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

0022–2836/98/440707–19 $30.00/0

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
	→ Correlation in the patterns of occurrence of genes in the comparison of multiple genomes points to functional relations between the genes
Feature:	Complete list of detected gene sequences
Usage	→ Identifying the optimal candidate gene in the whole genome for an observed enzymatic activity

Various types of patterns and (context) information that become available with the analysis of the complete genome can be used for function prediction at ''lower levels'', e.g. in the prediction of the function of single genes.

partners, and pathway context are currently being predicted, but only qualitatively. Expression patterns, regulation, kinetic properties, localization and concentration effects and, even more so, dysfunctions, environmental influence, fitness contribution or clinical symptoms can currently hardly be predicted. There is furthermore a relatively poor knowledge of the mechanisms of posttranslational modifications (Esko & Zhang, 1996). For example, although some sequence patterns for preferred glycolysation sites are known, the prediction accuracy is still limited and the assignment does not include the kind of sugar or carbohydrate that is attached, so that most of the functional features of the respective proteins will remain hidden.

The main goal will be to bridge the gap between genotype and phenotype (Figure 2), i.e. to understand the genotype to a degree that the phenotypic features can be predicted: What are the genes responsible for a certain disease phenotype and which proteins of the respective pathway (or an alternative one) are the best targets for a drug to be developed, or which variations at the DNA level are best suited for the respective diagnostics? Which genes have to be changed to achieve a desired phenotype? To answer such questions in a more general way, one needs a detailed understanding of the function of higher order processes, including the complex interaction between the heritable part of the phenotype and the environment. This will require a whole battery of novel types of experimental data with appropriate bioinformatics support.

Nevertheless, it is important to extract as much information as possible from sequence data using the already available (and inexpensive) computational tools to guide experimental work.

Functional prediction for gene products by annotation transfer from homologous sequences

When homologues of a query are identified in a database search (Bork & Gibson, 1996), the annotated information of the homologue and the taxonomic, biochemical and/or molecular-biological context of the query protein are used to extrapolate possible structural and functional features of the query protein. This approach has proven extremely successful although, from a formal point of view the hypotheses generated must be experimentally verified (Eisenhaber *et al*., 1995). The information transfer from well-studied proteins to uncharacterized gene products has to be done carefully since (i) a similar sequence does not always imply similar protein structure (Sander & Schneider, 1991) or function (in particular in important details such as recognition loops) and (ii) the annotation of the database protein might be incomplete or even wrong.

Often (particularly in the case of automatic prediction programs), the function is transferred from another member in a multigene family, but not exactly from the functional counterpart in a different species. Even orthologues (see below) can differ functionally in various organisms. It should also be emphasized that generally only the molecular functions of a protein can be transferred by analogy (Figure 1); it is rather rare that a particular sequence motif strongly correlates with cellular functions as in the case of the DEATH-domain, which is mainly contained in apoptose signalling

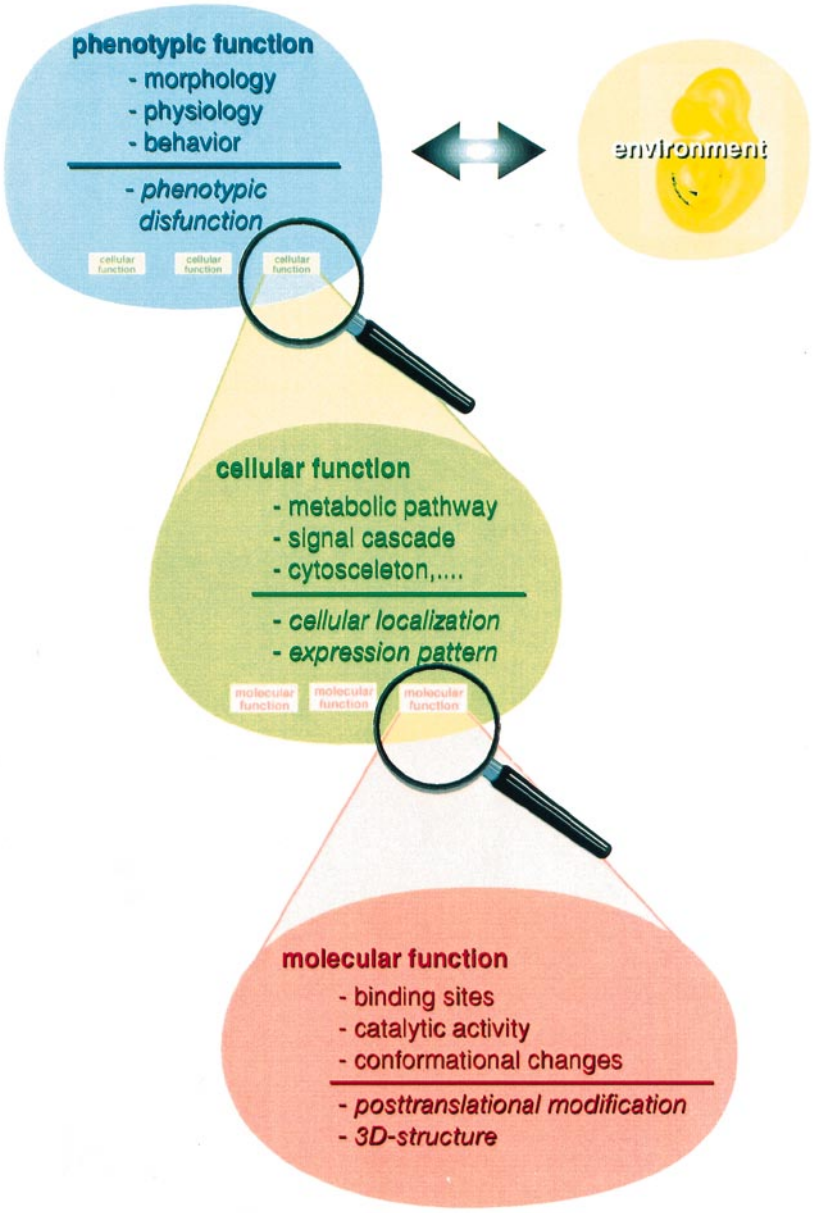

Figure 1. Characterization of protein function. Whereas nucleic acids fulfil the tasks of storage, transfer and processing of genetic information contained in the genome of living organisms, the proteins (gene products) form a complex (single- or multi-) cellular machinery for the realization of this 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 function), the morphology and physiology of the organism and its behaviour. Some phenotype properties may be traced to the activity of a single gene but most are determined by the co-operative action of many gene products. The absence of activity of a specific gene can result in phenotypic dysfunction. The knowledge of whole genomes will open a new era in the investigation of properties determined by many genes since the total set of genes influencing the phenotype is known.

proteins. Sometimes only the expression pattern and the tissue context determine the final functionality (for example, high sequence identity and even gene sharing between metabolic medium-chain dehydrogenases and eye lens crystallins; Piatigorsky & Wistow, 1991; Persson *et al.*, 1994; Serry *et al.*, 1998). Proteins (or more precisely, their domains) as structural and functional modules are multiply adapted by evolutionary processes and re-used in a different context. Thus, higher order

