is two mole ATP per mole glucose. Genome analysis shows that the complete glycolytic pathway is present in *E. coli*, as it is, incidentally, in most eukaryotic organisms and cells including all human cells. In contrast, in *H. pylori*, a causative

activities may then be overlooked in homology searches; Koonin *et al.*, 1996).

Databases increasingly facilitate the prediction of metabolic pathways, notably EcoCyc (Encyclopedia of *E. coli* Genes and Metabolism), HinCyc: (Encyclopedia of *H. influenzae* Genes and Metabolism), PUMA (see below), Biocatalysis/Biodegradation Databases, Enzyme Database (e.g. www.expasy.ch/sprot/enzyme.html), Ligand Databases (e.g. at the Japanese genome net, its compound section is a collection of metabolic compounds including substrates, products, and inhibitors; www.genome.ad.jp/dbget/ligand.html), Klotho, the biochemical compounds declarative database; KEGG (Kyoto Encyclopedia of Genes and Genomes page) and pathway pages on the Web such as the Boehringer Mannheim pathways chart, NetBiochem Welcome Page or pages on particular organisms such as for soybean metabolism (cgsc.biology.yale.edu/metab. html). Experimentally verified pathway databases have been collected for regulatory circuits such as cell cycle in yeast, human and budding yeast (BRITE project, www.genomes.ad.jp/brite/Cellcyclemaps.html), as cross-references (object oriented database management system ACEDB) between protein kinases, their interactions, 3D structure and pathways by Igarashi & Kaminuma (1997) and for fly genes involved in pattern formation (Jacq *et al.*, 1997).

Several software tools reconstruct metabolic pathways, usually in association with databases (see above). Early efforts (Seressiotis & Bailey, 1988; Mavrovouniotis *et al.*, 1990) required extensive pre-analysis of the genome, and the proteins encoded therein. More recent developments include Mosaic (Multipurpose Automated Genome

similar, the sequence similarity should not be restricted only to a functional domain; otherwise no complete function transfer possible). For the latter an efficient tool is the COGs server (clusters of orthologous genes; Tatusov *et al.*, 1997; http://www.ncbi.nlm.nih.gov/COG/). Profile alignments of important enzymatic activities such as signatures for pathways are also used and are being developed in several other laboratories (e.g. Rawlings & Searls, 1997).

The prediction of interdependencies of genes and metabolism

Utilizing the tools and approaches above together with methods for comparative sequence and genome analysis, a number of specific predictions in recently sequenced prokaryotic organisms have been made that go beyond the analysis provided in the publications on sequenced genomes for prokaryotes (Selkov *et al.*, 1996, 1997; Tatusov *et al.*, 1996; Koonin *et al.*, 1996; Strauss & Falkow, 1997) and eukaryotes (Oliver, 1997; Palsson, 1997). The plasticity and the enzyme variety even of very basic pathways turns out to be surprisingly high. Figure 6 illustrates this for variants from standard glycolysis encountered after genome analysis.

Predictions for protein functions and enzyme pathways just cover the repertoire of functions present. However, metabolic control analysis also considers quantitative aspects such as flux, flow, concentrations, stochiometric and allosteric effects, compartimentalization and regulation (see e.g. Schuster, 1996; Thomas & Fell, 1996; Bish & Mavrovouniotis, 1998 and references therein). Knowledge of possible metabolite flows (i.e. differ-

Robustness, modularity and interdependence

When considering all the levels discussed there seems to be a discrepancy between the complex nature of the networks of genes and their interdependence (e.g. *cis* regulation) on the one hand and the surprising robustness (e.g. horizontal gene transfer or gene loss) on the other. One way in which such robustness might be achieved is a highly modular organisation, the interdependencies

robustness can be observed at all levels, the balance of which might seem surprising given the shuffling, horizontal transfer, disruption, insertion etc. of genetic material. On the other hand, the robustness represents also hope that functional features are more significantly implicated and predictable from sequence than previously expected.

Prediction of function from sequence is a considerably more complex enterprise than a simple sequence database search which represented the

d'Aubenton-Carafa, Y., Brody, E. & Thermes, C. (1990). Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop, structures. *J. Mol. Biol.* **216**, 835–858.

Audic, S. & Claverie, J.-M. (1998). Self-identification of protein-coding regions in microbial genomes. *Proc. Natl Acad. Sci. USA*, **95**, 10026–10031.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence databank and its supplement TrEMBL. *Nucl. Acids Res.* **26**, 38–42.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status and progress. *Nucl. Acids Res.* **25**, 217–221.

