FEBS Letters 389 (1996) 96–101 FEBS 17138

Minireview

Genomes with distinct function composition

Javier Tamames[a], Christos Ouzounis[b], Chris Sander[c,d], Alfonso Valencia[a,*]

[a] *Centro Nacional de Biotecnologia – CSIC, Madrid, Spain*
[b] *AI Center, SRI International, Menlo Park, CA, USA*
[c] *EMBL, Heidelberg, Germany*
[d] *EBI-EMBL, Hinxton Hall, Cambridge, UK*

Received 10 May 1996

Abstract The functional composition of organisms can be analysed for the first time with the appearance of complete or sizeable parts of various genomes. We have reduced the problem of protein function classification to a simple scheme with three classes of protein function: energy-, information- and communication-associated proteins. Finer classification schemes can be easily mapped to the above three classes. To deal with the vast amount of information, a system for automatic function classification using database annotations has been developed. The system is able to classify correctly about 80% of the query sequences with annotations. Using this system, we can analyse samples from the genomes of the most represented species in sequence databases and compare their genomic composition. The similarities and differences for different taxonomic groups are strikingly intuitive. Viruses have the highest proportion of proteins involved in the control and expression of genetic information. Bacteria have the highest proportion of their genes dedicated to the production of proteins associated with small molecule transformations and transport. Animals have a very large proportion of proteins associated with intra- and intercellular communication and other regulatory processes. In general, the proportion of communication-related proteins increases during evolution, indicating trends that led to the emergence of the eukaryotic cell and later the transition from unicellular to multicellular organisms.

Key words: Database annotation; Functional class; Genome composition

1. Introduction: functional classifications of proteins

The transition from vitalistic to mechanistic theories was marked by a quest for the basic components of living systems. In earlier times, biological research revealed a large number of small molecules that participated in metabolic pathways. With the advent of molecular biology, the gene and its product became the compositional quantum for biological systems, now at the macromolecular level. While information about single genes and their functions is being accumulated for decades, a single pattern of composition for a whole organism can only emerge now with the availability of (large or complete fractions of) model genomes. In the sequence databases, some species already have a quite large number of sequences that can be considered to be a typical subset of their total genome. The goal is to classify these sequences into general functional classes and perform compositional comparisons between these model genomes. Surprisingly, meaningful patterns emerge, consistent with different data sets and for phylogenetically related organisms.

With the effort of many different groups and the contribution of different sequencing projects, we now know tens of thousands of proteins collected in databases such as Swiss-Prot [1]. This massive amount of data allows us to answer in a quantitative way questions about the functional composition of information storage in model organisms. Until now the amount of data available has precluded any comparative analysis, and only information about single organisms has been manually analysed in a systematic way. Riley has classified protein functions in different groups using the available data about *Escherichia coli* [2] and Venter and colleagues have classified the complete genomes of *Mycoplasma genitalium* [3] and *Haemophilus influenzae* [4] as well as expressed sequence tags collections of human libraries into functional groups [5].

Here we review and extend this type of analysis, by classifying all the available sequences of many different species in a few key functional groups. We have selected only three generic functional classes to gain maximal statistical significance within a meaningful biological context. The three classes selected are a generalisation of the ones proposed by Riley [2] and later used Venter et al. [3,4], the detailed cross-reference between them is given in Table 1. Our three classes represent the following processes:

ENERGY, representing the effort of living beings to maintain themselves against the medium. It includes proteins related to metabolism: anabolism, catabolism and intracellular transport. This class is in general related to binding to small molecules, e.g. cofactors or metabolites.

INFORMATION, representing replication and proliferation through time. It includes proteins related to DNA structure, replication and repair, transcription, splicing and translation. Most proteins in this class bind genetic informational macromolecules, e.g. DNA and RNA.

COMMUNICATION, representing the interaction with the medium and communication between cellular states. Channels and transporters but also proteins of the cell cycle are included in this category. Protein-protein and protein-complex carbohydrate interactions are very common in this class.

