F. Rechenmann and C. Gautier, "Interpreting the Genome", *La Recherche*, n° 332, June 2000, pp. 39-45

# Interpreting the Genome

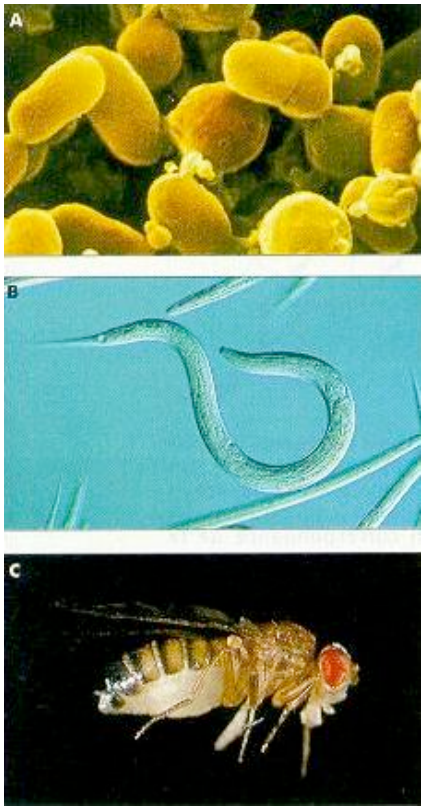
**by François Rechenmann and Christian Gautier**

*How can we interpret the billions of bases in the human genome? We can locate the genes of bacteria using bioinformatics, but for more complex organisms the error rate is close to 50%. Discovering their function is another matter altogether.*

When the first complete genome of a living organism, the bacterium *Haemophilus influenzae*, was sequenced in July 1995, there were mixed reactions from the biological research community. (1) Some welcomed it as a major event which opened up radically new avenues in the study of life, while others saw it as at best a purely technological and economic exploit, which distracted the attention of decision-makers and the public away from the real concerns of research. Five years later, now that the genome sequences of several dozen bacteria and three eukaryotic organisms – the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans* and the fruit-fly *Drosophila melanogaster* – have been obtained and published, and a draft version of the human genome sequence has been announced, there are still the same differences of opinion. (2,3)

**François Rechenmann** is director of research at INRIA Rhône-Alpes, France (the *Institut national de recherche en informatique et en automatique*).
**Christian Gautier** teaches at the *Université Claude-Bernard* in Lyon, France

The strategy employed is both systematic and exploratory, and it is this two-pronged approach which has prompted the debate. The expression “post-genome”, misleading in more than one respect, is often used to mark the end of a period of blind experimentation and to welcome the return to a hypothetico-deductive type of approach. It gives the impression that since the beginning of the 1970s, sequences have simply been collected, without any information about the function and evolution of living systems being learnt from them. In fact, the availability of sequences merely marks the beginning of the long and difficult job of analysing the data, frequently interrupted by a return to experimentation and even to new sequencing. How should this mass of information be used, to turn it into biological knowledge? A sequence has a formal structure and lends itself naturally to analysis by computer. What roles should computational analysis of genomic data and experimental approaches play?




The genomes of several dozen organisms have been sequenced to date. *Haemophilus influenzae* was the first bacterium (A), and the nematode *Caenorhabditis elegans* (B) was the first multicellular organism. The fruit-fly *Drosophila melanogaster* (C) is the most complex organism whose sequence has been published.

 A, CNRI; B, Cosmos; C, CNRI

Sequencing a genome means finding out the sequence of nucleotides which make up the DNA macromolecule. Each nucleotide is referred to by the first letter of the name of its nitrogen base, and the information carried by the genome is contained in the long text – nearly 4 billion characters for the human genome – written in an alphabet of these four letters, A, C, G, and T. The efficiency of a sequencing project is measured in kilobases per day, and is determined by the number of machines used. As these can only sequence relatively short sections of the DNA molecule at a time, powerful computers have to be used to put the partially overlapping sub-sequences obtained into the correct order, to reconstitute the complete genome sequence. As well as point mutations, where one base is missing or has been replaced by a different one, there may be errors in the order of the sub-sequences. What is more, certain parts of the DNA molecule are more difficult to sequence, and obtaining a sequence covering 100% of a genome is particularly expensive. This is why the part of the human genome which is now available, or the part which should soon be, is called a “working draft” It is thought that it will be at least two years before a complete, high quality sequence is available.