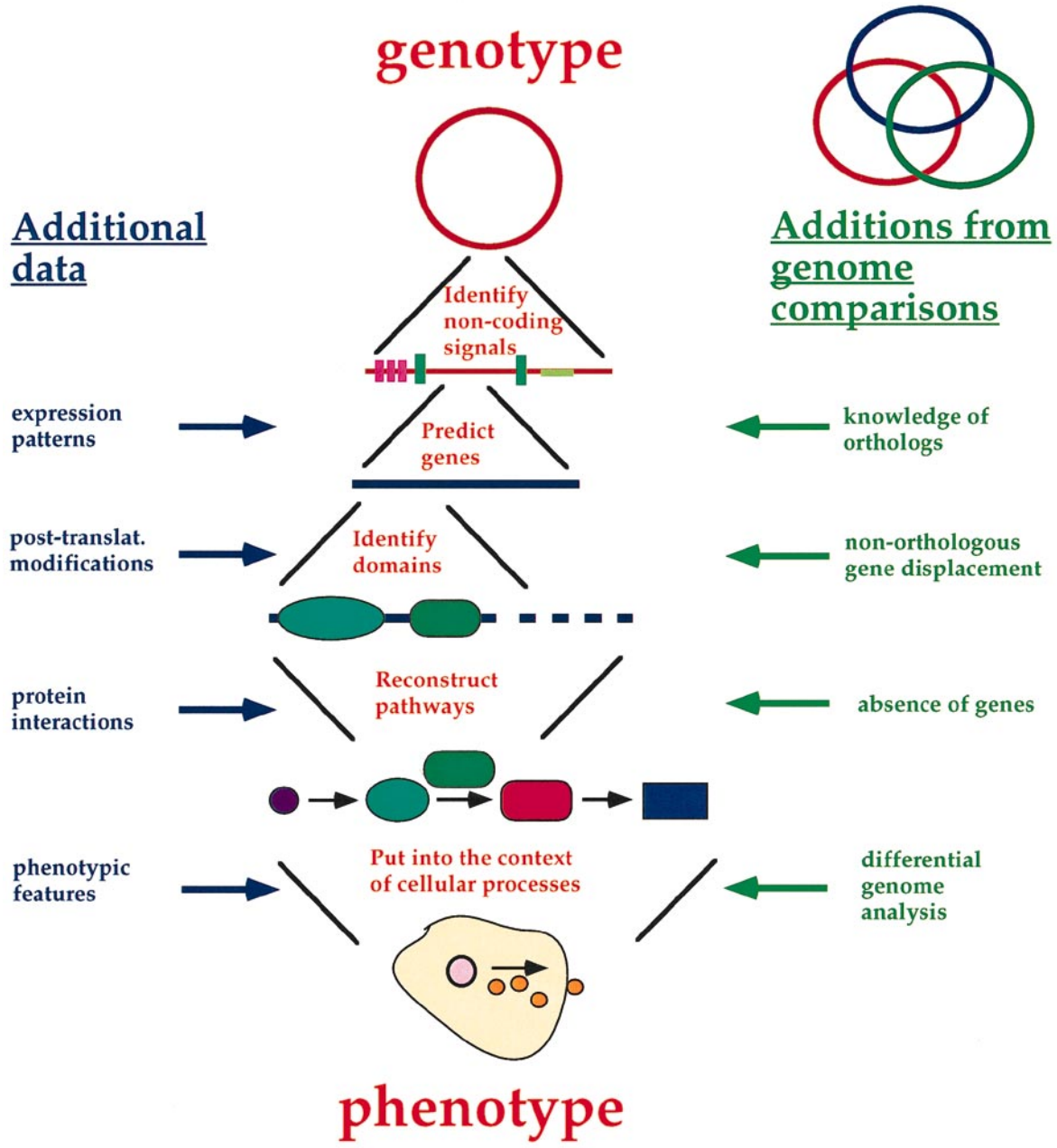

Figure 2. Function prediction scheme: zooming in and out. Whether gene, contig or genome: current methods concentrate on gene prediction and the annotation of individual genes that are then put into context. Due to our limited understanding of the genome, this is only possible by accessing complementary 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 will leave about one error per gene (mostly frameshifts) leading to considerable deviations at the protein level. Unless the accuracy is above 99.99% (the majority of the reading frames are sequenced without any error), a considerable error rate should be considered in the analysis.

Processing of raw genomic DNA includes identification of genes, their exon/intron structure, and the ''*in silico*'' translation into protein sequences by automatic methods. Given the limited accuracy of eukaryotic gene prediction methods (Burset & Guigo, 1996; Guigo, 1997, see below) and the impact of organelle- and species-specific translation tables, of pre- (RNA editing and splicing) and post- (propeptide cleavage and protein splicing, side-chain modifications) translational changes, the sequence quality of a given genomic segment is expected to be lower in protein databases than at the DNA level.

The value of sequences stored in databases is greatly increased by their functional annotation. However, automatic as well as manual annotations have all kinds of inaccuracies ranging from orthographic errors, simple spelling ambiguities, and incompleteness to semantic mistakes (Bork & Bairoch, 1996; Eisenhaber & Bork, 1998; Smith & Zhang, 1997). Function assignments obtained as a result of automatic homology searches are often not labelled as such and cannot easily be distinguished from true experimental data (Bork & Bairoch, 1996; Andrade & Sander, 1997). Furthermore, there is a gap between the current database annotation and the knowledge embodied in the scientific literature (Bork & Koonin, 1998).

Creating, updating, and correcting functional annotation is a costly effort absorbing a considerable amount of manpower. At the moment, there is no real alternative to manual input from experts. In the future, text analysis systems might support this process by automatically extracting abstracts of related articles from literature databases and selecting relevant keywords and text units for protein families (Guigo *et al.*, 1991; Guigo & Smith, 1993; Andrade & Valencia, 1997).

For analyses of genotype-phenotype relationships, the retrieval of complete sets of proteins from sequence databases with respect to their function is necessary. This can efficiently be achieved only by categorized protein function descriptions (Riley, 1998) for cellular (subcellular localization, involvement in metabolic pathways, signal transduction 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 SWISS-PROT (expasy.hcuge.ch/sprot/), and, for catalytic functions, the system of Overbeek *et al.* (1997) put enormous pressure on the database curators. Such classifications also have to be adapted and updated frequently in accordance with the increasing understanding of the biological relationships.

Rule-based automatic algorithms that parse written annotations for defined questions might be a solution since a much smaller effort (compared with database reformatting) is required for their updates. For the deduction of cellular localization, a system of about 1000 biological rules was able to classify 88% of entries of SWISS-PROT (currently seen as one of the best annotated general protein sequence databases; Bairoch & Apweiler, 1998) into subcellular localization categories. This is considerable progress given that only 22% of the entries can be retrieved using querying stems of keywords such as ''extracell'' or ''membrane'' (Eisenhaber & Bork, 1998).

Annotating genomes

Function prediction usually starts with already assembled genomic or cDNA data: at best a complete genome (Figure 2). Several features intrinsic to DNA can be recognized first, before identification of genes and pathways, although detection of the latter enhances also the annotation of non-coding features in genomes.

Nucleotide frequencies

Nucleotide frequencies are one of the oldest features of genomes that have been studied, even before sequencing was available (Chargaff & Davidson, 1955). Biases in nucleotide frequencies exist both within and between genomes, they have various uses in gene and function prediction. In warm blooded vertebrates and angiosperms, for example, the genome is divided in regions, so called isochores, that differ in G+C content. Isochores with a high G+C content are relatively rich in genes (Saccone *et al.*, 1996). Biases in G+C content can hence be used to find genes. A number of bacterial species show biases in the nucleotide frequencies of the leading and lagging strands in replication (Mrazek & Karlin, 1998; Freeman *et al.*, 1998); these biases can be correlated with a bias in

the coding density, e.g. in *Bacillus subtilis* (Kunst *et al*., 1997).

In the study of complete genomes, biases in nucleotide frequencies and codon usage provide an important clue for detecting recent horizontal transfers of genes into the genome (Figure 3; Medigue *et al*., 1991). Variations in the codon usage can be described with a principle component analysis which divides the variation among orthogonal axes. Different axes correspond to independent sources of variation; the variation in codon usage that results specifically from horizontal gene transfer can be identified using auxiliary functional information from the genes (see below). In *Helicobacter pylori* for example, it is the first principle component that reflects horizontal gene transfer (Figure 3). On the basis of the variation in the codon usage in *E. coli* its genome has been predicted to consist of at least 10–15% of recently horizontally transferred genes (Medigue *et al*., 1991). Biases of nucleotide frequencies within a genome also reveal information about their function apart from information about the evolutionary history of genes. Recently horizontally transferred genes are expected not to be involved in the core functions of the cell and to be relatively expendable (they were generally not present before the transfer). In *H. pylori* the regions with deviating nucleotide frequencies can be related to pathogenicity, or are prophages and/or are rich in insertion sequences (Figure 3). The same observation has been made in *Haemophilus influenzae* (Fleischmann *et al*., 1995; Huynen *et al*., 1997).