With this classification, we intend to give a quantitative answer to the following questions:

1. What is the proportion of different protein functions found in living forms?
2. Does this proportion differ between taxonomic levels (ancient vs recently evolved taxa)?

*Corresponding author. Protein Design Group, CNB-CSIC, Campus Universidad Autónoma, Cantoblanco, E-28049 Madrid, Spain. Fax: (34) (1) 585 45 06. E-mail: valencia@samba.cnb.uam.es

PII S0014-5793(96)00527-3

2. Automatic classification of sequences in functional groups

A small data set can be accurately classified by human experts [2–6], but the avalanche of new sequence data demands the application of automatic systems able to perform this task in an objective and reproducible way. We have recently developed such a system [7]. In short, the procedure starts with sequences classified by human experts, extracts from these sequences an initial set of keywords clearly associated with one functional class, and classifies all sequences from the database with this dictionary of keywords. The process can be iterated by extracting again the keywords from the sequences already classified, building a new dictionary and classifying again all the sequences in the database. As initial input, we have used the classification by human experts of sequence data sets from yeast, high vertebrates, *E. coli* and plants. The final dictionary used for this study contains 536 uniquely assigned keywords and covers 81% of the sequences in the database with some functional annotation (see Table 1 for a list of the most represented keywords). The accuracy of the system is around 80% when tested with a cross-validation procedure applied to randomly selected sets of 100 sequences from the previously assigned sequences [7]. The coverage obtained for the different species in the database is given in Fig. 1.

The automatic system for functional classification belongs to a very general class of text understanding systems [8]. These systems attempt to interpret text and extract relevant information: examples of such systems include the FASTUS or SCISOR systems [8]. From the experience in this area of research, it is evident that tasks performed accurately by experts, with insights about the domain problems and human generalisation capacity, are very difficult for computer systems that are limited to the present data and have difficulties generalising. On the contrary, and given their limitations, automatic systems can analyse vast amounts of data reproducibly. The idea of using database functional annotations about the sequences has been previously explored by Guigo and Smith [9]. They reported a double clustering of the database by sequence motifs and keywords with the goal of discovering new associations between functions (keywords) and sequence families (defined by motifs).

In the case of sequence databases, the current limitations are the presence of many sequences with scarce or totally absent functional annotations, lack of suitable annotations for functional classifications, and uneven representation of species in the databases.

Despite the various technical limitations, the results of the automatic classification are comparable to the ones reported by human experts, for example in the case of the classification of *E. coli* sequences [2], reportedly at 60% ENERGY, 30% INFORMATION and 9% COMMUNICATION when reduced to the three classes used here. These numbers match well with our results, at 51% ENERGY, 45% INFORMATION and 4% COMMUNICATION. A part of the discrepancies comes from the differences in the number of sequences used in both analyses, for a systematic comparison see [7].

3. Consequence of systematic sequencing

The analysis of the functional annotation in Swiss-Prot, the best annotated protein sequence database, show the existence

Table 1
Protein functions assigned to each functional class: (a) correspondence between standard protein function classification [2] and the three classes schema and (b) most representative keywords in each class

Energy	Information	Communication
(a) Type of protein function included in each class		
Transport through membrane	Replication/repair/DNA	Signal transduction/regulation
Transporters/symporters	Replication	Interaction with environment
Phosphorylation	Recombination	Recognition
Amino acid metabolism	Repair	Adhesion
Pathways	Cell division	Defense (toxic substances)
Modifications	Other nuclear	Extracellular degradation
Nucleotide metabolism	Gene expression/RNA	Carbohydrates
Pyrimidine	Transcription	Proteins
Purine	Translation	
Lipid metabolism	Ribosomal proteins	
Carbohydrate metabolism/energy conservation	Splicing	
Preglycolytic sugar modification/degradation	Others	
Pentosephosphate cycle	Proteins	
Glycolysis	Protein biosynthesis	
Electron transport/Respiration	Folding	
Gluconeogenesis	Internal transport/translocation	
Carbohydrate synthesis	Posttranslational modification	
Storage and starvation	Protein degradation	
ppGpp		
Carbamylphosphate		
Secondary pathways, other	(Structural proteins are not used in this classification)	
(b) The most represented keywords in different classes are:		
photosynthesis	activator	hormone
oxygen transport	zinc finger	amidation
respiratory protein	early protein	ser/thr protein kinase
monooxygenase	nuclear protein	G-protein coupled receptor
kinase	developmental protein	serine protein inhibitor
hydrophobic ion transporter		cell adhesion