It will take at least two years to go from the working draft of the human genome to a complete, high-quality sequence.

The Human Genome has almost 4 billion base pairs, spread out over 23 chromosomes. The invention of the sequencing gel (background) was the most important advance in the manual analysis of DNA sequences. J.J.P./ Eurelios

Part of the sequences is deposited in databanks which are freely accessible via the Internet. Three banks – EMBL in Europe (maintained by the European Bioinformatics Institute (EBI) at Hinxton near Cambridge), GenBank (maintained by the National Center for Biotechnology Information (NCBI) in the United States), and the DNA Data Bank of Japan (DDBJ) in Japan share their data, and in practice form a single bank with three entry points. GenBank’s February 2000 version holds 5.7 million sequences, a total of 5.8 billion nucleotides long, and the size of the bank now doubles every seven months, at a rate of 15 million new bases per day. It is obviously impossible to put a figure on the very large number of sequences not held in these banks, for confidentiality reasons related to the economic interests at stake. The human genome sequence which Craig Venter and his firm Celera say they have completed is not yet accessible either for the time being, but it should be soon - publication in the scientific journals is expected at the end of 2000.

**Each sequence has attached to it various information called “annotations”.** This naturally includes the source organism, but also, where some of the genes have been identified experimentally or by computational analysis, a brief description of their function, as well as bibliographical links. One good thing about these banks is that they bring together all the publicly available sequences, but they do have several shortcomings. The quality of the sequences varies, and some of the data are redundant – there may be several copies of the same section of the genome of a given organism, sequenced and deposited by different laboratories. There is little logical structure to the annotations, so it is difficult to interpret them by computer, and these too are of very variable quality. Because of this, a number of specialised databases are growing up

parallel to these banks. Some bring together sequences which relate to the same organism, for example SubtiList and NRSub for the bacterium *Bacillus subtilis*, Cyanobase for the bacterium *Synechocystis*, TAIR for the plant *Arabidopsis thaliana*. Others group together complementary annotations, cutting across various different sequence databases. This is the case with FlyBase, for the drosophila, MGD (Mouse Genome Database) for the mouse and GDB (Genome Data Base) for the human genome. Others concentrate on a particular class of sequences, but for a group of organisms. The Eukaryotic Promoter Database (EPD) brings together sequences for promoters* from eukaryotic organisms. Finally, there are several databases devoted to proteins. SwissProt in Geneva is maintained by the group led by Amos Bairoch, in collaboration with the EBI, and contains more than 80 000 sequences relating to several hundred different organisms. Access to all these data on the Web has significantly changed biologists’ research strategies.

* A **promoter** is a region upstream from the coding region, where the RNA polymerase – the enzyme responsible for transcribing DNA into RNA - binds to the DNA strand.

Each database addresses different biological questions, and this shapes the way the data are structured within them. They thus each have a different conceptual plan, so hoping to organise all the genomic data – the sequences and the various other data which are attached to them – within a single database is a lost cause. On the other hand, their integration does need to be improved; in other words it should be made easier to search these different bases at the same time, in response to a complex request from a biologist who has his own method of approaching a problem. This as much a conceptual issue as a

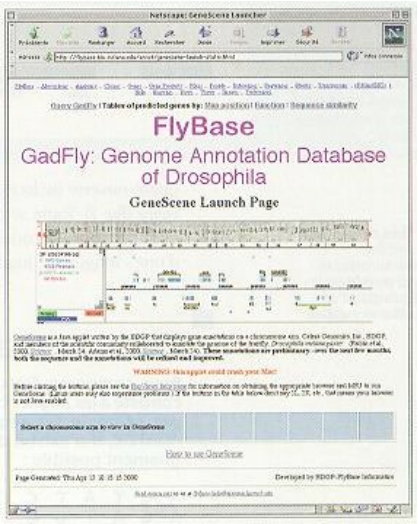
technical one. How can different databases be reconciled, when their structure is based on different definitions, (all too often in a way which is not even explicit), especially definitions of such fundamental concepts as the genes themselves? Some databases consider the gene to be limited to those regions of DNA which code for its product or products (protein or RNA) while for others it includes the various regions which come into play during transcription (from DNA into RNA) and translation (from RNA into proteins), that is, a large number of regulatory sequences.

