Computational comparisons of model genomes

Christos Ouzounis, Georg Casari, Chris Sander, Javier Tamames and Alfonso Valencia

Complete genomes from model organisms provide new challenges for computational molecular biology. Novel questions emerge from the genome data obtained from the functional prediction of thousands of gene products. In this review, we present some approaches to the computational comparison of genomes, based on sequence and text analysis, and comparisons of genome composition and gene order.

With the recent publication of the complete genome sequences from two bacteria, *Haemophilus influenzae* Rd[1] and *Mycoplasma genitalium*[2], new challenges are emerging for computational biology. Two such challenges are (1) to predict and annotate the functions of the gene products as rapidly and completely as possible, and (2) to derive adequate abstractions that make genomes comparable at a higher-than-molecular level.

Function prediction is a primary goal of genome-sequence analysis, as many newly determined sequences have no experimental information associated with them, while functional information can be derived by examining homology to proteins of known function. Prediction can be carried out by integrating and co-ordinating a number of well-tested methods that rapidly and efficiently identify sequences with the highest degree of similarity from complete databases, and can, therefore, assist function prediction using homology observations. The GeneQuiz system[3] automatically annotates protein-encoding sequences and can identify novel functions, e.g. for *H. influenzae*[4] and *M. genitalium*[5] sequences*. The use of integrated databases and software tools, combined with the application of a number of empirical rules that can automatically eliminate false annotations[6], make this possible.

If the functional annotations are known, what is the next step in genome analysis? We have been exploring new ways to make use of this information, and have been performing global comparisons of genomic data, addressing a number of questions. These analyses yield a profile of the composition of genomic functions of an organism, identifying some components that are common to other species, and some that appear to be unique. These predicted 'expression patterns' can help in the identification of novel or potential metabolic or regulatory pathways, and provide a faster route to the development of targets for drug design and discovery.

Orthologues: functionally equivalent genes across species

A gene with a certain level of sequence similarity to its homologue in the genome of another species may have the same function as its homologue; such genes are defined as orthologues[7]. How is it possible to determine whether the two proteins encoded by the genes of different species have the same function? Proteins change during evolution, forming families of related molecules that have similar primary, secondary and tertiary structures, but which have divergent functions. Algorithms for sequence comparison can detect genes encoding homologous proteins, but are unable to determine definitively whether two molecules have exactly the same function. Therefore, function prediction by detection of sequence similarity, especially in incomplete genomes, can only be approximate, because it usually makes use of genes encoding the most similar proteins; however, these genes might not yet have been identified.

In complete genomes, however, there is a finite number of genes, so it is possible to determine which genes share the highest degree of similarity, thus narrowing the set of experiments that must be performed to prove that proteins encoded by orthologous genes have the same function. Therefore, the average similarity value can be calculated, helping in the estimation of the rate of change in different families. The

C. Ouzounis (ouzounis@ai.sri.com) is at the Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA. A. Valencia and J. Tamames are at the Protein Design Group, Centro Nacional de Biotechnologia, CSIC, Campus U. Autonoma, E-28049 Madrid, Spain. G. Casari is at the European Molecular Biology Laboratory, Meyerhofstraße 1, D-69012 Heidelberg, Germany. C. Sander is at the European Bioinformatics Institute, EMBL, Hinxton Hall, Cambridge, UK CB10 1RQ.

*Analysis of the *H. influenzae*, *M. genitalium* and *S. cerevisiae* genomes, including functional classification, is available on the World Wide Web at <http://www.sander.embl-heidelberg.de/genequiz/>.

 0167 – 7799/96/$15.00. PII: S0167-7799(96)10043-3

redundancy of homologous families of proteins (e.g. transport proteins in bacteria) can also be measured, and the most conserved pairs of proteins can be identified. The first analysis of this kind was performed with a contig set from *Mycoplasma capricolum*[8] that encoded almost half of the total number of proteins. We have found that the average similarity between protein-encoding genes in *M. capricolum* and *Escherichia coli* is 40%, with a wide variance (Fig. 1). It should be noted that analyses of putative orthologues in incomplete genomes provide a lower estimate of sequence similarity between orthologues. There may be other genes, as yet unidentified in either genome, that display an even higher similarity.

