BIOINFORMATICS

Vol. 16 no. 1 2000
Pages 65–75

A brief history of genome research and bioinformatics in France

Antoine Danchin

Regulation of Gene Expression, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

Abstract

The development of in silico *genomics has progressed slowly in France for a number of political reasons. Two administrative organizations, the Groupement de Recherche sur les Génomes (GREG) and the Groupement de Recherche 1029 (GDR 1029) of the Centre National de la Recherche Scientifique (CNRS) have been established. These organizations have created the dynamics that hopefully will place France (which coordinated consortia that completed several of the first large microbial genomes) among the developed nations that support Large-Scale Biology.*

Contact: *adanchin@pasteur.fr*

Introduction

The development of scientific knowledge in many countries is deeply rooted in politics and economics. It is usual to assess the scientific production of a country using publication record alone (May, with appropriate caveats from Adams, 1998; Barreto *et al.*, 1997; Kotiaho, 1998; May, 1997a,b; Paris *et al.*, 1998; Price, 1998; Schoonbaert and Roelants, 1998). It is also important to consider the political structure and the economic situation of a country when assessing the impact and outcome of scientific programmes. In France during the last 25 years, biology, and especially genetics, has not been well supported compared with other sciences, for epistemological reasons (Burian, 1990; Burian *et al.*, 1988; Burian and Zallen, 1992). If comparison is made with the USA, the amount spent per scientist in biology is about three times that of France (Allègre, 1998), which possibly explains why only a few places in France are famous for biology. This does not require that famous scientists are particularly brilliant, but simply that they have been able to secure regular funding at a level comparable to that in the USA. In France this situation had an important consequence: the scientific judgement of the politically effective scientists led to an unbalanced impact on the scientific choices made. This is the first historical 'rule' in understanding the historical intricacies of the development of new fields such as 'genomics' and 'bioinformatics' in France. The history presented in this paper relies mainly on information that differs from the usual reference texts used in scientific articles. Many are reports, reviewing comments, white papers, etc., that have not been widely circulated. For this reason some of the historical information can only be alluded to, it is only partial and some names are not disclosed. I shall, nevertheless, try to make the situation understandable for scientists (or epistemologists or historians of science), who are not always aware of the arcane functioning of political and administrative systems.

The political situation had a considerable effect on the development of genomics in France. This effect was associated with the general organization of political power. In French politics, the interactions between the democratic processes and the exercise of power may not always be clear, but they can be of vital importance for the explicit development of the political choices made by the citizens. Ministries are not fixed entities, and in the domain of education and research there has been many changes in the funding of science following political change. Research and technology usually go together under one ministry; however, they may stand alone, be associated with the ministry of industry, even the ministry of education, or simply the universities. When politics separates 'research and technology' from 'education' (this happened quite often during the time when bioinformatics was created as a discipline placed at the core of genomics), this creates *two* independent places where research can be organized. Research carried out by civil servants is controlled by the Ministry of Research and Technology, while research performed by professors and lecturers in universities (civil servants also) is controlled by the Ministry of Education. The evaluation of the quality of research is mostly performed by Institutions ('Etablissements Publics à caractère Scientifique et Technique', EPSTs) such as the Centre National de la Recherche Scientifique (CNRS, research institution), the Institut National de la Santé et de la Recherche Médicale (INSERM, mainly medical research), the Institut National de la Recherche Agronomique (INRA, mainly agronomy research and linked to the Ministry of Agriculture), etc. It is obvious that, unless there is active coordination at the highest level (e.g. by the Prime Minister), the policies of the different institutions may be divergent, and this

gives room for personal lobbying and a variety of short cuts. The latter can allow those who are familiar either with the arcana of the system, or simply linked by any type of relationship [i.e. former students among the so-called 'Grandes Ecoles', or simply family relationships (Romains, 1925–1934)], to follow their own personal policy rather than that dictated by the democratic process.

A worse problem then follows, that of *organisation*. When, after an election or after a parliamentary vote and a new minister takes charge of his or her ministry, the first thing to be done is to organize the administration of the ministry. To do this, the minister has to go to the EPSTs he or she controls (where he or she has either confirmed or nominated new directors for each of these institutions), or to the presidents of universities, to ask them to allow some of their personnel to work in the ministry as advisors, chargés de mission or whatever. The usual reaction of these directors is to be extremely reluctant. However, if one understands all the options which can be pursued by having one's own personnel in the ministry, the situation changes entirely. It becomes feasible to circumvent the various political changes (and in fact, the more political changes the better, because each new minister will be more than happy to use personnel that are already aware of the functioning of the machine), and to create ones own set of private advisors. This allows one to participate in an unobtrusive but effective way in the real politics of France in the domain of research, whatever the political choices of the French citizen, whatever the politics of the ministries in charge! This is actually what happened from 1982 to 1996. This had extremely important consequences for bioinformatics and genomics, because it meant that genome programmes were not valued and should be avoided at all costs, except perhaps in the domain of cDNA construction and analysis.