How can databases be reconciled when they are based on different definitions of a term as fundamental as 'the gene'?

**Remember that the term 'genome' is not without ambiguity either**. Generally, it refers to the DNA macromolecule contained in the chromosomes, but there is also non-chromosomal DNA, in the plasmids of bacteria and the organelles (mitochondria or chloroplasts for example) of eukaryotic organisms. The term also applies to the whole set of genes of an organism. To return to sequencing, if we use the classic metaphor in which the DNA bases are seen as letters, then once the text (the sequence) has been obtained, the first difficulty is to identify the words (the genes) which make it up. Next comes the question of the meaning – the function of the genes.

A biologist's first reflex, when a new sequence is available, is to compare it, together with its potential translations into protein sequences, with those already held in banks and databases, looking for similar rather than identical sequences. With the exception of sequencing errors, any differences represent mutations which have accumulated in the

course of evolution. If there is enough similarity, the two fragments are considered to result from divergent evolution from one ancestral fragment, and they are said to be homologues. If the fragment includes a gene, homology suggests that the proteins it codes for have a similar function, but it does not prove this, as will be seen later. The search for similarity has led to a wealth of technical and methodological developments, both to shorten the computer run time, when a sequence is compared to all the sequences that are already known, and also to take prior knowledge about evolutionary mechanisms into account when designing algorithms. There are limits to what this strategy can achieve. A similarity search may fail simply because no homologous sequence has yet been identified. When the yeast genome was sequenced, almost half its genes were completely unknown, and they did not resemble anything found in the banks. Such genes are known as “orphans”. Besides, relying exclusively on the information in the databanks means that if this information is incorrect, as is all too often the case, the errors are propagated, resulting in what some researchers call a “house of cards”. So it is essential to have access to direct gene identification methods which do not rely on

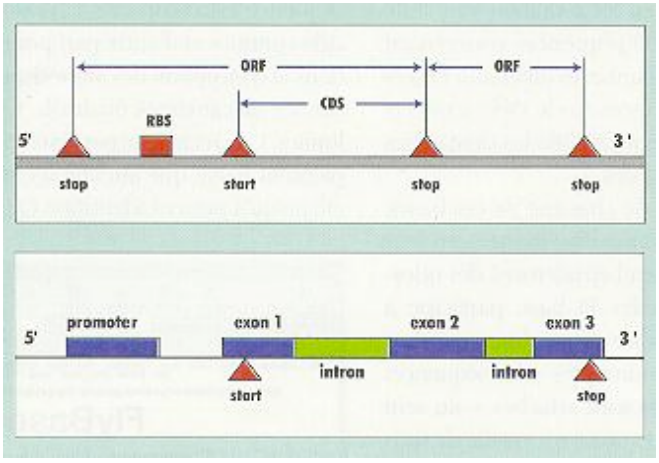



Database interrogation requires increasingly complex software tools (this is GadFly). At present there is no universal protocol.

homology. This research is much easier when the genome in question comes from a prokaryote (a bacterium) than if it comes from a eukaryote (any other organism).

**A prokaryotic genome is fairly dense** – almost the entire sequence corresponds to genes – and we know the codons (sets of three nucleotides) which mark the beginning and end of translation of a region which codes for a protein. But unfortunately it is not that simple, as there are certain ambiguities: for example, the codons which mark the beginning of translation also code for an amino acid. ATG, the most common start codon, codes for methionine. So there is only one possible “necessary condition” defining where to look for a coding sequence: between two codons which mark the end of translation (known as STOP codons), in what is called an Open Reading Frame (ORF) (Fig. 1).

**Figure 1**. An open reading frame (ORF) is the region between two STOP codons. Within this, a coding sequence (CDS) begins with a START codon, and is preceded by a ribosome binding site (RBS). In eukaryotes, a gene is interrupted by non-coding sequences called introns. The exons are the parts which are translated into proteins.