The identification of orthologues can be used in the reconstruction of metabolic pathways encoded by various bacterial genomes; *E. coli* metabolism has been used as the model[9]. This approach can provide a deeper understanding of the primary-sequence data, and can help in the identification of potentially interesting targets for medical, industrial and environmental applications.

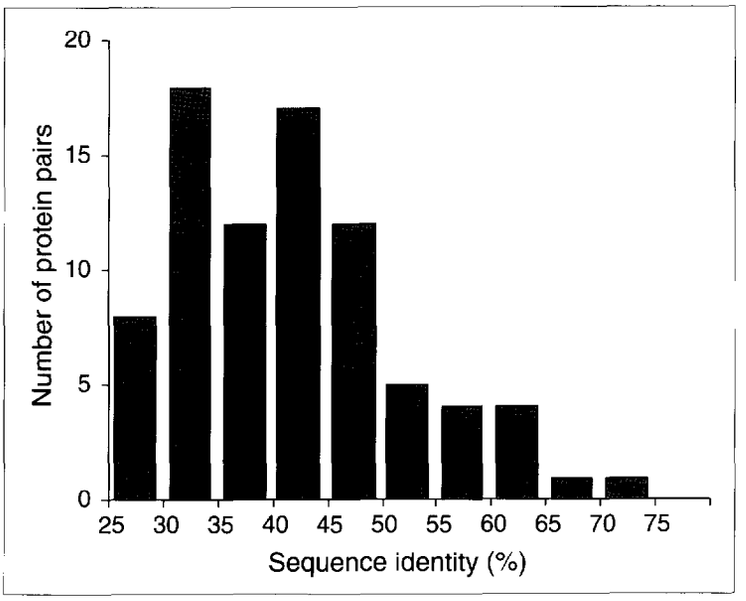


Figure 1

Frequency histogram indicating the sequence identity of genes encoding 82 *Mycoplasma capricolum* proteins with their putative orthologues in *Escherichia coli*[8]. These numbers represent lower estimates: if protein-encoding genes with higher sequence-similarities are found in either genome, the distribution would probably shift to the right. The three most conserved gene pairs between the two bacteria are those encoding FtsH, enolase and ClpB, all of which share ≥63% sequence identity over the length of available sequences (60–125 amino acids).

Genome compositions from automated function-classification

Functional classification of genes and gene products is another aspect of genome analysis that can provide a basis for comparative genomics. This type of

Box 1. Functional classes for genome comparison

Our definition of functional classes used for genome comparisons corresponds with the one that was previously proposed[23] for characterizing *Escherichia coli*. Of the nine classes, eight are shown, and these can be classified into three super-classes; the ninth class consists of unknown functions, and is not shown. Typical keywords for the super-classes are: **ENERGY** – photosynthesis, oxygen transport, respiratory protein, monooxygenase, kinase, hydrophobic ion transporter; **INFORMATION** – activator, zinc finger, early protein, nuclear protein, developmental protein; **COMMUNICATION** – hormone, amidation, serine/threonine protein kinase, G-protein coupled receptor, serine protease inhibitor, cell adhesion.

ENERGY

Metabolism
- Amino acid metabolism
- Cofactor biosynthesis
- Prosthetic groups and carriers
- Central intermediary metabolism
- Energy conservation (e.g. sugar modification and degradation, pentose phosphate cycle, glycolysis, gluconeogenesis, electron transport, respiration, secondary metabolism)
- Fatty acid and phospholipid biosynthesis
- Purines, pyrimidines, nucleosides, nucleotides

Transport
- Transport through membranes

INFORMATION

DNA and RNA
- Replication (including recombination and repair)
- Transcription (including splicing)

Translation
- Translation, ribosomal proteins

Proteins
- Protein biosynthesis, folding, internal transport, translocation, post-translational modification and degradation

COMMUNICATION

Signal
- Regulatory functions
- Cell division
- Cell killing

Environment
- Interaction with environment: recognition, adhesion, defense (toxic substances), extracellular degradation