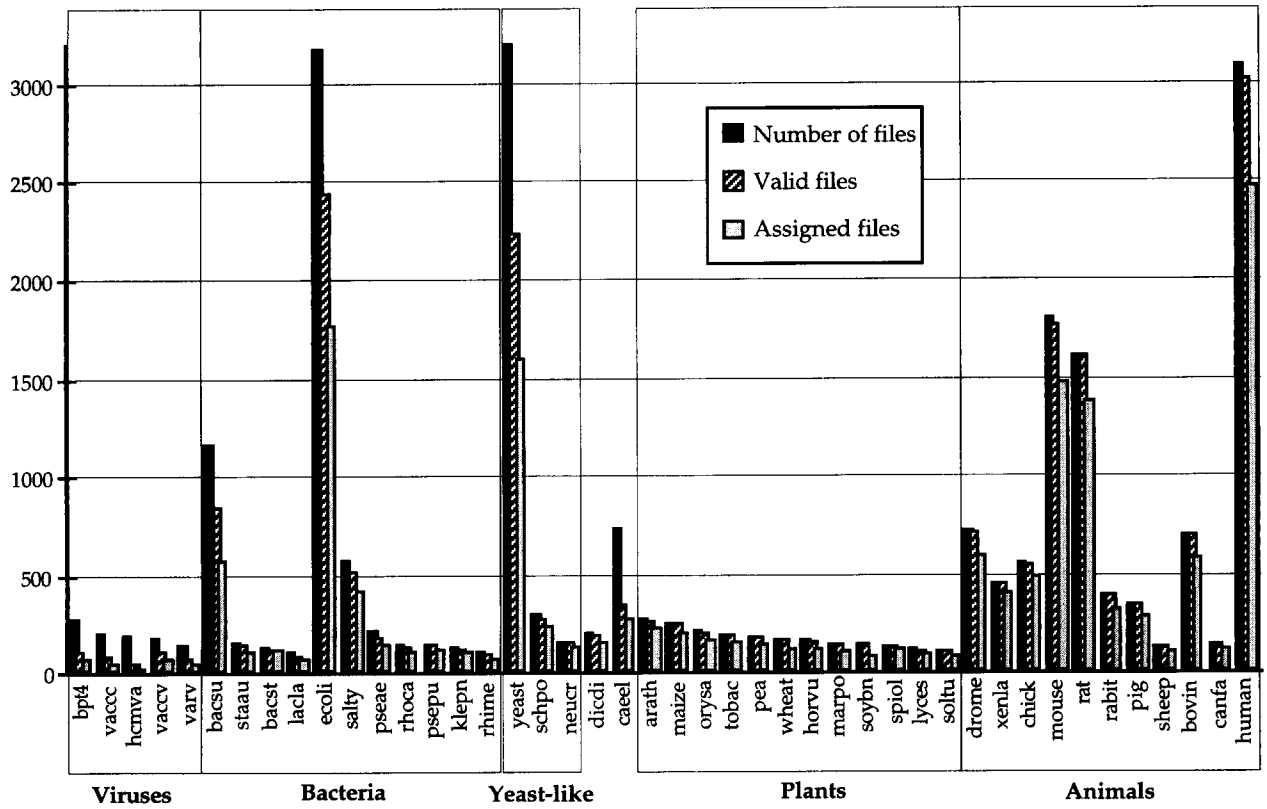

Fig. 1. Functional information content in the Swiss-Prot sequence database and coverage of the classification in three classes. All sequences deposited in the database were classified in their corresponding species. 'Number of files' (closed bars): total number of sequences for each species. 'Valid files' (hatched bars): number of entries annotated with meaningful keywords (i.e. annotations such as hypothetical are not considered). 'Assigned files' (shadowed bars): number of entries finally classified by the system. The average coverage of the classification method is 81% over the number of sequence with keywords (valid files). This proportion is much lower in those species subject to systematic sequencing. This probably reflects the lack of full functional analysis of all these new sequence data. In this analysis, only species with more than 100 sequences in the database were taken into account. At the bottom of the figure the classification of the species in five groups is given: viruses: 5, bacteria: 11, yeast-like: 3, plants: 12 and animals: 11 species. The study was done with Swiss-Prot version 31.0 from February 1995. This version did not contain the full sequences of *H. influenzae* and *M. genitalium*. The species represented are: bpt4: bacteriophage T4; vaccc: Vaccinia virus (strain Copenhagen); hcmva: human cytomegalovirus (strain adl69); vaccv: vaccinia virus (strain wr); varv: variola virus; bacsu: *Bacillus subtilis*; staau: *Staphylococcus aureus*; bacst: *Bacillus stearothermophilus*; lacla: *Lactococcus lactis* (subsp. lactis); ecoli: *Escherichia coli*; salty: *Salmonella typhimurium*; pseae: *Pseudomonas aeruginosa*; rhoca: *Rhodobacter capsulatus*; psepu: *Pseudomonas putida*; klepn: *Klebsiella pneumoniae*; rhime: *Rhizobium meliloti*; yeast: *Saccharomyces cerevisiae*; schpo: *Schizosaccharomyces pombe* (fission yeast); neucr: *Neurospora crassa*; dicdi: *Dictyostelium discoideum*; caeel: *Caenorhabditis elegans*; arath: *Arabidopsis thaliana*; maize: *Zea mays* (maize); orysa: *Oryza sativa* (rice); tobac: *Nicotiana tabacum* (tobacco); pea: *Pisum sativum* (garden pea); wheat: *Triticum aestivum* (wheat); horvu: *Hordeum vulgare* (barley); marpo: *Marchantia polymorpha* (liverwort); soybn: *Glycine max* (soybean); spiol: *Spinacia oleracea* (spinach); lyces: *Lycopersicon esculentum* (tomato); soltu: *Solanum tuberosum* (potato); drome: *Drosophila melanogaster*; xenla: *Xenopus laevis* (african frog); chick: *Gallus gallus* (chicken); mouse: *Mus musculus*; rat: *Rattus norvegicus*; rabit: *Oryctolagus cuniculus* (rabbit); pig: *Sus scrofa*; sheep: *Ovis aries*; bovin: *Bos taurus*; canfa: *Canis familiaris* (dog); human: *Homo sapiens*.

of a large number of sequences without any functional annotation (Fig. 1). These sequences are currently more than 1700. Their number is increasing with the expansion of systematic sequencing projects. This influence can be clearly noticed in species such as *Escherichia coli*, *Bacillus subtilis*, *Caenorhabditis elegans* or *Saccharomyces cerevisiae*.

4. Life as three basic processes

The plain average of all species have a relative composition of: 40.61 ± 20.14% ENERGY, 37.07 ± 17.77% INFORMATION and 22.31 ± 19.60% COMMUNICATION. The deviation of these numbers is remarkably high, and it originates from differences between species and taxonomic groups. The analysis of these differences could help us to understand how the functional composition is adapted in different forms of life. It is interesting that the class with more relative deviations is COMMUNICATION, as this class diverges drastically between different groups of organisms.

5. Different functional composition of groups and species

In Fig. 2A, the basic proportional composition of different species and groups is given. The groups chosen (viruses, bacteria, yeasts, plants and animals) are not homogeneous in the phylogenetic sense but the current database does not allow a more fine-grained classification. In this selection, all of the groups have similar numbers of species but not similar proportions of sequences in the different functional categories.