Any sequence included in an ORF which begins with a START codon and which is judged to be long enough (for example 300 nucleotides for a prokaryote, which corresponds to a protein of 100 amino acids) is considered to be a potential coding region. If significant sub-sequences, particularly a promoter or a ribosome* binding site, are found upstream from this region, this supports the hypotheses, as does the existence of similar sequences in the nucleotide and protein bases. Finally, the same sequence can be “read” in three different ways, grouping the letters in threes, codon by codon, and each of the two complementary strands of DNA can be read, so that in practice the search for coding regions must be carried out on six different virtual sequences. Together with Antoine Danchin’s group at the Institut Pasteur, the authors have developed software tools to facilitate genome analysis, but there are many others. (fig. 2) [4]

***Ribosomes** are macromolecules which allow the translation of messenger RNA into protein to take place.

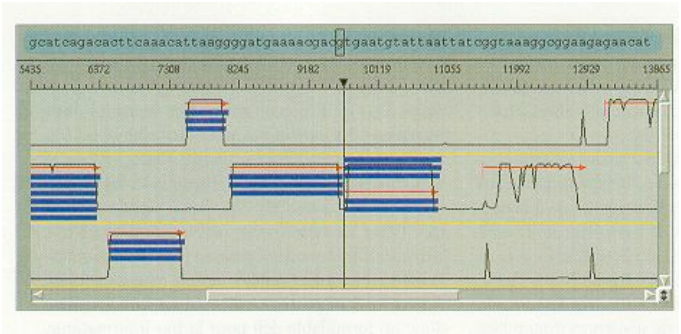
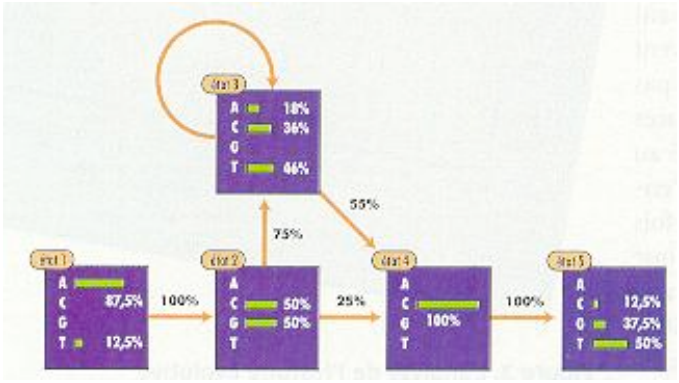



**Figure 2**. Gene location software often combines several different methods. The first method, represented by the arrows, looks for coding sequences preceded by ribosome binding sites (see figure 1). The second looks for similar sequences in the database, represented by blue rectangles. The third uses a Markov model - a sudden variation in the black trace indicates a coding region.

In eukaryotic organisms, the situation is a great deal more complicated, because the coding regions represent only a small percentage of the total genome sequence (3 to 5 % in mammals), mostly because a eukaryotic gene is made up of several coding regions called exons, separated by non-coding regions called introns (fig.1). So the strategy used for bacteria does not work, and in order to identify the coding sequences we have to turn to other properties of genes, which are less strictly defined and thus less efficient. Firstly, the fact that a sequence codes for a protein imposes constraints which make bases more likely to appear in certain orders than in others. Secondly, the cellular machinery recognises the boundaries between exons and introns thanks to particular arrangements of consecutive bases, which the software may learn from known examples. Of the mathematical tools currently available, Markov models seem to manage these two sorts of information most efficiently (see inset “Markov models”.) But there are many others. As none of them is completely satisfactory, it is advisable to combine the results of several complementary or even rival methods. It is only thanks to this strategy that it is becoming possible to make a reasonably accurate prediction of a complete gene (ie the succession of introns and exons) and then to reconstruct the coded protein or proteins, as well as the various regions involved in transcription and translation.

See following pages for insets
“Markov models” and
“Aligning two sequences.”