Structure
- Structural proteins

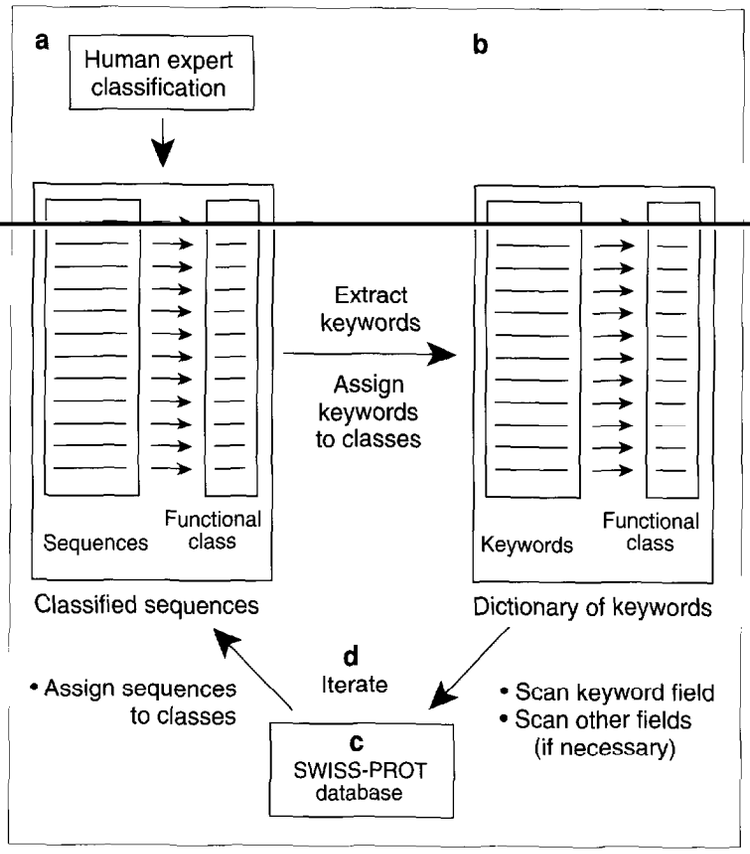


Figure 2

Schema of the method used to classify sequences in functional classes. (**a**) Sequences are classified into functional classes; (**b**) a dictionary with keywords corresponding to each functional class is created; (**c**) the database is searched and all sequences are classified into functional classes; and (**d**) the procedure is iterated. Classified sequences can be input manually, or from the automatic classification provided by the method, or from a combination of both methods.

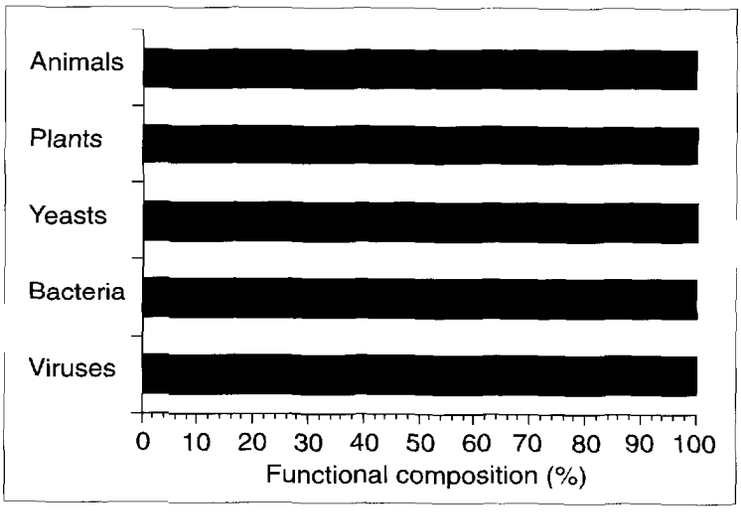


Figure 3

Basic functional composition of various taxa[12]. Only organisms with more than 100 sequences in SWISS-PROT are represented. Viruses include phages and eukaryotic viruses, both displaying very similar composition. For each taxon, (blue) energy-, (green) information- and (red) communication-related composition is shown as a percentage of the total. The three most interesting patterns are: (1) that viruses have a small, but significant, amount of metabolic enzymes (mostly involved in nucleotide biosynthesis), (2) that yeasts and plants are remarkably similar, and (3) that there is an increase in communication-related proteins during cellular evolution.