It has been very difficult for those who wished to promote genomics to appreciate this biased situation. From the outside, French politics regarding genomics is difficult to understand. The ministries often made declarations to journalists of how they would support both genome research and the associated informatics, and give large amounts of money for the purpose. Journalists published the figures without any verification and people believed this was true. However, reality was the opposite, with only little money available. Also traditional genetics was favoured over genomics, with bioinformatics of genomes relegated to the lowest possible level, with continuous financial support for the old trends imposed by the local establishment.

Returning to the evolution of this research theme in France, bioinformatics is not an homogeneous discipline. It comprises *at least* three well separated domains of biology. Bioinformatics was born when the use of computers was found to be necessary. Their use led to major improvements in structural biology—solving the phase problem in the crystal analysis by x-ray diffraction. Computers were also useful in the study of models of behaviour, ecology, physiology, as well as medical biology, including image processing.

The third domain, and the only one that I shall consider in this article, is the study of biological sequences, polypeptides or genes and genomes. It can be considered as a science in itself, a view of life complementary to that given by *in vitro* or *in vivo* experiments, an *in silico* view (Médigue *et al.*, 1991). This view subsequently spread fast, first through the European Union administration, then in the academic world and even in the commercial world. By mid-April 1999, more than 800 web pages contain the word 'bioinformatics'. Because computers can be used in a variety of biological problems, the word 'bioinformatics' is sometimes avoided by disciplines that already exist, to the detriment of the new field of genomics. Some disciplines even refer to 'structural genomics', not to investigate the structure of chromosomes as the expression would imply, but to structural biology under the cover of a field which now appears to be fashionable!

The outstanding importance of biological sequences was clear after 1977, when DNA sequencing was made possible. Nucleic acid sequences provide us with a fundamental starting point for describing and understanding the structure, function, and development of all organisms. It was discussed when the creation of the European Molecular Biology Laboratory was decided, with a program integrating biological activities for sequencing of the genome of *Escherichia coli* (Danchin, 1995). Because mathematics and statistics are needed for analysing sequences, the scientists involved at the beginning of this new science were mostly population geneticists and evolutionists. A starting point for genomic bioinformatics in France was the development of statistics involving a large number of data sets, in particular for population genetics. Multivariate analysis, stemming from the work of (Sneath, 1957) and Michener and Sokal (1957) in the UK, had a very strong influence, and was followed by the development of a school of statistics, led by Jean-Paul Benzécri, who invented and developed factorial correspondence analysis (FCA) derived from principal component analysis (PCA) (Benzécri, 1984). In contrast to standard PCA, that uses measures to analyse continuous data with homogeneous properties, FCA specifically uses aggregate data that can be both discontinuous (or even simply logical) and comprise objects clustering in classes varying widely in the number of constituent objects. Benzécri was soon aware of the need for computers in his work and pushed hard for their wide use. Curiously enough, FCA is scarcely known in English-speaking countries (Hill, 1974). The reason might be that Benzécri always refused to speak or to write in English!

A prehistory to genome programs and bioinformatics in France: 1968–1985

Formal genetics

From their beginning, several communities that contributed to the creation of bioinformatics in France worked independently of each other. The first one is rooted in the French school of genetics, around Philippe L'Héritier, Madeleine Gans, Boris Ephrussi and Piotr Slonimski at the Centre de Génétique Moléculaire (CNRS, Gif sur Yvette). This community was aware of the development of quantitative genetics in the UK, and was interested in population genetics, in multicellular organisms development and in cytoplasmic heredity. An important point in the underlying epistemological background of this community was its central interest in relationships between biological objects, rather than the objects themselves (although this is often overlooked, Ephrussi was certainly a major player in the discovery of the gene-protein (enzyme) correspondence). This background resulted in a school of qualitative and quantitative genetics that maintained the concepts of genetics, over a background heavily dominated by biochemistry. This school was not directly related to the Institut Pasteur school of François Jacob, André Lwoff, Jacques Monod and Elie Wollman, whose major objects were usually bacteriophages and bacteria rather than multicellular organisms. The first school developed the techniques of multivariate analysis derived from that of Benzécri, and began to apply these techniques, at an early stage to the analysis of molecular processes, as well as to taxonomy.