Bhatia, U., Robison, K. & Gilbert, W. (1997). Dealing with database explosion: a cautionary note. *Science*, **276**, 1724–1725.

Bish, D. R. & Mavrovouniotis, M. L. (1998). Enzymatic reaction rate limits with constraints on equilibrium constants and experimental parameters. *Biosystems*, **47**, 37–60.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J. *et al.* (1997). The complete genome sequence

in aligned sequences: application to the analysis of random sequences. *Nucl. Acids Res.* **19**, 353–358.

Chargaff, E. & Davidson, N. J. (1955). *The Nucleic Acids*, Academic Press, New York.

Chastian, P. D. & Sinden, R. R. (1998). CTG repeats associated with human genetic disease are inherently flexible. *J. Mol. Biol.* **275**, 405–411.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996). Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.

Colas, P., Cohen, B., Jessen, T., Grishina, I., McCoy, J. & Brent, R. (1997). Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. *Nature*, **380**, 548–550.

Dandekar, T. & Sharma, K. (1998). *Regulatory RNA*, Springer Verlag, Heidelberg.

Dandekar, T., Beyer, K., Bork, P., Kenealy, M.-R., Pantopoulos, K., Hentze, M. W., Sonntag-Buck, V., Flouriot, G., Gannon, F., Keller, W. & Schreiber, S. (1998a). Systematic genomic screening and analysis of mRNA in untranslated regions and mRNA precursors: combining experiment and computational approaches. *Bioinformatics*, **14**, 271–278.

Dandekar, T., Snel, B., Huynen, M. A. & Bork, P.

genitalium. *Proc. Natl Acad. Sci. USA*, **94**, 11929–11934.

Fitch, W. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Nature*, **269**, 496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., *et al.* (1998). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**, 580–586.

Freeman, J. M., Plasterer, T. N., Smith, T. F. & Mohr, S. C. (1998). Patterns of genome organization in bacteria. *Science*, **279**, 1827.

Frishman, D. & Mewes, H.-W. (1997). Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4**, 626–628.

Gaasterland, T. & Sensen, C. W. (1996). Fully automated

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucl. Acids Res.* **24**, 4420–4449.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. & Herrmann, R. (1997). Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucl. Acids Res.* **25**, 701–712.

Hood, D. W., Deadman, M. E., Jennings, M. P., Bisercic, M., Fleischmann, R. D., Venter, J. C. & Moxon, E. (1996). DNA repeats identify novel virulence genes in Haemophilus influenzae. *Proc. Natl Acad. Sci. USA*, **93**, 11121–11125.

Humphery-Smith, I. & Blackstock, W. (1997). Proteome analysis: genomics via the output rather than the input code. *J. Protein Chem.* **16**, 537–544.

Huynen, M. A. & Bork, P. (1998). Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.

Huynen, M. A., Perelson, A., Vieira, W. A. & Stadler, P. F. (1996). Base pairing probabilities in a complete HIV-1 RNA. *J. Comput. Biol.* **3**, 253–274.

Huynen, M. A., Diaz-Lazcoz, Y. & Bork, P. (1997). Differential genome display. *Trends Genet.* **13**, 389–390.

Huynen, M. A., Dandekar, T. & Bork, P. (1998a). Genomics: differential genome analysis applied to the species specific features of *Helicobacter pylori*. *FEBS Letters*, **426**, 1–5.

Huynen, M. A., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, A., Yuan, Y. & Bork, P. (1998b). Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* **280**, 323–326.

philic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*, 390, 364–370.

Koonin, E. V. & Galperin, M. Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* 7, 757–763.

Koonin, E. V., Mushegian, A. R. & Bork, P. (1996). Non-orthologous gene displacement. *Trends Genet.* 12, 334–336.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997). Comparison of archaeal and bacterial genomes: computer analysis of proteins sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* 25, 619–637.

Krishna, R. & Wold, F. (1997). Identifcation of common post-translational modifications. In *Protein structure – A Practical Approach* (Creighton, T., ed.), 2nd edit., pp. 91–116, Oxford University Press.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., *et al.* (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, 390, 249–256.

Küster, B. & Mann, M. (1998). Identifying proteins and post-translational modifications by mass spectrometry. *Curr. Opin. Struct. Biol.* 8, 393–400.

Liao, J. C., Hou, S.-Y. & Chao, Y.-P. (1996). Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.* 52, 129–140.