## Markov models




The transcription and translation machinery which produces proteins from genes is able to recognise particular sequences, such as those that mark the boundary between exons and introns. These sequences are not always the same, but they do share a “family likeness”. They are called consensus sequences. How can a program recognise whether or not a given sequence is likely to belong to a family? Most of the software techniques use Markov models. These are based on successions of states, linked by transition probabilities. Each state is itself described in terms of probability. In the case of DNA sequences, a state represents a position in a sequence. For example, in the first position there will be an A in 85.7% of cases and a T in 12.5% of cases. This might be followed in 100% of cases by a state 2, defined by other probabilities of finding certain bases, and so on. Calculating the values of these probabilities is the aim of the first phase in using a Markov model: the training phase. A set of sequences known to belong to a given family is entered into the computer so that it can identify the frequently occurring forms. Next, in the recognition phase, the model is used to determine the degree of similarity between a given sequence and the ones in its memory. The probability of a particular sequence corresponding to the model is the product of the probabilities of occurrence within each state, and the transition probabilities. It is compared with that obtained for a random sequence of the same length. Above a certain threshold, it indicates whether the sequence belongs to the family required, for example an intron-exon boundary.

## Aligning two sequences.

The simplest way to compare two sequences is to line them up side by side. If they are the same length, for example the nucleic acid sequences AGTATC and AGATGC, then in the simplest case there is only one possible alignment:

A G T A T C

A G A T G C

It is a different matter if insertions or deletions may have occurred, for example:

A G T A T – C  or alternatively  A G – T A T C

A G – A T G C  A G A T – G C

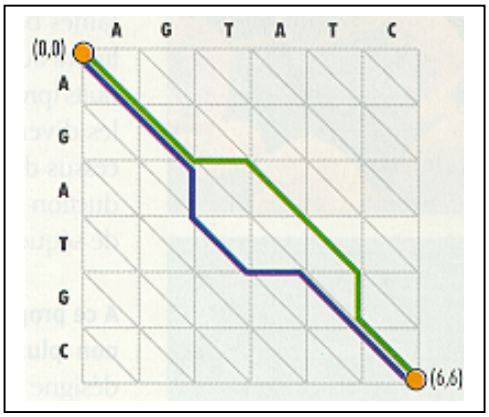
A basic mark is then given to each pair of bases: 2 if the top letter is the same as the bottom letter, 0 if they do not correspond, and –1 where a letter corresponds with a gap. The choice of basic marks is obviously important and stems from biological considerations. The basic marks of an alignment are added up to give a score: for the alignments shown above the scores are 6, 8 and 6 respectively. The second one is thus the best of the three.

What is the best possible score for an alignment? For two sequences six bases long, there are already 924 possibilities, and more than two and a half million for sequences 12 bases long. It is thus impossible in practice to try out all the possibilities.

The principles of the technique used, called dynamic programming, are illustrated in the diagram overleaf. The possible alignments for two sequences n bases long are represented as paths leading from the initial node (0,0) to the final node (n,n), on a grid where each column corresponds to a letter of the first sequence and each row corresponds

to a letter of the second sequence. Starting from the initial node, the score of the partial optimal alignment terminating at each node (i,j) is calculated one step at a time. The optimal alignment is then identified by working backwards from the final node, and selecting at each step the direction which leads back to the highest-scoring previous node, until the initial node is reached.

This algorithm can also be used for sequences of amino acids, but the calculation is more complicated as there are twenty different variables, not four as with nucleic acids. To work out the scores, a 20x20 matrix, called a substitution matrix, is used, which gives the different “costs” of substituting one amino acid for another. The values of the elements of this matrix represent biochemical and evolutionary considerations. For example, the replacement of a hydrophobic amino acid with another hydrophobic amino acid carries a lower penalty than its replacement with a hydrophilic amino acid. It has to be said that this basic algorithm still has a long run time. For the sake of efficiency, the most frequently used programs such as BLAST or FASTA, use heuristics which run even faster but which do not guarantee to identify an optimal alignment.




An alignment is a path leading from the node (0,0) to the node (6,6). A diagonal trace between the node (i-1, j-1) and the node (i,j) shows that the i letter (ie from the top row of the sequences compared) and the j letter (from the bottom row) are the same. A horizontal line on the graph indicates that a letter on the top row corresponds to a gap in the bottom row; a vertical line indicates a correspondence between a gap on the top row and a letter on the bottom row. A straight diagonal line across the grid thus represents an alignment without insertions. The blue path corresponds to the second alignment and the green path to the third one.