Repeats

For a large fraction of the DNA of multicellular eukaryotes no obvious function has yet been assigned. Most of it consists of repetitive elements. For example, *Alu* repeats may cover as much as 13% of the human genome (Mighell *et al*., 1997). Repetitive, non-coding DNA should be filtered out as one of the first steps in function prediction to reduce the search space for the finding of genes in eukaryotic DNA (Jurka *et al*., 1996). Coding regions contain repeats too, but these are hardly identifiable at the DNA level due to their divergence. They usually represent structural domains and should be detected at the protein level (see below). An exception are the trinucleotide repeats that are expanded in a number of disease genes (Chastian & Sinden, 1998); they can even specifically be used to search for such genes in DNA libraries (Pujana *et al*., 1998).

In prokaryotes repeats are much less frequent. However, tetranucleotide repeats have been found in some virulence genes that increase variability by frameshift mutations (Hood *et al*., 1996). More strikingly, repetitive elements even have been found in what are probably the smallest bacterial genomes, those of mycoplasmas. These have been hotspots for genome rearrangements *via* recombination, as can be deduced by whole genome comparison (Himmelreich *et al*., 1997).

Regulatory regions

Regulatory regions can indicate when and how genes are expressed, repressed or co-expressed. Their computational detection is a powerful complement to novel experimental approaches (see proteomics, below). If known structures provide a template, simple consensus searches, matrix approaches and also programs taking into account specific features, structural constraints and energy values are available (reviewed by Dandekar & Scharma, 1998).

If no genomic template structures are available, neural networks (Demeler & Zhou, 1991; Pedersen & Engelbrecht, 1995; Ogura *et al*., 1997), language based approaches (Trifonov, 1996) and other non-consensus search methods are important (e.g. Tiwari *et al*., 1997). One can, for example, search for so-called CpG islands, which are, relative to the rest of the genome, abundant in the regulatory regions of mammalian housekeeping genes (Wirkner *et al*., 1998). The combination of artificial *in vitro* evolution and genomic screening is another powerful way to identify a regulatory motif when no template structure or sequence is available. The computer based genomic screen delineates how close the *in vitro* selection procedure comes to the situation *in vivo* (Dandekar *et al*., 1998a).

The challenge from complete genome sequences is double: first, a comprehensive annotation of known regulatory elements using specific searching methods (i.e. either templates for particular elements such as promotors, attenuators, terminators and enhancers or RNA secondary structure fitting methods; d'Aubenton-Carafa *et al*., 1990; Brendel *et al*., 1986); second, the identification of novel elements using comparative analysis and experimental indications (co-expression, etc.). Knowledge of gene expression and changes in gene expression patterns at a complete genomic level may revolutionize drug discovery processes. An overview of the complete genome allows much better tailoring of drugs and the discovery of correct, condition specific targets (Gelbert & Gregg, 1997).

A comparison of complete genomes identifies orthologous genes (see below). Their upstream regions can be screened for common regulatory signals in a much reduced search space. When co-expression patterns or functional interactions of genes are known, one can also search within the non-coding regions of a single genome. Unfortunately, regulatory regions in prokaryotes seem to be little conserved (Figure 4; Diaz-Lazcoz *et al*., unpublished), thus it is necessary to include several species to increase the signal to noise ratio *via* multiple alignments. In the case of putative RNA structures, one can utilize methods that include base-pairing information (cf. Chan *et al*., 1990; Han & Kim, 1993). These approaches, however, require

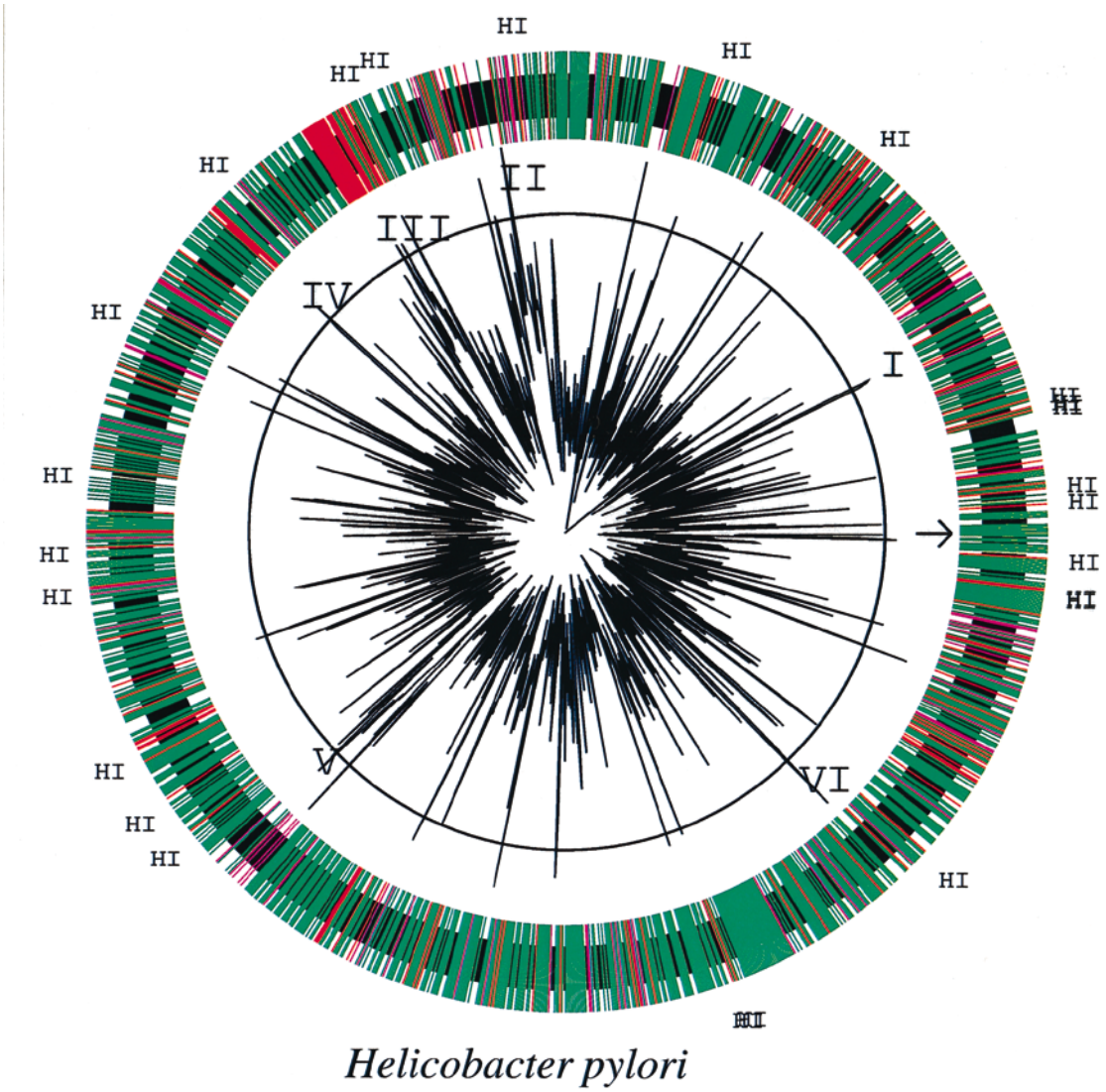

Figure 3. Differential genome display of *H. pylori versus E. coli, H. influenzae*. The genes of *H. pylori* are divided into sets. Set 1 (green) are genes with a homologue in *E. coli*, set 2 (HI), genes with a homologue in *H. influenzae* but not in *E. coli*. Set 3 (red) are genes without a homologue in *E. coli* that are (putatively) involved in interaction with the host like virulence factors, outer membrane proteins and toxins. Set 4 (purple) are genes without a homologue in *E. coli* or *H. influenzae* that are not host interaction factors. A large fraction (63%) of the genes in *H. pylori* that have no homologue in *E. coli*, but for which some functional classification is possible, can be considered host interaction factors. The star-figure in the centre gives the values of the codon usage of the genes on the first principle component in distance to the centre. The first principle component corresponds roughly to the usage of A and to a lesser extent T in 3D codon positions. Hence, genes with a high value on this axis have a relatively high A+T content in third coding positions. Six clusters (I-VI) of at least three consecutive genes with, on average, the codon usage that deviates the most from the genomic mean were further analysed. The genes in these tend not to have any homologue in *E. coli* or *H. influenzae*, their closest relatives for which complete genome sequences are available. This observation supports that the genes in clusters I-VI result from horizontal gene transfer into the genome. Proteins from I and VI are hypothetical proteins with no known homologues other than proteins in *H. pylori* itself. Region II contains homologues of VirB4, a virulence factor and of transposases. Region III is the CAG pathogenicity island, whereas region V again is rich in transposases. Region IV consists of three proteins, HP0611-HP0613. Sequence analysis reveals a frameshift that would merge HP0611 with HP0612. The resulting protein is an ABC type 2 transporter, the only one that can be observed in *H. pylori*. ABC-2 transporters are involved in export of complex carbohydrates and play an important role in virulence.