Link, A. J., Hays, L. G., Carmack, E. B. & Yates, J. R. (1997). Identification of the major proteome composition of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis*, 18, 1314–1334.

Lukashin, A. V. & Borodovsky, M. (1998). GeneMark. hmm: new solutions for gene finding. *Nucl. Acid Res.* 26, 1107–1115.

Lupas, A. (1997). Predicting coiled coil regions in proteins. *Curr. Opin. Struct. Biol.* 7, 388–393.

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1989). *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Mavrovouniotis, M. L., Stephanopoulus, G. & Stephanopoulus, G. (1990). Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.* 36, 1119–1132.

Medigue, C., Rouxel, T., Vigier, P., Henaut, A. & Danchin, A. (1991). Evidence for horizontal transfer in *Escherichia coli* speciation. *J. Mol. Biol.* 222, 851–856.

Melendez-Hevia, E., Waddell, T. G., Heinrich, R. & Montero, F. (1997). Theoretical approaches to evolutionary optimization of glycolysis. *Eur. J. Biochem.* 244, 527–543.

Mighell, A. J., Markham, A. F. & Robinson, P. A. (1997). Alu sequences. *FEBS Letters*, 417, 1–5.

Mrazek, J. & Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl Acad. Sci. USA*, 95, 3720–3725.

Mushegian, A. R. & Koonin, E. V. (1996a). Gene order is not conserved in bacterial evolution. *Trends Genet.* 12, 289–290.

Mushegian, A. R. & Koonin, E. V. (1996b). A minimal

signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6.

Nuño, J. C., Sánchez-Valdenebro, I., Pérez-Iratxeta, C., Meléndez-Hevia, E. & Montero, F. (1997). Network organization of cell metabolism: monosaccharide interconversion. *Biochem. J.* 324, 103–111.

Ogata, H., Goto, S., Fujibuchi, W. & Kanehisa, M. (1998). Computation with the KEGG pathway database. *Biosystems*, 47, 119–128.

Ogura, H., Agata, H., Xie, M., Odaka, T. & Furutani, H. (1997). A study of learning splice sites of DNA sequence by neural networks. *Comput. Biol. Med.* 27, 67–75.

Oliver, S. G. (1997). From gene to screen with yeast. *Curr. Opin. Genet. Dev.* 7, 405–409.

Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E., Jr & Yunus, I. (1996). The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucl. Acids Res.* 24, 26–28.

Overbeek, R., Larsen, N., Smith, W., Maltsev, N. & Selkov, E. (1997). Representation of function: the next step. *Gene*, 191, GC1–GC9.

Palsson, B. O. (1997). What lies beyond bioinformatics? *Nature Biotechnol.* 15, 3–4.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* 276, 71–84.

Pedersen, A. G. & Engelbrecht, J. (1995). Investigations of *Escherichia coli* promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. *Intelligent Systems Mol. Biol.* 3, 292–299.

Persson, B., Zigler, J. S., Jr & Jörnvall, H. (1994). A super-family of medium-chain dehydrogenases/reductases (MDR). *Eur. J. Biochem.* 226, 15–22.

Piatigorsky, J. & Wistow, G. (1991). The recruitment of crystallins: new functions precede gene duplication. *Science*, 252, 1078–1079.

Pollack, J. D., Williams, M. V. & McElhaney, R. N. (1997). The comparative metabolism of the mollicutes (*Mycoplasmas*): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit. Rev. Microbiol.* 23, 269–354.

Pujana, M. A., Volpini, V. & Estivill, X. (1998). Large CAG/CTG repeat templates produced by PCR, usefulness for the DIRECT method of cloning genes with CAG/CTG repeat expansions. *Nucl. Acids Res.* 26, 1352–1353.

Rawlings, C. J. & Searls, D. B. (1997). Computational gene discovery and human disease. *Curr. Opin. Genet. Dev.* 7, 416–423.

Richterich, P. (1998). Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res.* 8, 251–259.

Riley, M. (1998). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.* 8, 388–392.

Rodriguez, R. & Vriend, G. (1997). Professional gambling. In *Proceedings of the NATO Advanced Study Institute on Biomolecular Structure and Dynamics: Recent Experimental and Theoretical Advances*, 1, 1–10.

Rost, B. & Donoghue, S. (1997). Sisyphus and the prediction of protein structure. *Comp. Appl. Biosci.* 13, 345–356.