The differences are very striking. Viruses, including big viruses and bacteriophages, have the overall lower proportion of proteins classified in the ENERGY class, with many of

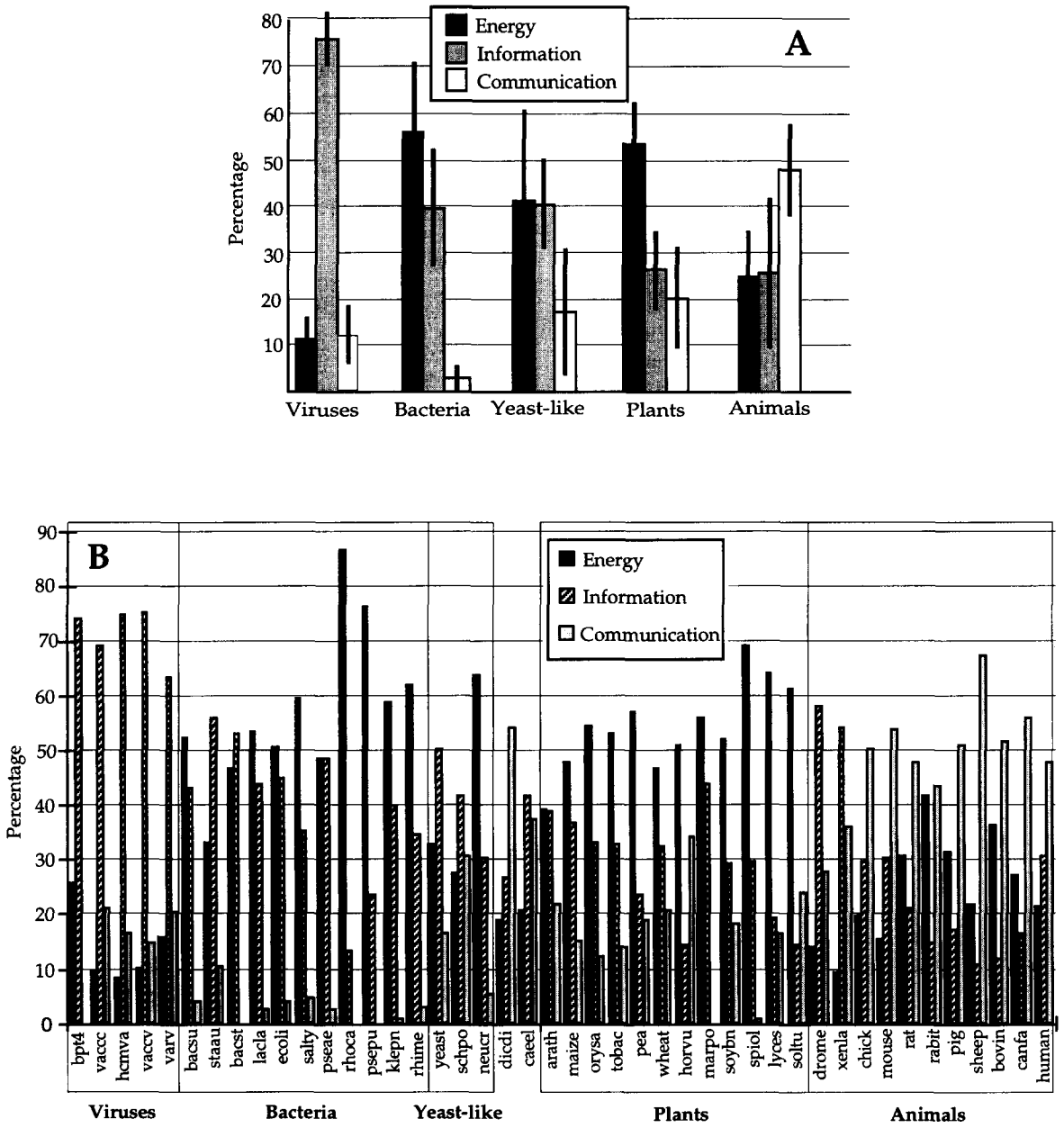

Fig. 2. Basic composition of living beings. The relative proportions of proteins classified in the different categories ENERGY, INFORMATION and COMMUNICATION are given. Only sequences represented with more than 100 sequences in Swiss-Prot have been considered. (A) Basic composition of different groups of species. The proportion of proteins classified in the three different classes is shown for five groups of species (Viruses, Bacteria, Yeast-like, Plants and Animals). The groups have been selected to contain similar species in the restricted possibilities imposed by current databases. Perhaps the two most striking observation is the gradual increase in the proportion of functions classified as COMMUNICATION, from bacteria to animals. This increase is done mainly at the expenses of the INFORMATION class. (B) Functional composition of individual species. All species used in the analysis are shown in the same order as in Fig. 1. The same trend observed in panel A is now spliced at the level of species. Two observations are interesting: species subject to systematic sequencing have a behaviour very similar to their neighbour species (i.e. ecoli versus other bacteria). Therefore the bias introduced by the arbitrary decision about what proteins to sequence seems not to affect the analysis at this level. There are particular species with outstanding behaviours, i.e. dicdi. It is here more difficult to separate their biological particularities, i.e. the cAMP-dependent regulated signalling in *D. discoideum*, from the experimental bias to study this particular cycle.

them in INFORMATION, e.g. polymerases. That is an obvious consequence of their adaptation to use other cells' machinery. It is interesting, however, that there is still a number of COMMUNICATION-related proteins in viruses.

Bacterial cells have a very high proportion of proteins in the ENERGY group and very few of them in the COMMUNICATION class. This may be the standard composition of a free cell with a prokaryotic lifestyle.

Yeasts, represented by only three species, have a larger proportion of functions in COMMUNICATION than bacteria with a smaller proportion of functions in the ENERGY class.