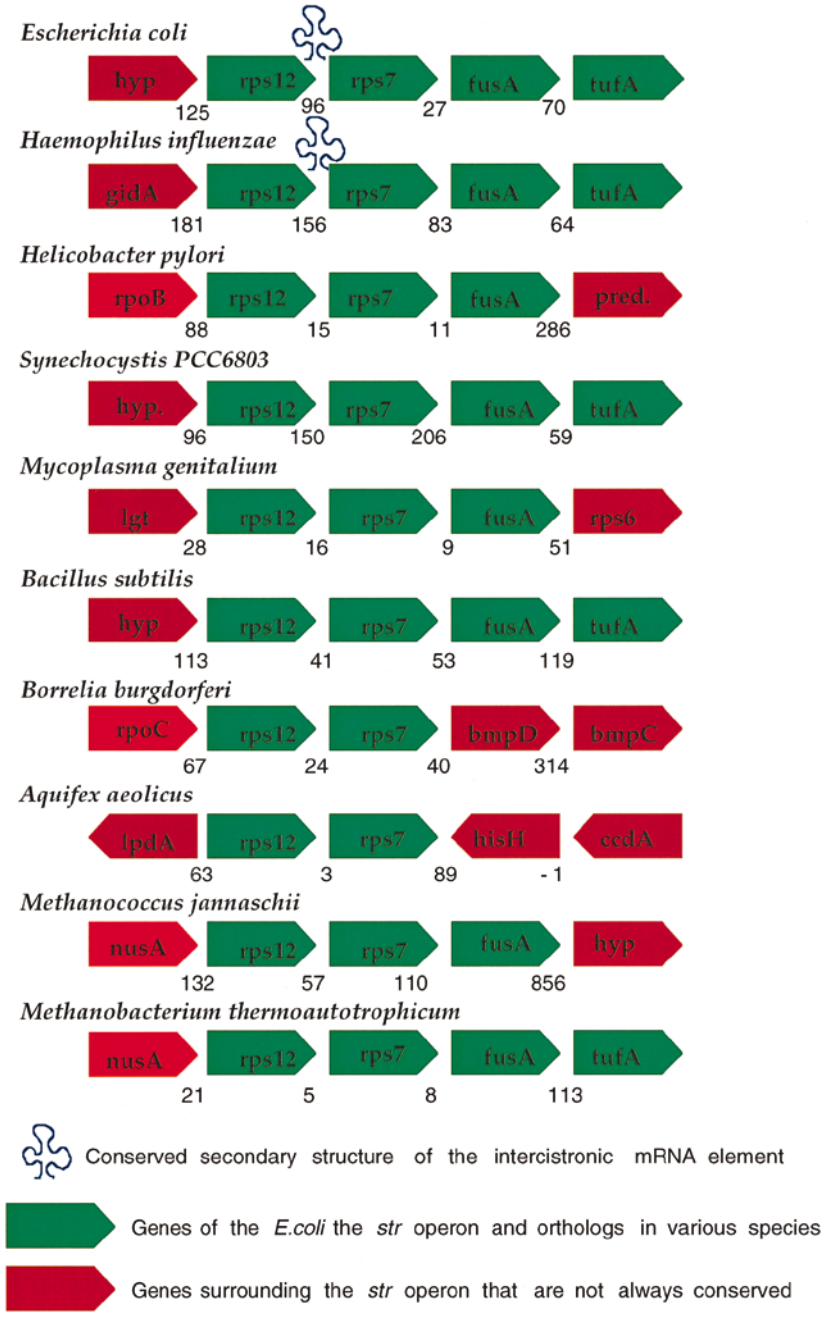

Figure 4. Variability of regulatory regions. Structure of the *str* operon and surrounding genes in eight Bacteria and two Archaea. The organisation of the *str* operon in *Archaeoglobus fulgidus* is essentially the same as in *M. jannaschii*, while the structure of the *str* operon in *Mycoplasma pneumoniae* 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 between two or more genomes other than *E. coli*. Gene names may vary in the official genome annotations, but were kept constant in this Figure for clarity purposes. Operon structure is generally not well conserved in prokaryotic evolution. The conservation of two genes besides each other in all the prokaryotic genomes that have been sequenced thus far can be regarded as an exception. Note that the conservation of the *str* operon does not follow the standard phylogenetic pattern: i.e. the operon structure in *E. coli* is more similar to that in the Archaea than it is to the operon structure in the closer related bacteria *B. burgdorferi* and *A. aeolicus*. In *E. coli* expression of the *str* operon is regulated by an RNA secondary structure located between the *rpsL* and *rpsG* genes (Saito & Nomura, 1994). A similar structure is present in *H. influenzae*, but is absent from the other species. Hence, regulatory elements appear even less conserved than gene order.

longer and relatively strong signals. Weaker motifs can be identified using statistical approaches without prior alignment of sequences (cf. Staden, 1989; Hertz *et al.*, 1990; Wolfertstetter *et al.*, 1996). Reliable statistics require, however, many orthologous sequences. The comparison based on orthologous regions in complete genomes, from different Gram-negative bacteria for instance, offers a new way to identify regulatory motifs without a preconception of the regulatory motifs revealed. In due course this and other approaches (see above) will improve quantitative predictions on expression and regulation in complete genomes and should also yield probabilities for tissue distribution of expression patterns and regulatory factors.

Gene prediction

The prediction of protein coding genes from DNA sequences can become a major bottleneck in genomics as currently there is quite a lot of information loss when genes cannot be identified correctly. In eukaryotes the situation is particularly complicated, due to the generally low coding density (probably as low as 2% in human) and the presence of introns surrounding the relatively short coding regions. Various different, but weak signals have to be combined such as promotors, splice sites, translational start and stop sites; different knowledge-based methods complemented by homology searches are applied to utilize them (Guigo,

1997). However, an analysis of the accuracy of all available packages for the prediction of coding sequences for a region of human DNA showed a low accuracy for the prediction of coding sequences and specifically the prediction of intron/exon boundaries (Burset & Guigo, 1996; Gelfand *et al*., 1996; Guigo, 1997; Lukashin & Borodovsky, 1998).

In Archaea and Bacteria the situation is, relative to most eukaryotes, less complicated due to the almost complete absence of introns. In predicting protein coding regions, several methods make use of the information in the complete genome (Borodovsky *et al*., 1994; Fraser *et al*., 1998). Such bootstrapping methods use the genes that can easily be predicted, e.g. on the basis of the length of open reading frames and/or similarities to genes from other species to establish (i) taxon-specific patterns in codon usage, hexanucleotide frequencies and local complexity (information content), and (ii) taxon-specific signal sequences like poly(A) signals, regulatory sequences such as ribosome-binding segments (Shine-Delgarno segments) and promoters and start codons, etc. These patterns are currently implemented into Hidden Markov Models (HMMs) to predict the other genes in a genome. A method that relies on no *a priori* information to divide the genome into coding and non-coding regions has been shown to be successful (Audic & Claverie, 1998). These gene finding approaches should be complemented by another round of homology searches to find shorter or frameshifted genes that do not follow the codon usage of the organism.

Annotating individual proteins

Although homology searches are often already integrated into the gene prediction procedures, they are fully exploited only at the protein level with its higher sensitivity. Database searches are a standard technique for annotating proteins, but should be used in context with other methods (Bork & Koonin, 1998).