Molecular phylogeny

The second school was born, from Emil Zuckerkandl's work and interest in molecular phylogeny. It started when the first protein sequences were known and grew when nucleic acids sequences were included. Numerical taxonomy, referred to above, played a major role. Zuckerkandl's while working with Linus Pauling, prime contribution was published in 1965: the concept of a 'molecular clock' added time to the trees of relationships that could be built up using statistical clustering and ordering techniques (Ayala, 1986). The underlying hypothesis was that the mutation rate is constant for all living species (and independent of the generation time of the organism), so that it is proportional to the time of divergence between two species. This hypothesis, which permitted phylogeny to be constructed from biochemical data as a complement to geological data, made Zuckerkandl quite famous, and the French government supported him by creating a research centre for him.

In the mid 1960s, Zuckerkandl came to Montpellier in the South of France and set up the Centre de Recherche en Biochimie Macromoléculaire (CRBM) in a new building that opened in 1968. At this time there was much discontent spreading throughout France and Europe. The proposed research was multidisciplinary (a fashionable concept of general use but notoriously difficult to implement), to structure the exploration of phylogenies. At the CRBM one could find biologists, molecular biologists, biochemists, chemists, computer scientists, NMR spectroscopists (but no geneticists because of little support for this discipline in France). This created an active and innovating community, but unfortunately also many human problems. In 1973, after a series of clashes with the community in his institute, Zuckerkandl returned to the USA where he continued his work on phylogenies. Among the work developed there, in parallel with the work of Margaret Dayhoff on the family of globins, there were investigations of the phylogenetical relationships of intracellular calciproteins (calmodulin and parvalbumins) using parsimony methods that challenged the concept of a molecular clock (Goodman *et al.*, 1979) and started a debate that is still active today.

Derived from the work developed in Montpellier, a major trend for bioinformatics in France was instigated by Richard Grantham with one of his first students, Christian Gautier, in Lyon. There, with Manolo Gouy, he created ACNUC, the first relational database for nucleic acid sequences (Gouy *et al.*, 1984). With its original management system and data structure, ACNUC has been the most advanced and efficient data base for biological sequences for many years. It is still used today (Perriere *et al.*, 1997) and although it is aging, it remains one of the most efficient sequence database management systems.

Neuronal networks

A completely different pathway allowed my participation in the work that led to genomics. I was trained as a 'pure' mathematician and in the mid 1960s I worked in this field at the Institut Henri Poincaré, next to the Institut de Biologie Physico-chimique. At the latter Institute, Marianne Grunberg-Manago proposed that I conduct experiments in biochemistry in her laboratory. In doing so, I set up a small group with the mathematician Philippe Courrège, trying to translate biological problems into formal terms: this was the tradition of the Bourbaki's group that had an important role in the development of modern mathematics and we wondered whether this would help biology. As a starting point we worked with the formalization of selective properties of the association of biological molecules. In 1971, I met Jean-Pierre Changeux who asked me whether we would extend our theoretical work to the formal properties of learning and memory in the animal nervous system. Thus we founded a group that met every Wednesday afternoon at IBPC for almost 6 years. Many scientists interacted with our group (Alain Chenciner, Jean-Michel Lasry,

Pierre Lusson, Michka Naïditch, Jean Petitot, Pierre Rosenstiehl, Gabriel Ruget, Bernard Saint-Loup, and even Benoît Mandelbrodt, who exposed us to the first ideas of what later became the theory of fractals). We started work on the selective properties of neuronal networks. With J.F.Blanchard suggesting that we work in terms of discontinuous objects and phenomena, we set up a diagram of synapse evolution (from labile to stabilized or regressed) that became the basis of our model of learning and memory (Changeux *et al.*, 1973). Later on, with Ruget, I extended this work to the field of immunology. Written in terms of networks, using the concepts of large intervals in the law of large integers (Azencott and Ruget, 1977), this led to a model formally similar to neuronal networks with emphasis on irreversibility of the learning steps (a rare but important feature in neuronal networks) (Danchin, 1979a). A result that was to play a significant role in the launching of the *Bacillus subtilis* genome programme had been obtained by Rosenstiehl and his colleagues at the Maison des Sciences de l'Homme: a network of local automata can have global properties (the 'firing squad' theorem), so that it might seem that there is no need for a specific organization of the genes in a cell to ensure a global property. This work demonstrated that this property (global behaviour derived from local properties) is highly unstable, so that each network needs a more or less *ad hoc* solution for each global behaviour (Rosenstiehl and Petitot, 1974). This placed cells in a perspective where the existence of global properties of gene organization appears necessary. Interaction of levels in gene expression might be a way to organize the appropriate hierarchy (Danchin, 1974), but the most likely consequence is that genes are organized in the chromosomes for fine tuning of their collective expression.