**Using these advanced research strategies produces fairly reliable results** in prokaryotic genome analysis, but there is still a long way to go for eukaryotic genomes. How can we be sure that the computer predictions are correct? Computational data (*in silico* or ‘dry lab’) must be compared against biological data (*in vitro* and *in vivo* or ‘wet lab’). For example, when a gene is expressed it is transcribed into RNA before being translated into proteins. This RNA can be recovered and sequenced. It does not contain introns, and can be compared to the genome sequences. Jean Thierry-Mieg, who took part in sequencing the nematode *C. elegans*, has shown that about 50% of the predictions were wrong, sometimes significantly so. (5) It also appears that rather than the 18 000 genes originally predicted, there are only in fact 12 000. An error rate of 50% was also found for one of the very first prokaryotes to be sequenced, *Mycoplasma pneumoniae*, even though the process should theoretically be simpler, as we have seen. This last figure takes into account errors in gene function attribution. (6)

Clearly the difficulties are not over once a gene is discovered. Its function or functions still have to be discovered. Structural homology suggests functional homology, so the strategy is based on a database search for genes with a similar sequence. But this method also has its limits. Once a certain similarity has been identified, genes which are orthologues must be distinguished from those which are paralogues. What does this difference mean in real terms? It is quite common for some genes to duplicate themselves. While the original copy of the gene generally retains

Structural homology does not necessarily mean functional homology. We need to know how the gene has evolved as well.

its original function (this is a true homologue, hence its name of ‘orthologue’), the duplicate or duplicates (paralogues) may evolve independently and acquire completely different functions. These two cases can only be distinguished through an evolutionary analysis, by constructing phylogenetic trees.

The first step is to “align” the sequences of homologous genes, that is, to estimate what mutations have appeared during their divergent evolution from a common ancestor. If only two sequences are available, a dynamic-programming algorithm is used (see inset “Aligning two sequences”). Where large numbers of sequences are available, as is the case with certain genes coding for ribosomal RNA, higher-speed heuristics have to be used, but these are not guaranteed to find an optimal alignment. After deciding on an evolutionary model, it is

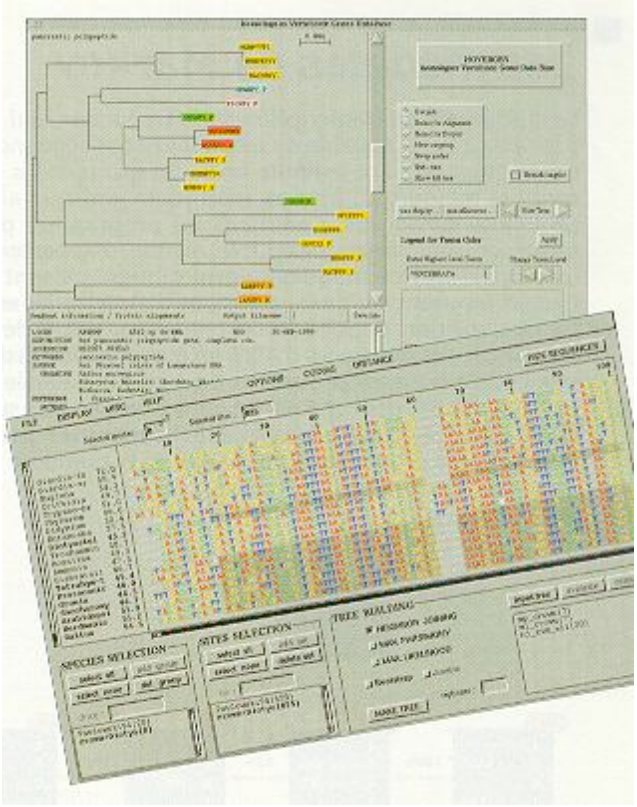



**Figure 3**: The analysis of a gene family’s evolutionary history begins by lining up their sequences (bottom) . Their phylogenetic tree can be drawn up after estimating the number of mutations necessary to change from one gene to another. (top)

normally possible to differentiate between paralogues and orthologues by estimating the total number of changes along the branches of the phylogenetic tree linking each pair of sequences. However it is impossible to validate the resulting tree experimentally. At best it can be checked against prior knowledge from the field of systematics * (see figure 3).