Domain analysis

Due to the modularity of many proteins, i.e. their multidomain architecture, the first step in functional annotation should be a scan for known domains in a query protein. Several databases exist that comprise patterns or profiles, i.e. fingerprints of already classified domains, and are well-suited for this first scan. Although somewhat redundant, they each have their individual strengths. PROSITE (Bairoch *et al*., 1997) is one of the oldest and probably most widely used. It is well-annotated and covers more than 1000 different domains. With the inclusion of PROFILESCAN there is now also access to more than 250 domains that cannot easily be described with the classical PROSITE consensus string. A drawback, perhaps, is that the profiles are not yet fully integrated and most of them are not exhaustively annotated (which is a huge amount of work). BLOCKS (Henikoff *et al*., 1998) is derived from PROSITE and offers ungapped alignments that are, in turn, used for a pattern matching approach which is more sensitive than the consensus string matching method of the original PROSITE database.

PRINTS describes a protein domain with a set of several motifs separated along the sequence (Attwood *et al*., 1998). Version 17.0 is extensively annotated and comprises about 800 fingerprints with in total about 4500 motifs.

PFAM contains a collection of accessible multiple alignments that are translated into hidden Markov models; version 2.1 is a large collection covering 527 families that match at least once 47% of all SWISS-PROT entries in release 34 (Sonnhammer *et al*., 1998); it more sensitive than the classical PROSITE or PRINTS, has poorer annotation, but has many entries crosslinked to other domain databases. SMART concentrates only on mobile domains and hence is not exhaustive, but has a high sensitivity and selectivity, takes care of domain borders and provides additional annotation features (Schultz *et al*., 1998). Each of the databases offers search software on the web and there are efforts under way to overcome the difficulties of different formats and annotation styles.

Intrinsic feature analysis

Current database search techniques are all hampered by compositionally biased (low complexity) regions with a reduced residue alphabet. This includes (1) transmembrane regions: accumulations of ten hydrophobic residues in segments of length 20 of non-homologous transmembrane proteins are treated as homologous, (2) coiled coil segments (widespread heptarepeats with patterns of hydrophobic and polar residues) that pollute database search outputs with high scoring similarities to analogous (but probably not homologous) coiled coil regions in other proteins, (3) small repeats that lead to a bias in amino acid composition and (4) other regions with biases towards one or several amino acids such as proline-rich or glutamine-rich regions.

Methods exist for the identification of all those features. Most of these use many sequences with the feature as training sets and identify the feature knowledge based. A general method for finding low complexity regions, SEG (Wootton & Federhen, 1996) is already integrated into BLAST (Altschul *et al*., 1997) in the form of a filter. Special types of composition bias can, of course, be predicted better by specialised methods such as coiled coil predictors (for a review see Lupas, 1997), transmembrane helix recognition (e.g. TOPPRED2, von Heijne, 1992) or even for subclasses of those such as signal sequences (e.g. SIGNALP, Nielsen *et al*., 1997). For transmembrane regions, a variety of methods exists with widely varying outputs. It is worrying that when using different methods for

genome analysis the results vary greatly, e.g. the fraction of transmembrane proteins in *Mycoplasma genitalium* was predicted to be 18% (Fischer & Eisenberg, 1997), 24% (Koonin *et al*., 1997), 30% (Arkin *et al*., 1997) and 36% (Frishman & Mewes, 1997).

To avoid spurious hits and thus erroneous transfer of functional information, such regions can be filtered out, for example using the SEQ option in BLAST (Altschul *et al*., 1997; replaced by ''neutral'' Xes for ''any amino acid''). One has to bear in mind though that also such residues can contain useful functional and structural information and need to be annotated.

Another functional feature, especially in eukaryotic proteins, is the presence of posttranslational modifications. The EXPASY server provides useful software tools to detect and describe these (www.expasy.ch/www/tools.html).

Homology analysis

A classical database analysis should only be performed after identification and masking of domains and intrinsic features described above. This has the advantage of search space reduction and of better annotation quality. A database search using BLAST (Altschul *et al*., 1997) or FASTA (Pearson, 1998) often reveals significant homologues, but this is then only the beginning of a complicated, and mostly manual transfer of functional information from the homologue in the database to the query sequence as one does not know how many of the functional features are shared (Doerks *et al*., 1998). Averaged over all species, the chance that a newly sequenced gene has a homologue in sequence databases detectable by BLAST is already above 70% (e.g. 84% for yeast chromosome III; Bork & Koonin, 1998; 70–85% for Bacteria and 73% for Archaea; Koonin *et al*., 1997, but lower for animals), while the fraction for which some functional features can be predicted is at least 70% in Archaea and Bacteria (Koonin *et al*., 1997). For more than a third of all bacterial proteins, some homology-based fold assignments can be done with high confidence (Huynen *et al*., 1998b; M.A.H. *et al*., unpublished). Knowing the 3D structure of a protein is crucial in the understanding of the relation between sequence and function. In the case that amino acid identity levels to sequences with known 3D structures are higher than 50%, homology modelling can be used to further elucidate the roles and interactions of individual amino acids (Johnsson *et al*., 1994; Eisenhaber *et al*., 1995; Sanchez & Sali, 1997; Rodriguez & Vriend, 1997). Other predicted structural features such as secondary structure elements (Rost & O'Donoghue, 1997) can also be used in functional characterization. Characterization of a potential protein or RNA secondary structure can help to assess whether an open reading frame codes for a protein or a sequence codes for a functional RNA structure (Huynen *et al*., 1996), respectively, or to test hypotheses based on other, independent observations.

Only in the minority of cases can functional and structural features of a homologue be transferred to the query sequence as is (see above, Figures 1, 2) because often only some of the features are shared. Functional equivalence is only likely for orthologues.

Finding orthologues

Orthologues (Figure 2, top right) are genes whose independent evolution reflects a speciation event rather than a gene duplication event (Fitch, 1970). They are likely to perform the same function in various species, and hence represent a refinement over homologues in sequence analysis and annotation. Knowledge of the complete genome and of its protein coding regions improves the detection of orthologues. Orthologues are expected to have the highest level of pairwise similarity between all the genes in two genomes (Tatusov *et al*., 1996, 1997; Huynen & Bork, 1998), having diverged relatively recently compared to non-orthologous homologues. One needs to know all the proteins in two genomes to use relative levels of sequence identity to identify orthologues. Methods for the finding of orthologues rely both on relative similarity of genes from various genomes, and on information from the context of a gene in a genome. If two genes from different genomes share the same context, e.g. in the form of being a neighbour to a gene that also has the highest pairwise similarity between the two genomes, this supports them being orthologues of each other. The comparison of the sequence tree and the species tree can help in identifying orthologues (Yuan *et al*., 1998), assuming that the genes have not been subject to horizontal transfer. Apart from information about the ''functions'' present in the genome, orthologues also provide information about the evolution of gene regulation. Specifically by comparing the 5′ and 3′ regions of orthologous genes one can obtain information about the evolution of promotors and operator/repressor sequences, and about the evolution of RNA secondary structures involved in gene regulation (see above). Orthologues should be the basis of subsequent reconstruction of pathways, rather than proteins for which we only know that they are homologous. Within the current databases, only a minor fraction of homologous relations can be classified as orthologous and thus one has to incorporate external data (Figure 2, left) for further function characterization.

Searching genes for a function

A tool that further exploits the information from comparing genomes for function prediction is differential genome analysis (Huynen *et al*., 1997, 1998a). The genes that are not shared between two genomes are probably responsible for species-

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.