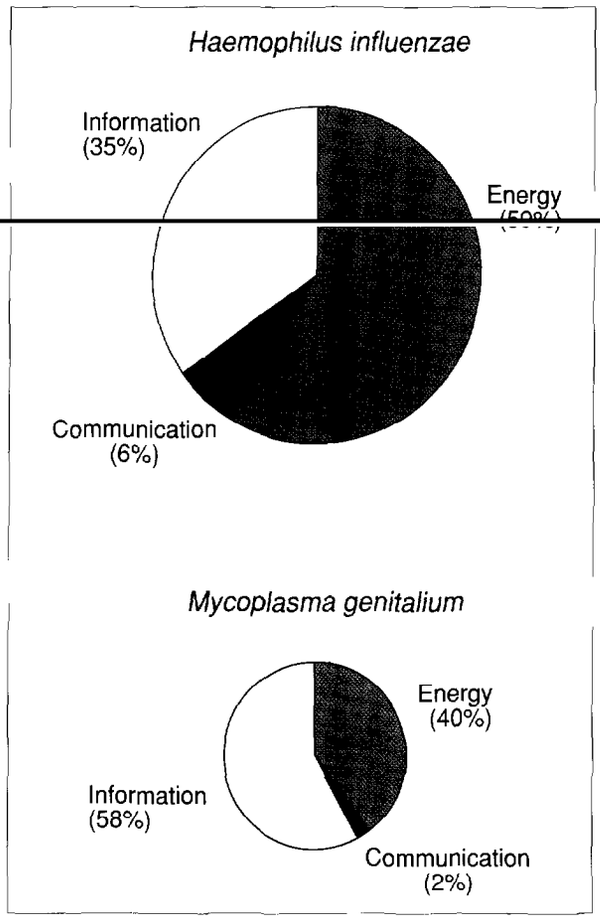


Figure 4

The functional compositions of two complete genomes[4,5]. The relative sizes of the charts reflects genome size. While *Haemophilus influenzae* displays a typical composition for a bacterial genome (compare with Fig. 3), *Mycoplasma genitalium* appears to devote most of its coding potential to replication and translation, while its metabolic capacity is restricted, owing to its parasitic lifestyle.

approach raises the following questions: How can we devise useful abstractions that allow model genomes to be compared? In other words, how can we classify proteins beyond the similarity of the sequences encoding them? The problem is, how can genomes be compared when sequence similarity is not sufficient; for example, if most homologous proteins have not been identified, or if they are too divergent or simply not as abundant between species (e.g. between mycoplasmas and mammals). We have approached this problem by categorizing protein functions into nine classes (eight classes of major functions, and one unknown; Box 1). These classes, which correspond with a number of cellular processes, can be clustered into three super-classes: energy-, information- and communication-related genes and proteins[10], in a simplification that has a physical basis; i.e. the substrates of these super-classes are usually small molecules, nucleic acids or proteins, correspondingly. This situation is analogous to that seen for expression patterns obtained from expressed sequence tag (EST) sequencing[11], except that the compositional profiles define a potential, and not an actual, genetic pattern.