***Systematics** involves studying and describing the diversity of organisms, investigating the nature and causes of the differences and similarities between them, demonstrating family relationships between them, and developing a classification system which reflects these relationships”.

Translated from L. Matile, P. Tassy, D. Goujet, ‘Introduction à la systématique zoologique’ in *Biosystéma N°. 1* SFS Editions, Paris, 1987

Another problem in the quest for gene functions stems from the fact that the merging of fragments originating from different genes allows totally new functions to emerge. This is what François Jacob meant by “evolutionary tinkering”. In addition to these problems linked to the way living systems function, there are others which arise from the fact that the available sequence databases are incomplete and contain errors. For these reasons, the results produced by software are no more than hypotheses, which must in turn be experimentally tested in the laboratory, in particular by observing the effects of the substitution or deletion of a gene in the organism, or in one related to it. This is why the priority given to the human genome has sometimes been criticised. Some think it would be better to begin by sequencing and analysing the mouse genome, which has large numbers of genes homologous with human genes, and which can be experimented on, rather than to tackle the human genome straight away, with the risk of accumulating

hypotheses which cannot be validated in the short term. Whatever the answer, given the inadequacy of a purely computational approach, determining the function of genes (or rather of the proteins they code for) is now a matter for the experts. As soon as the drosophila genome had been sequenced, Craig Venter hosted what he called a “jamboree” for forty-five of the world’s top specialists in fly genetics, bio-informatics and proteins, where they spent eleven days comparing their opinions on the raw sequence he had just obtained in collaboration with more than thirty teams around the world. It was only after this brain-storming session that an annotated sequence was submitted to the rest of the scientific community, and published in the journal *Science*. Clearly, systematising this “annotation” process is a considerable challenge for bioinformatics. Once we think we have identified the sequence of a gene, what is the best way to fit together data and knowledge of various kinds and various origins, relating to several organisms, in order to predict the functions of that gene?

Annotation is still a matter for the experts.
45 specialists held an 11 day meeting on the drosophila sequence.

**In the “anything goes” strategy**, one key element is the way data and information are structured within computer systems, whose powerful capabilities allow the researcher to search and browse, to visualise data from a different perspective, and thus to draw new inferences. Although it is easy to store basic data such as sequences, the computational representation of data about functions, for example those which relate to metabolic pathways, is still a problem for bioinformatics research. A look at the KEGG database

will confirm this – here, the data are only presented as images, available “at a click”, certainly, but impossible to process using software.

The function of certain proteins is to interact with the “regulatory regions”, generally found upstream from the genes, and to switch those genes on or off – in other words to regulate their expression. Directly or indirectly, the products of these genes are then likely to have an effect on the expression of other genes. This creates networks of molecular interactions which adapt protein production to the cell’s needs in a given context. A knowledge of these networks is of great importance, because it could explain cell specialisation within a multicellular organism, and more generally, its development and morphogenesis.

At present, these interactions are described in the specialist literature, and less often in databases, in the form of text, diagrams or graphs. (see fig 4.) But it will only be possible to check how consistent they are, to compare them against

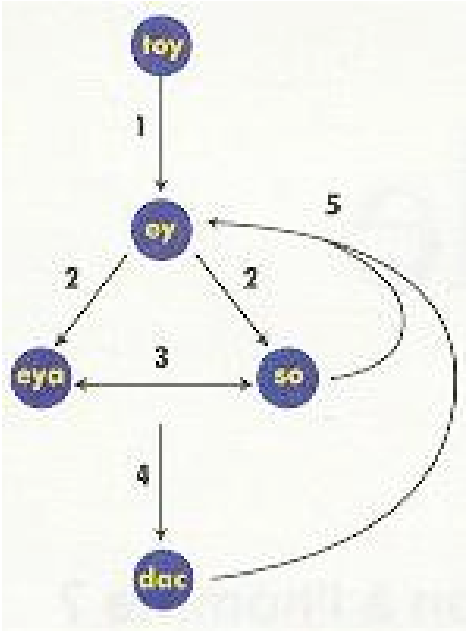



**Figure 4**. Genetic control over the formation of the eye is based on a gene cascade. Toy activates ey (eyeless) which can, together with dac (dachshund) activate itself (the arrows represent activation). These interactions are still only represented in the databases to a very limited extent.