The proportion of functions in plants is between yeast and bacteria, with the proportion of functions in COMMUNICATION very similar to the yeast class. This low number of proteins in the COMMUNICATION class may reflect the simpler organisation of plants with fewer cell types than animals, and communication processes carried out by small molecules and not proteins in many cases. The values between different plants species are remarkably homogeneous, a fact that adds credibility to the interpretation.

Different animal species have a much larger proportion of proteins related to COMMUNICATION and INFORMATION than any of the other groups. Correspondingly, they have an smaller proportion of protein functions related with ENERGY. Our interpretation is that animal cells have more regulated processes (INFORMATION) and participate in more processes of relation with the medium or other cell types (COMMUNICATION).

The similarity between sequences that belong to the same group can be better seen in the full display of all species (Fig. 2B). Viruses show a very similar behaviour, in bacteria some species deviate more from the average behaviour. For example, *Rhodobacter capsulatus* (rhoca) and *Pseudomonas putida* (psepu) have a very high proportion of ENERGY. In plants, the most deviating cases are *Spinacia oleracea* (spiol), *Lycopersicon esculentum* (lyces) and *Solanum tuberosum* (soltu), in which again ENERGY dominates. Finally, of the different animal species, *Drosophila melanogaster* (drome) and *Xenopus laevis* (xenla) have a higher proportion of functions in INFORMATION. This is partially due to the high number of homeobox genes specifically cloned in both species for the study of development.

These examples and some other species with exceptional behaviours could be interesting from the biological point of view. *C. elegans* is similar to higher animals with a large proportion of proteins in the COMMUNICATION class and *Neurospora crassa* has a composition that overall is more similar to bacteria than to eukaryotes. It is still too early to know if the specific features reflect real biological facts or are a consequence of experimental bias introduced from sequencing.

One of the most striking cases is *Dictyostelium discoideum* with a large percentage of sequences in COMMUNICATION. This stems from proteins associated with the cAMP signalling pathway, highly specific to *D. discoideum* for the control of cell migration and organisation.