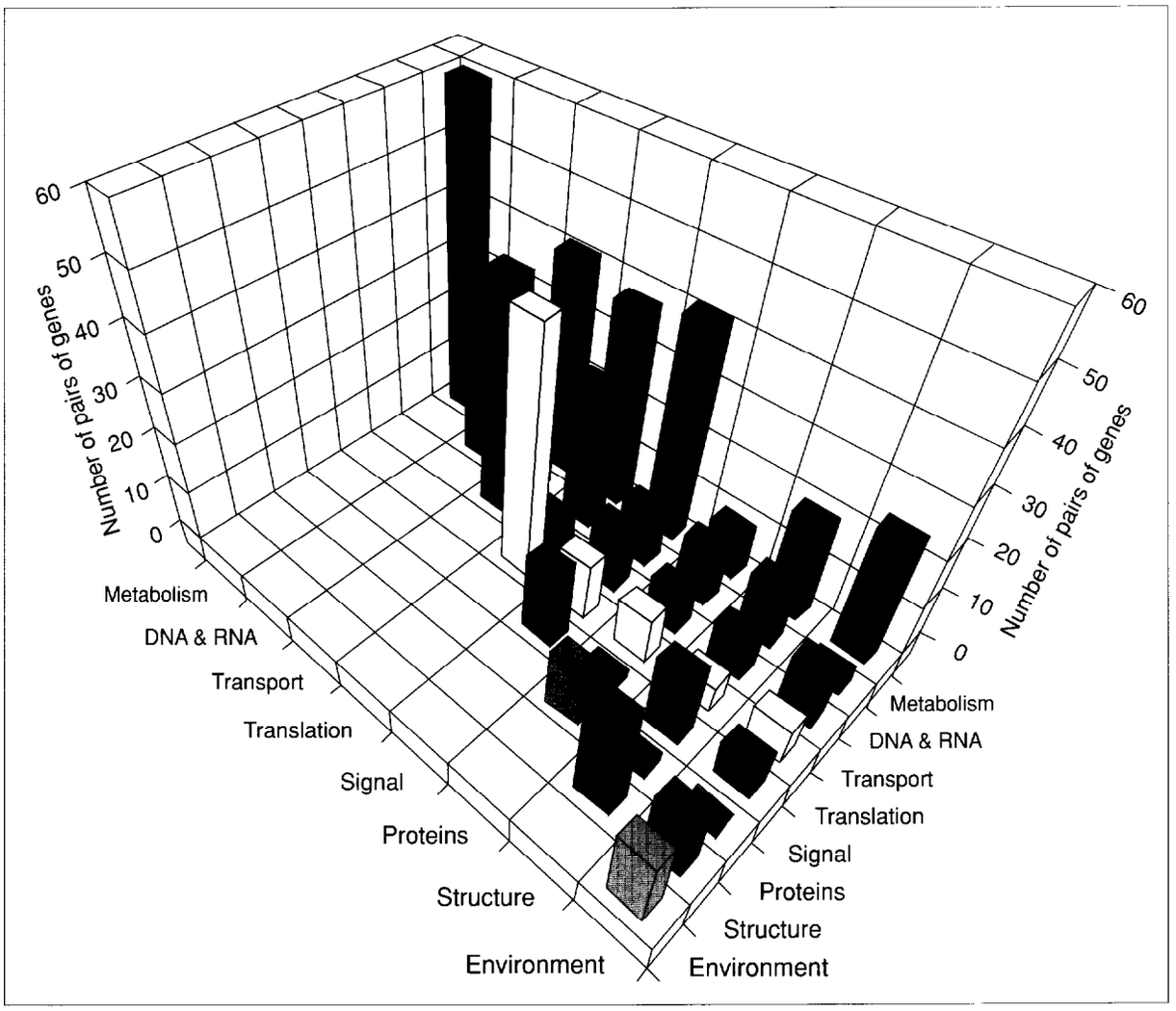


Figure 5

Numbers of pairs of functionally related genes in the genome of *Haemophilus influenzae*[1]. The two planar axes represent the classes in which neighboring genes belong, the third axis represents the number of gene pairs. It is clear that the numbers in the diagonal value dominate, corresponding to adjacent genes of the same functional class[22]. Genes for which there is no function assignment, or cases where the function assignment is not clear, are classified as unknown (and are not shown); values above 60 (only for pairs of metabolic genes) are also not shown.

We have developed an automatic functional classification system (Fig. 2; J. Tamames, G. Casari, C. Ouzounis, C. Sander and A. Valencia, unpublished) to classify thousands of predicted protein functions as rapidly and reliably as possible. A set of pre-classified sequences is used to build a dictionary of keywords characteristic of each functional class, and uses this dictionary to classify other sequences based on their keywords. The process is repeated, starting with the enlarged set of classified sequences, deriving new associations between keywords and functional classes, and classifying more sequences. In this way, the dictionaries are enriched with unique and distinctive patterns of keywords that describe each functional class. Inconsistencies, conflicts and other problems are dealt with, and the procedure converges when no more keywords can be uniquely associated with a class (Fig. 2).