The research activity in the formalism of biological problems was subsequently interrupted, until 1985 when I decided to explore whether the then fashionable field of 'artificial intelligence' was relevant to biology (Gascuel, 1985; Rodier and Sallantin, 1985; Sallantin *et al.*, 1985; Soldano and Moisy, 1985). The underlying hypothesis was that a genome is not a collection of genes competing with each other, but that there is cooperation between the genes, allowing for a harmonious development of the cell processes in an ever-changing environment (Danchin, 1979b). It was, therefore, time to undertake a program for sequencing genomes, with an idea similar to that which cosmologists have used when they build up a map of the sky. It was absolutely necessary to use elaborate informatics to understand the meaning of the genomic text. To establish whether it was possible to bring together a community of scientists working in very different themes, I met with Olivier Gascuel, who had been involved in making medical expert systems with J.-F. Boivieux, and we chose a very specific example for a test of the efficiency of our collaboration. It was known at that time that proteins were secreted in eukaryotes and in bacteria using signal peptides. The consensus was that the mechanism and the structures were very similar in both cases. However, in industry it was well known that this could not be that simple: human growth hormon, even with appropriate transcription and translation signals, was not well secreted in *E.coli*. There was a difference in both cases. We therefore undertook to attempt to describe, using techniques of learning by discrimination, the particular features of signal peptides of *E.coli* as compared with human signal peptides (Gascuel and Danchin, 1986). This gave us a non-ambiguous picture of 17 descriptors of what was a signal peptide in *E.coli* (this work was often used by patent attorneys to reject patent applications on heterologous protein secretion). This convinced me that new approaches in computer science would bring many benefits to the analysis of genomes, once known.

Services to the community

As the number of biological sequences increased steadily it became necessary, along with research, to develop services for the molecular biology community. This was first achieved by the foundation of informal clubs of computer users. In particular, Francis Rodier created the club BISANCE at the Institut Jacques Monod in 1979, on the campus of Jussieu (Universities Paris 6 and Paris 7). The need for significant computer power led to the transfer of the centre to the Centre Interuniversitaire de Traitement de l'Information 2, University Paris 5 (CITI2). In 1983, BISANCE became an official service and the laboratory of biochemistry at the Ecole Polytechnique asked Philippe Dessen, who had long been interested in computing in Jean-Pierre Waller's laboratory, to participate in its activity (Dessen *et al.*, 1990). BISANCE played an important role in allowing isolated scientists, especially in the medical field, to have access to software necessary for sequence analyses.

At the Institut Pasteur, Alain Rambach [who with Pierre Tiollais, had been one of the leaders of early genetic engineering in France (Rambach and Tiollais, 1974)], started to computerize Jacob's *E.coli* bacterial collection. He was soon followed by Francis Schaeffer (Henri Buc's laboratory), who advocated the introduction of significant computing facilities in this institute. It soon became clear that the Institut Pasteur needed a service devoted to what was to become bioinformatics, and in 1982, under the strong influence of Henri Buc, a computation centre was created with Jean-Michel Claverie at its head. A service similar to BISANCE, SASIP, was set up and developed there until 1985. The underlying philosophy of these services was that mainframe computers were an absolute must. With the invention and rapid spreading of the MacIntosh computer

in 1984, a community of self-taught biologists rapidly developed sofware for the local analysis of biological sequences [this led to a programme that quickly spread in the international community of molecular geneticists, DNA Strider (Marck, 1988)]. This initiated a new type of collaboration between scientists who needed to communicate their experience and coordinate their activities. At the same time this triggered a controversy (as in other domains of science where computers play an important role), between those who only believe in large centres organized around mainframe computers, and those who advocate a more distributed activity, organized into a network.

Meeting clubs

In 1983 at the Institut Curie in Paris, Jean Sallantin, a computer scientist interested in artificial intelligence, and Jean-Louis Moisy organized regular meetings—the 'Points de Curie'. At the latter, the use of general methods in mathematics and statistics, as well as the general trends of the then fashionable artificial intelligence approaches, were discussed in relation to their application in biology, and in particular to sequence management and analysis (e.g. collective, 1984). Many future workers of *in silico* experimentation participated in these informal meetings: Frédérique van Bockstaele, Bernard Caudron, Olivier Gascuel, Christian Gautier, Manolo Gouy, Jacques Haiech, Alain Hénaut, Philippe Marlière, Jacques Ninio, Joël Quinqueton, Jean-Loup Risler, William Saurin, Henry Soldano, Alain Viari. The meetings stimulated H. Soldano, in 1986, to organize the Atelier de Bio-Informatique (ABI), an informal structure where scientists worked together (still informal in 1999!), located at the top of the small building where Marie Curie used to work (the basement of which was still quite radioactive) and now rue Cuvier, at the University Paris 6.