gene expression data, or to simulate their function, if they exist in an computer-compatible form. Unfortunately, not enough is known at present about these interaction networks to allow them to be modelled in detail, for example in the form of systems of differential equations whose variables would be a function of the concentrations of different products. Only basic models are possible. In the simplest form, a network can be represented as a set of Boolean variables, with a value of 1 or 0 according to whether the corresponding gene is expressed or not, and a set of transitions between the values of these variables. Despite its simplicity, a model of this kind, which has more sophisticated variants, is capable of exploring the dynamics of interaction networks, for example to predict the existence of feedback circuits or steady states. This allows the analysis and simulation of well-defined systems such as the network of ten genes involved in flower formation in the plant *A. thaliana.* (7)

Other models are currently being developed, with the aim of producing more realistic behaviour patterns by including graduated information about reactions, of the type “the more gene A is expressed, the less gene B is expressed”. At the same time, technological progress in “DNA chips” heralds the availability of vast amounts of gene expression data. Thanks to these chips, it will be possible to work out the structure of underlying interaction networks, although this will need the help of methods of analysis which have not yet been devised. Within the same timescale, there are already ambitious projects aiming to create models linking the genomic and metabolic levels. (8)

Far from being an end in itself, the availability of a complete genome sequence opens up the possibility of a systematic approach to the genes within it. But progress is still needed. The human genome sequence is now available, and the mouse genome soon will be, but making reliable predictions of eukaryotic genes on the basis of those sequences is a classic example of a problem which is still wide open.

It is the extreme variety of the information available, and way in which it is interrelated, which causes the problem, rather than its volume. In fact, improvements in the efficiency of comparative techniques are keeping pace with the availability of sequences of new organisms. But to reap the benefit of this multiplier effect, whereby new information is produced on the basis of existing information and the analysis of new data, it is no longer good enough to record this information only in textual form and in natural language, even if it is stored in IT format. This form is an obstacle to wide-ranging, integrated searches of large numbers of databases, even when powerful search engines are used. The key issues in bioinformatics research are therefore not only to design new, increasingly powerful and above all appropriate algorithms or heuristics, but also to provide tools which will make it easier to model, structure, examine and visualise biological knowledge. **FR and CG ○**

**Further reading:**

- Académie des Sciences, *Rapports sur la science et la technique, No. 1* "Développement et applications de la génomique", Tec et Doc, 1999
- Antoine Danchin, *La Barque de Delphes. Ce que révèle le texte des génomes,* Odile Jacob, 1998
- Alain Bernot, *L'Analyse des génomes. Cartographie, séquençage, identification des gènes*, Nathan Université, 1966
- Steven Salzberg, David Searls, Simon Kasif, *Computational Methods in Molecular Biology*, Elsevier 1998
- Pierre Baldi and Søren Brunak, *Bioinformatics. The Machine Learning Approach*, The MIT Press, 1998.

**References**

[1] Fleischmann, RD. *et al*; *Science 269*, 496, 1995

[2] 'The *C. elegans* sequencing consortium'; *Science, 282*, 2012, 1998

[3] Adams, M.D. *et al*; *Science,287*, 2185, 2000

[4] Médigue, C. *et al*; *Bioinformatics vol 15*, *N° 1*, p2, 1999

[5] Thierry-Mieg, J. *et al;* Programme Génome/CNRS, http://www.cnrs.fr/SDV/mieg.html

[6] Venter, C. & Bork, P. Conference papers, Pasteur Institute conference *Génome 2000*

[7] Mendoza, L. *et al*; *Bioinformatics vol 15, N°7/8*, p593, 1999

[8] Tomita, M. *et al*; *Bioinformatics vol 15*, *N° 1*, p72, 1999