This system has been used to estimate the compositions of functional classes of a number of species that are well-represented in the sequence databases[12]. For example, it appears that during evolution, the fraction of (intra- and extra-cellular) communication-related proteins increases monotonically with the phylogenetic status of major taxa (Fig. 3). This pattern of increase clearly reflects the differences between prokaryotic and eukaryotic cells, and the changes that occurred during the transition from unicellular to multicellular organisms. In our analysis, differences in genome composition are observed between the genes encoding the complete sets of identified proteins of *H. influenzae* and *M. genitalium* (Fig. 4), and for the genes encoding the proteins of *Saccharomyces cerevisiae* (yeast), the first completely sequenced eukaryotic genome[13]. Functional compositions correspond well with the estimates previously obtained from the taxa in which these species belong. These compositional patterns are useful general descriptors of the genomic make-up of model organisms, and implicitly define a similarity measure between genomes that can be used in future comparative studies.

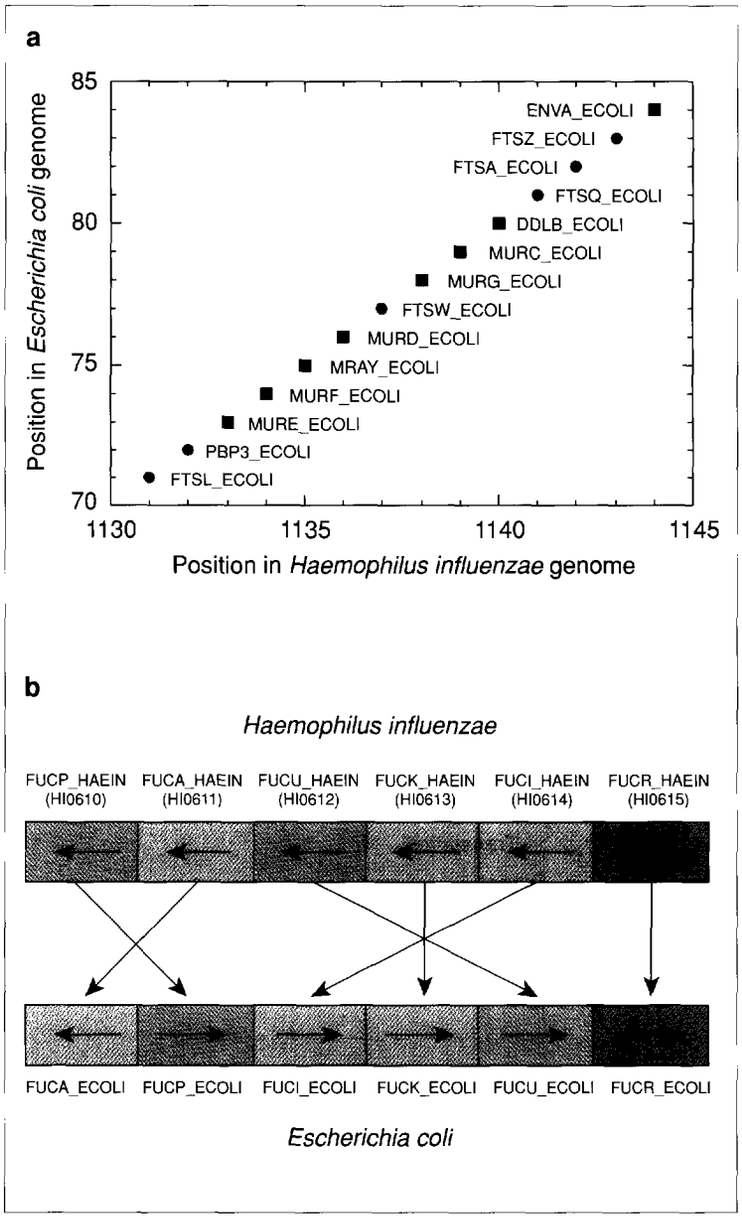


Figure 6

(**a**) An illustration of the conserved *dcw* cluster present in *Escherichia coli* and *Haemophilus influenzae*[22], containing a number of functionally related genes. The direction of transcription is colinear. Genes are labeled according to their name in *E. coli*, and markers represent functional classifications: red circles are used to denote genes encoding proteins involved in signaling, blue squares are used to denote genes encoding proteins with structural functions. Axes display the absolute position of genes in the *H. influenzae* genome paired with their orthologues from *E. coli*. This example demonstrates the patterns that emerge at this level of sequence comparison. (**b**) The conserved fucose (*fuc*) operon. While the genes in the *H. influenzae fuc* operon (top) are all transcribed in the same orientation, most genes in the *E. coli fuc* operon (bottom) are transcribed in the opposite orientation, with the exception of the first gene. Some rearrangements are also visible. Genes encoding transport proteins are colored orange, genes encoding proteins involved in metabolism are colored green, and genes encoding proteins involved in signaling are colored blue.