A similar idea was developed at many other places in France, along with existing laboratories (such as the Laboratories of Biometry in Lyon), ABIL and then ABIM in Marseilles and ABYSS in Caen. In 1985, Jean Sallantin decided to move from Paris to Montpellier. With this move, a CNRS laboratory of informatics was founded. This was very innovative because, although the laboratory did not depend on the CNRS department devoted to life sciences, it initiated the creation of a Groupement de Recherche for the investigation of biological problems (GDR 'Développement de l'intelligence artificielle en biologie et Robotique en Languedoc-Roussillon'). This GDR worked for 4 years. It triggered research on the user-friendly interface for the analysis of biological sequences—the 'Biostation', that incorporated the concept of user interface and hypertext (Haiech, and Sallantin), the use of artificial intelligence to learn from families of sequences (Quinqueton, Sallantin and Gascuel) and the development of phylogenetical approaches (Gascuel).

A difficult start: 1986–1990

Robert Sinsheimer, Renato Dulbecco and Charles DeLisi, each in his own way, proposed to sequence the human genome in 1985–1986 (Danchin, 1998). The project was presented as a technical programme whose outcome might help solve some problems in human health (Dulbecco, 1986). In 1986, André Goffeau proposed to the European Commission a programme aimed at sequencing the yeast genome as a typical illustration of the principle of subsidiarity. Both programmes were not directly related to the aim of answering a specific biological question, and this was probably the major reason for strong rejection by the community of biologists. In France the conceptual situation was initially different, since the reason why I proposed sequencing the genome of *B. subtilis* (at the spring meeting of the Société Française de Microbiologie in 1987) was a conceptual one (anonymous, 1996). The aim was to try to understand how genes can function collectively to allow the harmonious development of the cell. The underlying assumption was that this should be revealed as a prominent feature of the genome (which could therefore not be perceived as a simple collection of genes). The reaction to this proposal was almost universally negative, but Simon Wain-Hobson, who had recently sequenced the HIV genome, was interested. We proposed to sequence the genome of a universally spread sexual disease agent *Chlamydia trachomatis*, and approached the advisors of the Ministry of Research to suggest this as a genome programme (see below). In June of the same year Raymond Dedonder, then the director of the Institut Pasteur, attended the regular meeting on the biology of *B. subtilis* in California. There, James Hoch proposed to the community of specialists the sequencing of the genome of this bacterium. Dedonder remembered my proposal from the beginning of the year and asked me whether I was still interested. He was willing to set up a programme at the Institut Pasteur if I would take charge. Philippe Glaser was just completing the sequencing of a piece of DNA which we had identified as coding for the toxic adenylate cyclases of *Bordetella pertussis* (Glaser *et al.*, 1988). He was asked to set up a sequencing laboratory in my unit at the Institut Pasteur. Late in 1987, I was commissioned by the directorate Biology, division Biotechnology of the Commission of the European Communities to write a report for their white paper on the Biotechnology Action Programme for sequencing genomes. I was asked by André Goffeau, who, rather than see it as a competitor of the yeast sequencing programme supported the *Bacillus subtilis* programme, to give a conceptual justification for this type of research (Danchin, 1988). All this led to the birth of the *B. subtilis* genome programme, a history in itself: starting from a collaboration between five European and five US

laboratories, it ended as a collaboration between Europe and Japan, with no American group (anonymous, 1996; Danchin, 1998; Kunst *et al.*, 1997).