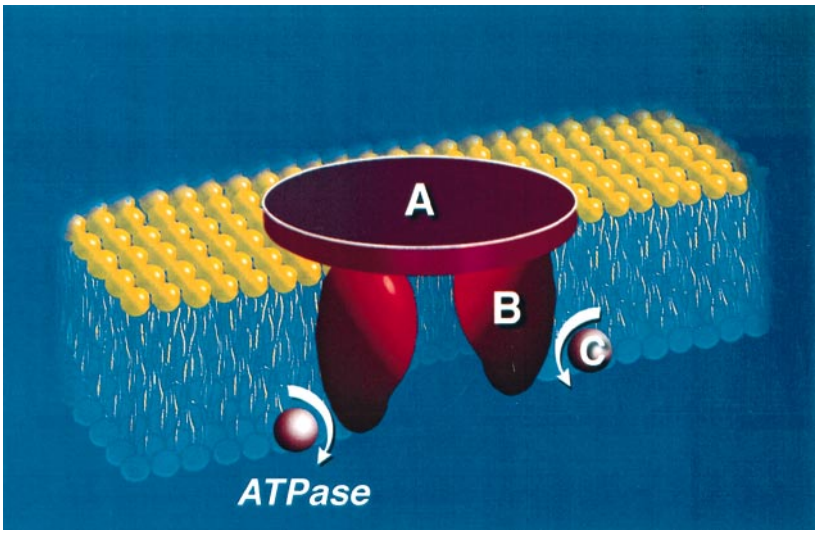

Figure 5. Protein interactions. Shown are ABC transporter proteins in the membrane of Gram-negative bacteria. Encoding genes are found as conserved gene clusters in the same sequential order in the complete genome sequences of *E. coli*, *H. influenzae* and *H. pylori*. They are an example of protein 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

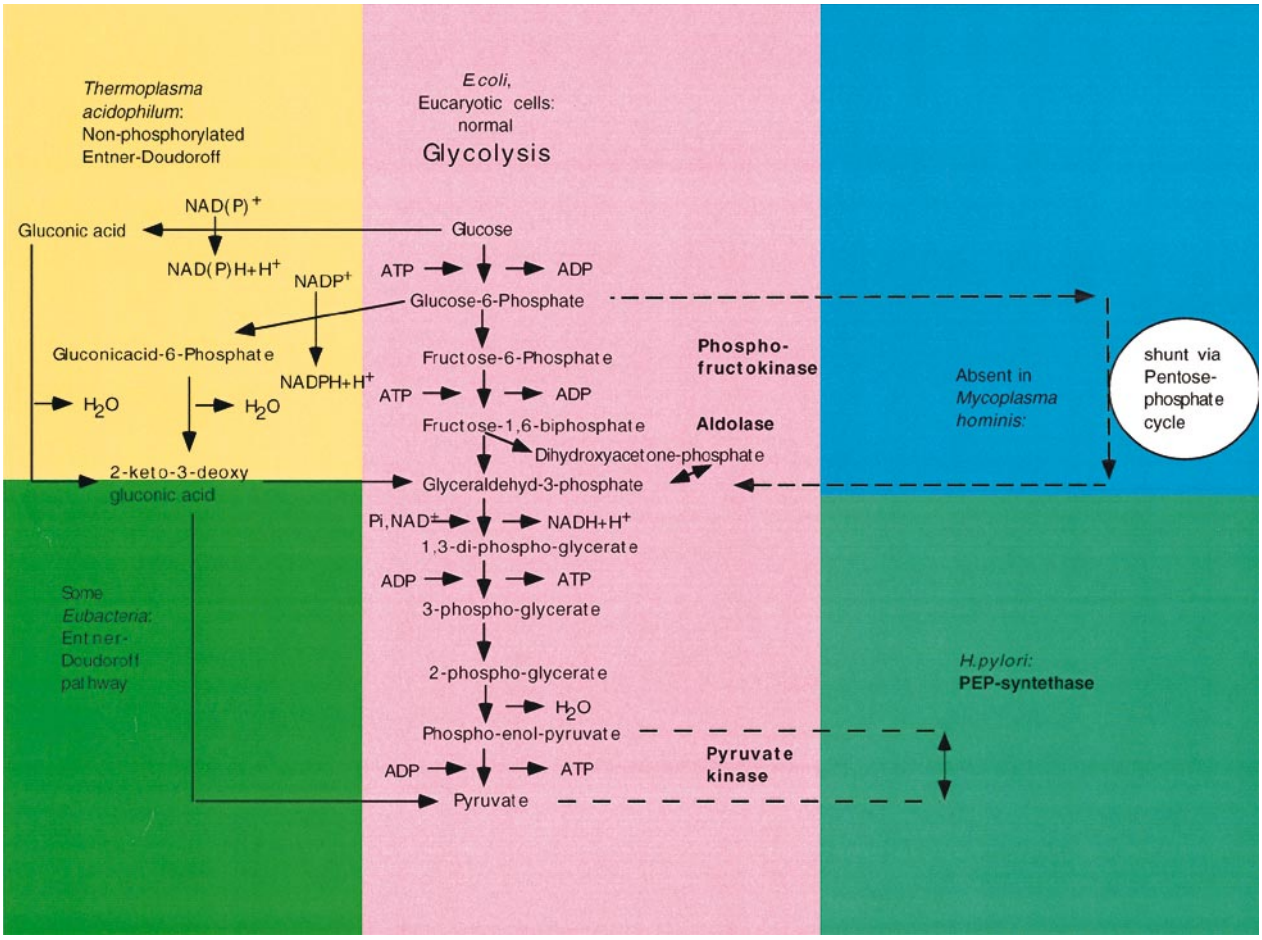

Figure 6. Prediction of metabolic pathways and pathway alignment. The glycolytic pathway (centre) and alternative other routes (sides) predicted from the genome and observed in several microorganisms are shown and compared (pathway alignment) to illustrate the often underestimated variability of metabolic 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 agent for stomach ulcer and chronic ulcerative gastritis, phosphofructokinase in the upper part of the glycolytic pathway and the important enzyme pyruvate kinase in the lower part seem to be missing. Thus a different route has to be taken in *H. pylori*. According to our analysis (right bottom), a homologue of phosphoenol-pyruvate (PEP) synthetase is present, which may support the missing step of the pyruvate kinase albeit at a reduced energy yield. The taking over of the role of pyruvate kinase by phophoenol-pyruvate (PEP) synthetase could be an adaptation to the highly acidic environment of the stomach in which *H. pylori* has to survive. More, and also more complex phenotypic features of *H. pylori* can be understood in this way by a pathway analysis utilizing differential genome comparisons (Huynen *et al*., 1998a). What about alternatives for the first part of glycolysis? This is illustrated in Figure 6 for *Mycoplasma* species which are non-glycolytics: Phosphofructokinase is missing (as probably is the case in *H. pylori* where only some homologue to pfkB from *E. coli* is present but is likely to be utilized differently) and also aldolase is absent, for instance in *Mycoplasma hominis*. These species seem to channel instead glucose by the pentosephosphate cycle (Pollack *et al*., 1997), which also yields glyceraldehyde-3-phosphate plus ribose and NADPH for nucleotide synthesis and is thus less dispensable than parts of glycolysis in these very compact genomes (Himmelreich *et al*., 1997). Our own investigations indicate that this should only be stochiometrically possible if there are additional enzymes in the genome or additional functions of known enzymes which serve to replenish the pool of sugar phosphates in the pentose phosphate pathway. There are several further alternatives for converting glucose to pyruvate. The Entner-Doudoroff pathway (bottom left), is used instead of glycolysis in some bacteria (Danson & Hough, 1992). Furthermore, genome analysis by us and others shows that this route is present as a backup pathway for instance in all Gram-negative genomes analysed to date. The ATP yield is only one mole per mole glucose. Probably it survived as an exclusive pathway in some genomes due to its simplicity and direct yield of NADPH. Top left shows the non-phosphorylated Entner-Doudoroff-pathway. This is an example of paleo-metabolism and due to the direct conversion of glucose to gluconic acid not yet optimized to obtain any net ATP yield per mole glucose (Melendez-Hevia *et al*., 1997). It is present in some Archaea such as *Thermoplasma acidophilum* (Fields, 1987).

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 Magpie (Multipurpose Automated Genome Project Investigation Environment), an automated genome analysis tool (Gaasterland & Sensen, 1996) that accesses several databases through an object and attribute viewer. Reaction equations, and compounds are taken from the Enzyme and Metabolic Pathway Database (Selkov *et al*., 1996) and have been assigned *via* homology to proteins from several organisms. The precomputed reconstruction can be accessed *via* the Web. The WIT (What Is There) system (Overbeek *et al*., 1997) is similar in concept, but offers a wide range of query options. It is a useful toolkit (http://www.cme.msu.edu/WIT) to briefly check for pathways that might be present in the genome of interest. Also with the KEGG database pathway computations are possible, for instance testing the completeness of an enzyme list (e.g. from a genome sequencing project) with regard to a certain pathway (Ogata *et al*., 1998).