Conserved clusters of functionally related genes

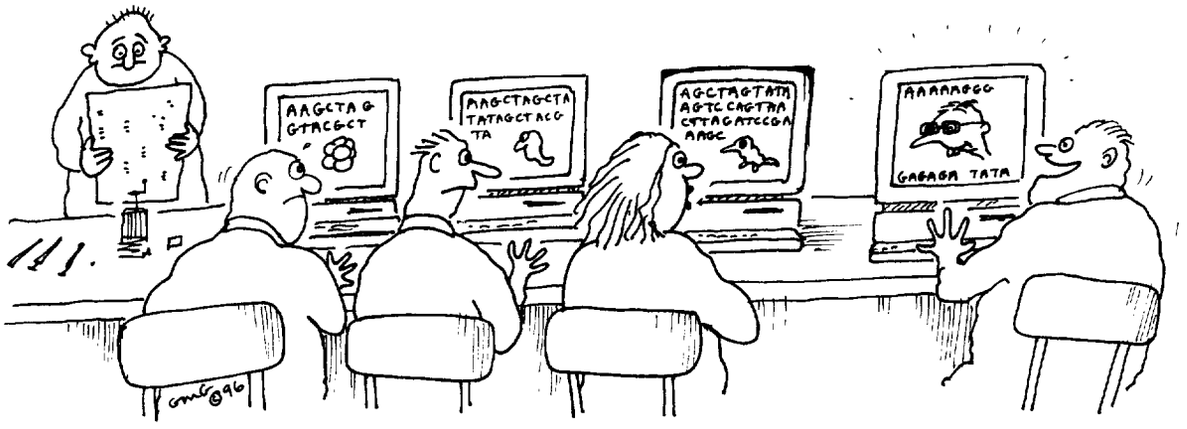
Is chromosome organization conserved in related species? This question, which was first raised in relation to genomic maps of plastids[14,15] and bacteria[16–19], was investigated using only homologous sequences. However, using the comprehensive functional classification described above, we can ask whether functionally related, but not necessarily homologous, genes of any two genomes are in a similar sequential order along chromosomes.

To facilitate this, we have compiled data from intra- and inter-genomic comparisons of *E. coli* and *H. influenzae*. All *E. coli* sequences were classified into the nine functional classes, using annotation from the sequence databases, in conjunction with keywords – a functional class dictionary was created. The relative locations of genes (not their absolute order) were taken from genome repositories, and genes for which there is no accurately mapped location were omitted. The positions of genes, which were deduced from the order of the genes along the chromosome, were used for comparing the two genomes[14,20]. The numbers of gene pairs are compared with the results from a random model[21], which estimates the probability of two functionally related genes being adjacent on the genome, in order to establish the statistical significance of these observations.

The main conclusion is that clusters of genes belonging to the same functional classes (that do otherwise not share sequence similarity) have patterns of aggregation along chromosomes that appear to be statistically significant[22]. For example, in *H. influenzae*, there is a clear aggregation of genes encoding metabolism- and translation-associated proteins (Fig. 5). In addition, patterns of linear conservation between genomes are observed (Fig. 6). Two examples of genome-order comparison in regions of the *H. influenzae* and *E. coli* genomes are presented. The first example illustrates the case for the *dcw* cluster in the 2′ region on the *E. coli* chromosome. In this region, all sequences belong to the signal or structural classes, and the order is preserved (Fig. 6a). The second example is that of the *fuc* operon, at 63′ in *E. coli* and 35′ in *H. influenzae*: together with *fucA*, *fucO* (upstream) is transcribed in the negative orientation with respect to the origin in *E. coli*, while *fucP* (adjacent to *fucA*) and the downstream genes are transcribed in the positive orientation; in *H. influenzae*, all genes are transcribed in the opposite direction to their direction of transcription in *E. coli* (Fig. 6b). Not all of the clusters are operons, and these clusters may represent newly identified regulatory units. In addition, it may be possible to predict the functional class of unknown genes and formulate a hypothesis for the cellular processes with which they are associated, further guiding experimental analysis.