An entirely different story, the human genome project in France, had started at the same time, and the pathways of the workers crossed in the Ministry of Research in 1987–1988. Daniel Cohen, an active collaborator of Jean Dausset at the Centre d'Etude du Polymorphisme Humain (CEPH), became interested in participating in the sequencing of the human genome, using the DNA libraries that were deposited at the CEPH. He tried to convince the Ministry of Research that this was a worthwhile enterprise, and that the CEPH, with its private structure, could begin a sequencing programme in a more efficient way than public organisations, with direct financial help from the ministry. Cohen began to plan the construction and general experimental planning of a CEPH building entirely devoted to the sequencing of the human genome. As early as 1989 the CEPH could recruit scientists and engineers, obtain robots and industrial equipment for the mapping and sequencing of the human genome on a large scale. Alongside this, an application to the Commission of the European Communities for a Eureka programme, together with the private society Bertin (and in partnership with Amersham and the British ICRF), aimed at the creation of an industrial provider of the machines and techniques necessary for the mapping and sequencing of genomes. This project, named Labimap, aimed at providing rapid oligonucleotide synthesizers, robots and reactors for automatic preparation of plasmids, sets for large-scale hybridization and miniature sequencing gels for electrophoresis. Clearly, it was then correctly perceived by Cohen that successful genome programmes had to incorporate a significant scaling up of the molecular biology techniques. It would be interesting to analyse the reasons why Labimap was a failure and did not lead, as was hoped, to laboratory equipment products on the European market.

At this point the reader can, at least in part, understand the paradoxical situation in France in the period 1988–1990. On the one hand there were scientific workers willing to start genome programmes with apparent support from the Ministry of Research (at least in the discourses of the minister and its direct advisors). On the other hand there were the majority of scientists, and especially those who had provided services to the ministry, who, in the background, were acting strongly against the development of these programmes. Under these circumstances, of course, things did not move very fast: the Institut Pasteur provided some support to the *B. subtilis* programme, the BAP EU supported the yeast programme, and the CEPH the human genome programme. This was too little for Cohen. By chance in 1987, Bernard Barataud, the ebullient chairman of the Association Française contre les Myopathies (AFM), had successfully started a Téléthon (a charity show) in France, and was thinking of using the fund-raising money he collected every year for an ambitious programme in human genetics. Cohen understood all the advantages that could be extracted from this interest and persuaded Barataud that sequencing the human genome would pave the way for the identification of genetic diseases. Barataud identified places where a centre, the size of a factory, could be set up in Evry (south of Paris, where the AFM had offices not far from his home) and, with the first prototypes created by Bertin for Labimap, the first Généthon was created at the end of 1990. Its main aim was to establish three major programmes: libraries of yeast artificial chromosomes carrying a shotgun of the human genome (Cohen), a detailed genetic map of the human genome (Jean Weissenbach, then at the Institut Pasteur) and a complete collection of human cDNA clones (Charles Auffray). Because of the lack of community support, the beginning of the Généthon had a low profile, with a scientific council comprising only scientists interested in what would later become known as genomics.

Except for the case of the *B. subtilis* programme (Danchin, 1988), not much thought had been given to the need for an important informatics infrastructure for these programmes. However, Alain Hénaut from the Centre de Génétique Moléculaire of the CNRS organized a new Diplôme d'Etudes Supérieures Spécialisées de Bioinformatics at the University Paris 6, the first training programme of this type in Europe. As a follow up, in 1989 Jacques Haiech created the Diplôme d'Etudes Approfondies Ingenia associating the universities of Montpellier and Marseilles. The Ministry of Education supported this policy by creating fellowships for the financial support of students following a double training biology/informatics and for spring or summer schools where students could become familiar with both disciplines (Hénaut and Marie-Odile Delorme). It is worth pointing out that the students trained during that period in Lyon (biometry), Paris or Montpellier–Marseilles, are the ones who are today responsible for the training in bioinformatics in a variety of universities or institutes in France or at the EMBL/EBI. This demonstrates both the enormous importance of training and the long inertia of the training system. The lack of support for training in recent years is the reason for the present-day lack of scientists and engineers in bioinformatics.

A false start?: 1991–1995

Because of the many conflicts created by the misunderstandings of the major participants in the research community in France, the situation at the end of 1990 was not favourable for the development of genomics, despite

the apparent support of the minister of research Hubert Curien. That year, the minister launched a 'Programme National Génome Humain', that was supposed to serve as a 'Groupement d'Intérêt Public', a structure noteworthy for being unwieldy and very long to establish (and especially easy to paralyse by the eventual unwilling partners). It was also clear that this programme, based on a stereotyped report that had not taken into account the work in genomics, did not understand the need for the sequencing of model genomes, nor the real purpose of sequencing the human genome. In March 1991 Jacques Hanoune was asked to prepare the 'GIP génome' programme, that was supposed to associate the various EPSTs (recalling that they were not in favour of genome programmes) involved in biology and medicine (INSERM, CNRS, INRA). Money was distributed in haste by the ministry, but much of it was, for some unaccountable reason, not available for use! In parallel, Généthon, which had succeeded in obtaining significant data, asked the ministry to provide direct support. In Spring 1992, at the annual Cold Spring Harbour meeting, Cohen presented a first complete map of chromosome 21, and in the autumn of the same year published a first contig map using YACs with up to 1 Mb human DNA inserts. This map, which used important computation facilities at INRIA (G. Vaysseix, J.-J. Codani) had a big international impact and suggested that France was at the forefront of genomics (Weissenbach *et al.*, 1992). At a meeting of the EC held at Elounda in Crete in the spring of 1991 the sequencing of the first yeast chromosome (chromosome III) had been presented, together with that of a large fragment of the *B. subtilis* genome. Both organisms revealed that at least one half of the genes did not have counterparts in other organisms. These 'elusive, esoteric, conspicuous genes', as described by Piotr Slonimski at the meeting (EEC genes) demonstrated that Europe was actually doing well in what was not yet known as genomics (Oliver *et al.*, 1992).