Finally, *E. coli*, *B. subtilis* and other bacteria have very similar compositions. This is an important observation, since it demonstrates that species subject to systematic sequencing (*E. coli*, *B. subtilis*) show a behaviour similar to other species for which the sequence data have not been obtained systematically. Apparently, non-systematic sequencing is not producing a strong bias towards certain functions favoured by experimental groups. This conclusion is only valid for the three functional classes analysed here and in the future it may be challenged by larger data sets and finer classifications schemes.

6. Increasing complexity: more COMMUNICATION-related proteins

There is a clear trend from simple to complex organisms to increase the number of functions in the area of COMMUNICATION. This higher proportion of proteins dedicated to communication with the environment can be seen as the natural consequence of the complex organisation of multicellular organisms. It is interesting that plants have a composition more similar to bacteria, since their communication processes are in many cases carried out by small molecules instead of proteins.

7. Importance of metabolism for evolutionary studies

It is also obvious that all living beings devote a large proportion of their proteins to energy-associated processes (energy transformations, transport). This shows the importance of studying metabolism as a preserved mechanism along evolution. With more sequences in the database and better annotations, it will become possible in the future to update the current classification. A new schema with more functional groups will add further details to the general picture outlined here.

In a different context, the method and dictionaries for functional classification now developed could be an invaluable tool in the analysis of how information is organised at other levels, i.e. full chromosomes or chromosome regions. It has also been used during sequence analysis of genome information [10]. We are in the process of integrating it with a system for large-scale sequence analysis [11].

8. Towards comparative genome analysis

There is an obvious and interesting parallelism between the problem of genome comparison and molecular sequence comparison. Different methods have been described for comparing proteins and DNA by their composition, for example see [12]. Analogously, we describe the composition of genomes by their composition in functional classes.

To proceed further, and describe these entities at a more detailed level, order in addition to composition can be taken into account. A simple form of this approach is nearest-neighbour information. Karlin and collaborators have pioneered this type of comparison for DNA and protein sequences [13]. This level of description is sufficient to describe many of the important properties of biomolecular organisation. We have recently extended a similar idea for genome comparison (Tamames et al., submitted).

Finally, it remains to be seen how many aspects of traditional sequence analysis are applicable to whole genome information, since basic operations of transposition and inversion have no counterpart in DNA or protein sequence comparison. New methodological approaches are needed for comparative genome analysis.

References

[1] Bairoch, B. and Boeckmann, B. (1994) Nucleic Acids Res. 22, 3578–3580.
[2] Riley, M. (1993) Microbiol. Rev. 57, 862–952.
[3] Fraser, C.M. et al. (1995) Science 270, 397–403.
[4] Fleischmann, R.D. et al. (1995) Science 269, 496–512.
[5] Adams, M.D., Kerlavage, A.R., Fields, C. and Venter, J.C. (1993) Nature Genet. 4, 256–267.
[6] Ouzounis, C., Valencia, A., Tamames, J., Bork, P. and Sander, C. (1995) In: Proceedings of the Third European Conference on Artificial Life (Moran, F., Moreno, A., Merelo, J.J. and Chacon, P., Eds.) pp. 843–851. Springer-Verlag, Berlin.

[7] Tamames, J., Casari, G., Ouzounis, C., Sander, C. and Valencia, A. (1996) submitted.
[8] Allen, J. (1995) Natural Language Understanding. Benjamin/Cummings, Menlo Park, CA.
[9] Guigo, R., Johansson, A. and Smith, T.F. (1991) Comput. Appl. Biosci. 7, 309–15.
[10] Casari, G., Andrade, A., Bork, P., Boyle, J., Daruvar, A., Ouzounis, C., Schneider, R., Tamames, J., Valencia, A. and Sander, C. (1995) Nature 376, 247–248.
[11] Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. and Sander, C. (1994) In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology (Altman et al., Eds.) pp. 348–353, AAAI Press, Menlo Park.
[12] Hobohm, U. and Sander, C. (1995) J. Mol. Biol. 251, 390–9.
[13] Karlin, S. and Ladunga, I. (1995) Proc. Natl. Acad. Sci. USA 91, 12832–12836.