Nevertheless, current reconstruction of metabolic pathways from sequence is mostly done manually using various tools that guide the decisions with consideration of accumulated biochemical and biological knowledge. For example the EcoCyc WWW Server (Karp *et al*., 1998) is used as a reference and each possible hit there is carefully checked for orthology (the whole protein function should be 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. different paths and orders of reactions given a constant number of enzymes; ''elementary modes'', Schuster & Hilgetag, 1994, 1995; Liao *et al*., 1996; Nuño *et al*., 1997; Bonarius *et al*., 1997) should improve the understanding of the context of identified enzymes in the near future. This requires well-studied systems. However, exactly these can be achieved by extensive genome and proteome analysis.

Comparative analysis of complete genomes provides further tools to study gene interdependence. For example, genes that depend on each other are expected to occur together in genomes or to be absent altogether. By doing large scale comparative genome analysis such correlations between genes become apparent and provide an extra tool for finding connections in metabolism or signalling cascades. An example are sets of genes shared by *M. genitalium* and one of either *M. jannaschii* (set 1) or *M. thermoautotrophicum* (set 2), but not by the other. Set 1 encodes among others, the functionally related proteins phosphoglucose isomerase, glyceraldehyde 3-phosphate dehydrogenase and pyruvate kinase, that are all involved in glycolysis, whereas set 2 contains the genes for DnaK and DnaJ, parts of a chaperone pathway (Huynen & Bork, 1998).

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. *via* 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 of genes would then be limited to small sets. As yet we do not have a quantitative understanding of the modularity of cellular organisation, including the genome, and its implications for the flexibility and robustness of evolution. One also needs to keep in mind that the examples of robustness we see are a selected set: evolution does not report negative results. We have tried to show here the powers of using information contained in entire genomes, i.e. context information and the interdependencies of genes within a genome. These rules affect the function prediction process in various ways.

Limited prediction accuracy at all levels and interdependence

Although many methods exist for various aspects of each prediction step, one has to bear in mind that they are not perfect and have only a limited accuracy. In addition, most of the methods have (sometimes hidden) parameters that influence the search result drastically (just switch in BLAST the matrix from default BLOSUM62 to PAM250 and watch the changes in the output). Fortunately, the loss of information in each step is compensated by the fact that data are produced by experimental methods in all the different levels (Figure 2). Thus, the errors do not add up and can be compensated by information from different levels e.g. by using genome information to improve the prediction of protein function as described here. Experimental validation of hypotheses can also be conducted at all levels, the interdependence allows even the interpretation of cellular data for molecular features and *vice versa*.

Modularity at each level and robustness

Modularity already is present at the DNA sequence level in repeats, ubiquitous promotors, duplicated segments, etc. The limited set of domains, used again and again as structural and functional scaffold, documents modularity at the gene and protein level. Displacement of non-homologous but functionally equivalent enzymes and the distinct pathway variants that all lead to the same compounds (see Figure 6) are evidence for modularity at the cellular level. Complex systems such as the cytoskeleton (animals *versus Mycoplasmas*) or even specialized organs (vertebrate *versus* octopus eye) do not represent unique solutions and reveal that even tissues can be re-invented on the basis of lower level modules. Thus, a remarkable 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 entire repertoire of tools a few years ago. In particular, with the arrival of multiple entirely sequenced genomes and experimental input at various complexity levels we have the chance to approach a new quality of understanding of cellular processes and their evolution.

Acknowledgements

The order of the authors is alphabetically. This work was supported by Deutsche Forschungsgemeinschaft (Bo 1099/3-1) and BMBF (grants 01KW9602/6; 0311748; 0311617). We thank Enrique Morrett and Shamil Sunyaev for critical reading of the manuscript and David Thomas for stylistic corrections. Most of all we acknowledge the work and efforts of all our colleagues who could not be mentioned in this review due to limitations of space and time and the limited selection such a review necessarily has to make.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped Blast and PSI-Blast, a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Anderson, L. & Seilhamer, J. (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis,* **18**, 533–537.

Andrade, M. A. & Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotechnol.* **8**, 675–683.

Andrade, M. & Valencia, A. (1997). Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. *Intelligent Systems Mol. Biol.* **5**, 25–32.

Andrade, M., Casari, G., de Daruvar, A., Sander, C., Schneider, R., Tamames, J., Valencia, A. & Ouzounis, C. (1997). Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function. *Comput. Appl. Biosci.* **13**, 481–483.

Annan, R. S. & Carr, S. A. (1997). The essential role of mass spectroscopy in characterization of protein structure: mapping post-translational modifications. *J. Protein Chem.* **16**, 391–402.

Arkin, I. T., Brunger, A. T. & Engelman, D. M. (1997). Are there dominant membrane protein families with a given number of helices?. *Proteins: Funct. Genet.* **28**, 465–466.

Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. & Selley, J. N. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucl. Acids Res.* **26**, 304–308.

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 of *Escherichia coli* K-12. *Science,* **277**, 1453–1462.

Boguski, M. S., Tolstoshev, C. M. & Bassett, D. E., Jr (1994). Gene discovery in dbEST. *Science,* **265**, 1993–1994.

Bonarius, H. P. J., Schmid, G. & Tramper, J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnol.* **15**, 308–314.

Bork, P. & Bairoch, A. (1996). Go hunting in sequence databases but watch out for the traps. *Trends Genet.* **12**, 425–427.

Bork, P. & Gibson, T. J. (1996). Applying motif and profile searches. *Methods Enzymol.* **266**, 162–184.

Bork, P. & Koonin, E. V. (1998). Predicting function from protein sequence: Where are the bottlenecks? *Nature Genet.* **13**, 313–318.

Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393–403.

Borodovsky, M., Koonin, E. V. & Rudd, K. E. (1994). New genes in old sequence: a strategy for finding genes in the bacterial genome. *Trends Biochem. Sci.* **19**, 309–313.

Brown, P. O. (1994). Genome scanning methods. *Curr. Opin. Genet. Dev.* **4**, 366–373.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., *et al.* (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science,* **273**, 1058–1073.

Brendel, V., Hamm, G. H. & Trifonov, E. N. (1986). Terminators of transcription with RNA polymerase from *Escherichia coli*: what they look like and how to find them. *J. Biomol. Struct. Dynan.* **3**, 705–723.

Brent, R. & Finley, R. L., Jr (1997). Understanding gene and allele function with two-hybrid methods. *Annu. Rev. Genet.* **31**, 663–704.

Burset, M. & Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics,* **34**, 353–367.

Chan, L., Zuker, M. & Jacobson, A. B. (1990). A computer method for finding common base paired helices 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. (1998b). Conservation of gene order: a fingerprint of physically interacting proteins. *Trends Biochnol. Sci.* **23**, 324–328.

Danson, M. J. & Hough, D. W. (1992). The enzymology of archaebacterial pathways of central metabolism. *Biochem. Soc. Symp.* **58**, 7–21.

Demeler, B. & Zhou, G. (1991). Neural network optimization for *E. coli* promoter prediction. *Nucl. Acids Res.* **19**, 1593–1599.

DeRisi, J. L., Vishwanath, R. I. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science,* **278**, 680–686.

Doerks, T., Bairoch, A. & Bork, P. (1998). Protein annotation: detective work for function prediction. *Trends Genet.* **14**, 248–250.

Dongre, A. R., Eng, J. K. & Yates, J. R. (1997). Emerging tandem-mass-sepctroscopy techniques for the rapid identification of proteins. *Trends Biotechnol.* **15**, 418–425.

Eisenhaber, F. & Bork, P. (1998). Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol.* **8**, 169–170.

Eisenhaber, F., Persson, B. & Argos, P. (1995). Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *CRC Crit. Rev. Biochem. Mol. Biol.* **30**, 1–94.

Esko, J. D. & Zhang, I. (1996). Influence of core protein sequence on glycosamino-glycan assembly. *Curr. Opin. Struc. Biol.* **6**, 663–670.