Global views

The approaches described above are posing new questions for the computational comparison of model genomes, and are helping us to obtain different global views of complete genome data. Having accurate estimates for the expected sequence similarity between orthologues in different species will be the basis for more-reliable function predictions. Automatic annotation and classification of sequences, possibly based



'It's Wilkinson!'

on alternative classification schemes, can supply global views of complete genomes. Finally, gene-order comparisons contribute towards the understanding of chromosome structure and evolution, complementing traditional sequence comparison.

Acknowledgements

We thank our colleagues in the GeneQuiz consortium, especially Reinhard Schneider and Antoine de Daruvar, for their valuable contributions to our continuing efforts in genome analysis. C.O. is a recipient of a long-term fellowship from the Human Frontiers Science Program Organization.

References

1 Fleischmann, R. D. *et al.* (1995) *Science* 269, 496–512
2 Fraser, F. C. *et al.* (1995) *Science* 270, 397–403
3 Schorf, M. *et al.* (1994) in *Intelligent Systems for Molecular Biology 1994* (Altman, R., Brutlag, D., Karp, P. D., Lathrop, R. and Searls, D., eds), pp. 348–353, AAAI Press
4 Casari, G. *et al.* (1995) *Nature* 376, 647–648
5 Ouzounis, C., Casari, G., Valencia, A. and Sander, C. (1996) *Mol. Microbiol.* 20, 897–899
6 Casari, G., Ouzounis, C., Valencia, A. and Sander, C. (1996) in *First Annual Pacific Symposium on Biocomputing* (Hunter, L. and Klein, T. E., eds), pp. 707–709. World Scientific
7 Fitch, W. M. (1970) *Syst. Zool.* 19, 99–106
8 Bork, P. *et al.* (1995) *Mol. Microbiol.* 16, 955–967
9 Karp, P. D., Ouzounis, C. and Paley, S. (1996) in *Intelligent Systems for Molecular Biology 1996* (States, D., Gaasterland, T. and Smith, R., eds), pp. 116–124, AAAI Press
10 Ouzounis, C., Valencia, A., Tamames, J., Bork, P. and Sander, C. (1995) in *European Conference on Artificial Life 1995 (ECAL95)* (Morán, F., Moreno, A., Merelo, J. J. and Chacón, P., eds), pp. 843–851, Springer-Verlag
11 Adams, M. D., Kerlavage, A. R., Fields, C. and Venter, J. C. (1993) *Nat. Genet.* 4, 256–267
12 Tamames, J., Ouzounis, C., Sander, C. and Valencia, A. *FEBS Lett.* (in press)
13 Ouzounis, C., Bork, P., Casari, G. and Sander, C. (1995) *Protein Sci.* 4, 2424–2428
14 Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F. and Cedergren, R. (1992) *Proc. Natl Acad. Sci. USA* 89, 6575–6579
15 Boudreau, E., Otis, C. and Turmel, M. (1994) *Plant Mol. Biol.* 24, 585–602
16 Taylor, D. E., Eaton, M., Yan, W. and Chang, N. (1992) *J. Bacteriol.* 174, 2332–2337
17 Kumisawa, T. (1995) *J. Mol. Evol.* 40, 585–593
18 Liu, S. L. and Sanderson, K. E. (1995) *Proc. Natl Acad. Sci. USA* 92, 1018–1022
19 Lopez-Garcia, P., St Jean, A., Amils, R. and Charlebois, R. L. (1995) *J. Bacteriol.* 177, 1405–1408
20 Tatusov, R. L. *et al.* (1996) *Curr. Biol.* 6, 279–291
21 Karlin, S. and Ladunga, I. (1994) *Proc. Natl Acad. Sci. USA* 91, 12832–12836
22 Tamames, J., Ouzounis, C., Casari, G. and Valencia, A. *J. Mol. Evol.* (in press)
23 Riley, M. (1993) *Microbiol. Rev.* 57, 862–952