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

Ewing, B., Hiller, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequncer traces using phred. I. Accuracy assesment. *Genome Res.* **8**, 175–185.

Eym, Y., Park, Y. & Park, C. (1996). Genetically probing the regions of ribose-binding protein involved in permease interaction. *Mol. Microbiol.* **21**, 695–702.

Fields, J. H. A. (1987). Fermentative adaptations to the lack of oxygen. *Can. J. Zool.* **66**, 1036–1040.

Fischer, D. & Eisenberg, D. (1997). Assigning folds to the proteins encoded by the genome of *Mycoplasma*

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 genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie,* **78**, 302–310.

Gelbert, L. M. & Gregg, R. E. (1997). Will genetics really revolutionize the drug discovery process? *Curr. Opin. Biotechnol.* **8**, 669–674.

Gelfand, M. S., Mironov, A. A. & Pevzner, P. A. (1996). Gene recognition via spliced sequence alignment. *Proc. Nat. Acad. Sci. USA,* **93**, 9061–9066.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. (1996). Life with 6000 genes. *Science,* **274**, 546–567.

Guigo, R. (1997). Computational gene identification: an open problem. *Comput. Chem.* **21**, 215–222.

Guigo, R. & Smith, T. F. (1993). Inferring correlation between database queries: analysis of protein sequence patterns. *IEEE Trans. Patt. Analysis Mach. Intell.* **15**, 1030–1041.

Guigo, R., Johansson, A. & Smith, T. F. (1991). Automatic evaluation of protein sequence functional patterns. *Comp. Appl. Biosci.* **7**, 309–315.

Han, K. & Kim, H.-J. (1993). Prediction of common folding structures of homologous RNAs. *Nucl. Acids Res.* **21**, 1251–1257.

Henikoff, S., Pietrokovski, S. & Henikoff, J. G. (1998). Superior performance in protein homology detection with the Blocks Database servers. *Nucl. Acids Res.* **26**, 309–312.

Hertz, G. Z., Hartzell, G. W. & Stormo, G. D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comp. Appl. Biosci.* **6**, 81–92.

Hieter, P. & Boguski, M. (1997). Modern modifications. *Science,* **278**, 601–602.

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.

Igarashi, T. & Kaminuma, T. (1997). Development of a cell signaling networks database. *Pac. Symp. Biocomp.* 187–197.

Jacq, B., Horn, F., Janody, F., Gompel, N., Serralbo, O., Mohr, E., Leroy, C., Bellon, B., Fasano, L., Laurenti, P. & Röder, L. (1997). GIF-DB, a WWW database on gene interactions involved in *Drosophila melanogaster* development. *Nucl. Acids Res.* **25**, 67–71.

Johnson, M. S., Srinivasan, N., Sowdhamini, R. & Blundell, T. L. (1994). Knowlegde-based protein modelling. *CRC Crit. Rev. Biochem. Mol. Biol.* **29**, 1–68.

Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. (1996). CENSOR: a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20**, 119–121.

Kahn, P. (1995). From genome to proteome: looking at a cell's proteins. *Science,* **270**, 369–370.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., *et al*. (1996). 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.* **3**, 109–136.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1998). EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **26**, 50–53.

Klenk, H. P., Clayton, R. A., Tomb, J-F., White, O., Nelson, K. E., Ketchum, K. E., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., Richardson, D. L., Kerlavage, A. R., Graham, D. E., Kyrpides, N. C., Fleischmann, R. D., *et al*. (1997). The complete genome sequence of the hyperthermo-

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 gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA,* **93**, 10268–10273.
Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic 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.
Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L. & Bernardi, G. (1996). Identification of the gene-richest bands in human chromosomes. *Gene,* **174**, 85–94.
Saito, K. & Nomura, M. (1994). Post-transcriptional regulation of the *str* operon in *Escherichia coli;* struc-

tural and mutational analysis of the target site for translational repressor S7. *J. Mol. Biol.* **255**, 125–139.
Sanchez, R. & Sali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins: Struct. Funct. Genet.* **Suppl. 1**, 50–58.
Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signalling domains. *Proc. Natl Acad. Sci. USA,* **95**, 5857–5864.
Schuster, S. (1996). Control analysis in terms of generalized variables characterizing metabolic systems. *J. Theoret. Biol.* **182**, 259–268.
Schuster, S. & Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* **2**, 165–182.
Schuster, S. & Hilgetag, C. (1995). What information about the conserved-moiety structure of chemical reaction systems can be derived from their stoichiometry? *J. Phys. Chem.* **99**, 8017–8023.
Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., 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.
Selkov, E., Maltsev, N., Olsen, G. J., Overbeek, R. & Whitman, W. B. (1997). A reconstruction of the metabolism of *Methanococcus janaschii* from sequence data. *Gene,* **197**, GC11–GC26.
Seressiotis, A. & Bailey, J. E. (1988). MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways. *Biotechnol. Bioeng.* **31**, 587–602.
Serry, L. T., Nestor, P. V. & FitzGerald, G. A. (1998). Molecular evolution of the aldo-keto reductase gene superfamily. *J. Mol. Evol.* **46**, 139–146.
Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, P., Pothier, B., *et al*. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155.
Smith, T. F. (1998). Functional genomics - bioinformatics is ready for the challenge. *Trends Genet.* **14**, 291–293.
Smith, T. F. & Zhang, X. (1997). The challenges of genome sequence annotation or ''The devil is in the details''. *Nature Biotechnol.* **15**, 1222–1223.
Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl. Acids Res.* **26**, 320–322.
Staden, R. (1989). Methods for discovering novel motifs in nucleic acid sequences. *Comp. Appl. Biosci.* **5**, 293–298.
Strauss, E. J. & Falkow, S. (1997). Microbial pathogenesis: genomics and beyond. *Science,* **276**, 707–712.
Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6**, 279–291.
Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science,* **278**, 631–637.
Thomas, S. & Fell, D. A. (1996). Design of metabolic control for large flux changes. *J. Theoret. Biol.* **182**, 285–298.
Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. & Ramaswamy, R. (1997). Prediction of probable genes by Fourier analysis of genomic sequences. *Comp. Appl. Biosci.* **13**, 263–270.
Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., *et al*. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature,* **388**, 539–547.
Trifonov, E. N. (1996). Interfering contexts of regulatory sequence elements. *Comp. Appl. Biosci.* **12**, 423–429.
Tsukamoto, Y., Kato, J. & Ikeda, H. (1997). Silencing factors participate in DNA repair and recombination in *Saccharomyces cerevisiae*. *Nature,* **388**, 900–903.
Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science,* **270**, 484–487.
Vietor, I. & Huber, L. A. (1997). In search of differentially expressed genes and proteins. *Biochim Biophys Acta,* **1359**, 187–199.
von Heijne, G. (1992). Membrane protein structure prediction, hydrophobicity analysis and the positive inside rule. *J. Mol. Biol.* **225**, 487–494.
Wirkner, U., Voss, H., Ansorge, W. & Pyerin, W. (1998). Genomic organization and promotor identification of the human protein kinase CK2 catalytic subunit alpha. *Genomics,* **48**, 71–78.
Wolfertstetter, F., Frech, K., Herrmann, G. & Werner, T. (1996). Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comp. Appl. Biosci.* **12**, 71–80.
Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.
Xu, C. W., Mendelsohn, A. R. & Brent, R. (1997). Cells that register logical relationships among proteins. *Proc. Natl Acad. Sci. USA,* **94**, 12473–12478.
Yates, J. R., III (1998). Mass spectrometry and the age of the proteome. *J. Mass Spectrom.* **33**, 1–19.
Yuan, Y. P., Eulenstein, O., Vingron, M. & Bork, P. (1998). Towards detection of orthologues in sequence databases. *Bioinformatics,* **14**, 285–289.
Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997). Gene expression profiles in normal and cancer cells. *Science,* **276**, 1268–1272.
Zweiger, G. & Scott, R. W. (1997). From expressed sequence tags to 'epigenomics': an understanding of disease processes. *Curr. Opin. Biotechnol.* **8**, 684–687.

Edited by P. E. Wright

(Received 22 May 1998; received in revised form 13 August 1998; accepted 13 August 1998)