In this context the Ministry of Research decided to accelerate the creation of the GIP genome, now changed to Groupement de Recherches et d'Etudes sur les Génomes (GREG, this name indicated that, at last, model genomes had been included), and asked Slonimski to take its helm. The GREG was formally installed at the end of January 1993, although it had been partially funded in 1992 (54.2 MF) (anonymous, 1994). During its short mission (25 January 1993–29 October 1996) the GREG financed a variety of genome programmes, but it only received funds from 1993 to 1995 (142 million F, including support to the CEPH) (Slonimski, 1996). Too small, and not supported by the EPSTs that were supposed to organize it, this effort did not allow the establishment of a new community in France which would devote a significant part of its activity to genomics. Even worse, the support of bioinformatics, that started 1992–1993 with the creation of a scientific committee of the GREG and supported the work of Jean Thierry-Mieg (the famous ACeDB database), as well as calls for proposals specially devoted to the task of developing this discipline, was suddenly, on the order of the Ministry of Research (by fax!), forced to stop (Thiellement, 1995). This reveals explicitly how certain scientists made efforts to try and stop the development of genomics in France, despite the fact that the academic level of bioinformatics was of a high standard compared with that in other countries.

At the same time, work at the Généthon had proceeded well. However, from the start of the programme, Barataud had explicitly said that the AFM would only help for a period of 3 years—it is an association of patients that does not have the support of public research as a primary goal—and that it would subsequently shift to other types of research, in particular to applied research. Three years is a very short period of time for programmes as ambitious as the sequencing of the human genome and, when it appeared that the map produced by the first period of the Généthon was not as useful as originally claimed (but nevertheless a very significant success), and that the programme for producing cDNA had not been as efficient as comparable programmes elsewhere in the world, tensions started to creep in at the Evry site. The AFM agreed to support Généthon I for a fourth year, but irrevocably set up a Généthon II, then a Généthon III, where genotyping and gene therapy were focused for the identification and cure of muscle genetic diseases. Under the lead of Weissenbach, thousands more markers were identified in order to help locate the genes of diseases, and Généthon participated in the preliminary mapping of 400 disease genes and in the identification of 60 of them, a very significant success (http://www.afm-telethon.asso.fr/). Cohen followed a different pathway and turned to industry, finally becoming the scientific director of the company Genset (http://www.genxy.com/Business/bus_pharmaco.html) (Balter, 1997). These movements should have placed the government initiative creating the GREG as a major player in genomics and in particular in bioinformatics in France. Unfortunately, this was not so.

As described above, a series of meeting clubs had been developing techniques for exploiting and managing sequence data. There were services to help isolated scientists work with sequences. In 1987, at the onset of the first model genome sequencing programme, it became obvious that it would be of the utmost interest to coordinate the activity of the various groups involved in this type of research (Danchin, 1990). I tried, therefore, to involve the CNRS, through the creation of a 'Groupement Scientifique', the body then existing to coordinate research between different groups. Unfortunately, despite numerous attempts, this was not successful until 1991, when the director of the department of life sciences in

this institution, Claude Paoletti, finally discovered that genomes were becoming somewhat fashionable and were probably important (this was stimulated by the sudden involvement of the then rival Ministry of Education, see below). The CNRS had at that time already created several 'Programmes Interdisciplinaires de Recherche' (PIRs), without considering genome programmes (despite explicit recommendations given by the 'schéma directeur' produced by the Comité National evaluating research at the CNRS). The only possibility left was to integrate genome bioinformatics into an existing programme. The chosen one was the PIR IMABIO (for Ingénierie des MAcromolécules BIOlogiques). The main difficulty was that informatics in this programme was devoted to structural biology (a thoroughly inadequate support of genomics, see above) and that the philosophy of genetics and genomics, which deal with relationships between objects rather than with objects, was not fully understood. The creation of a Groupement de Recherche, the GDR 1029 Informatique et Génomes (Danchin and Rechenmann), was therefore delayed (it only existed for 4 years from April 1992, with retroeffect to January of the same year). The white paper giving the conclusions of the audit programme summarizes well the misunderstanding between genomics and structural biology, and the comments of the auditing committee would still be relevant today in the case of other genomics programmes:

> this complementary action (i.e. bioinformatics at the GDR1029) to structural biology remained marginal with respect to the programme (IMABIO). Its integration in IMABIO appears to have been purely opportunistic, and was meant to compensate for the almost complete absence of the CNRS in programmes aiming at deciphering genomes. Everything separated the GDR from the rest of the PIR: a different theme, centered on nucleic acids, while the programme centered on proteins, an approach that led to concentrate on the study of links between biological objects rather that on objects, a functioning resting on a network structure rather than on geographical poles

etc. (Kahane, 1994).

In the mean time, it had been possible to organize schools dedicated to a multidisciplinary approach to genome sequences analysis and management (two IMABIO schools, Spring 1991 and Summer 1992). The reason for the change of direction of the CNRS and INSERM was that a new player entered the picture and played a positive role for the future of genomics: the Direction de la Recherche et des Etudes Doctorales, from the Ministry of Education (differing from the Ministry of Research) became suddenly interested in biological sequences and decided to be active in this domain. The development of this policy was concomitant with a change in the Ministry of Education (Lionel Jospin, today's Prime Minister of France), with his personal advisor Claude Allègre (today's Minister of Education *and* Research), and Vincent Courtillot (today's director of the research sector in the ministry) as director of the DRED, Philippe Vigier (now retired) and Alain Hénaut (now advisor to the minister's cabinet) taking charge of biology. The first action of the DRED was to organize the future of universities around Paris (Universités 2000), with significant emphasis on genomics. Associated with some development of bioinformatics in future universities, the problem of training (in particular mutual training of geneticists and computer scientists) was taken into account. This resulted in a first meeting in Seillac in November 1990, organized by INSERM with the DRED, an IMABIO school at the ABI in April 1991, followed by a spring school with computer scientists in Massy, near Paris (8–12 April 1991).

In spite of its limited resources, the GDR 1029 had a significant impact on scientists interested in *in silico* analysis of genomes (more than 100 scientists were at some point involved in the GDR). The GDR organized bioinformatics along four main geographical axes: Lyon–Grenoble, Montpellier–Marseilles, Paris, Toulouse–Strasbourg. Its scientific action can be summarized as follows (http://cosmos.imag.fr/GDR-INFOGENOMES/GDR-home.html). It devised a network of collaborations between groups interested in the same problem. The unifying theme was the genome analysis of the model genomes, *E.coli*, *B. subtilis* and *S. cerevisiae*, chosen to be sequenced. Some emphasis was placed on the mutual training of the community of geneticists and computer scientists. A central focus of the coordination was the development of tools which would help identify crucial experiments that had to be performed *in vivo* (using reverse genetics) and to validate or nullify the hypotheses made *in silico*. Three major conceptual themes: combinatorials and statistics for sequence analysis, artificial intelligence and methods for phylogenetic analysis, combined into eight general sub-themes, were defined. These were sequence alignments and phylogeny, secondary structures and regularities in sequences, identification of motifs in genes, identification of coding sequences in eucaryotic nucleotide sequences, knowledge acquisition, help in discovery, knowledge representation, constraints fulfilment and molecular biology (collective, 1995). Research in these areas were published (Biochimie, 1993, 1996) at an international conference held in Lyon (proceedings, 1994), and in many specialized articles. When the GDR closed at the end of 1995 it had produced more than 100 publications and had initiated a large interest in genomics amongst the

community of computer scientists. Unfortunately, this coincided with a change in the French government that gave more power to those who had used much of their energy to prevent the development of genomics in this country.

True or false start?: 1995–1996

Two major bodies, an agency that was in principle able to grant funds, the GREG, and a structure of coordination, the GDR 1029, were beginning to organize genomics in France when one of those frequent changes in the politics of the various ministries of research was initiated. The first outcome was that the committee that had financed bioinformatics at the GREG was disbanded, and there was no money for a follow up of the GDR 1029. The new ministry introduced new grant agencies, the 'Actions Concertées Coordonnées dans les Sciences du Vivant' which influenced the actions initiated by the GREG in bioinformatics, and tried to undo what had been constructed (Thiellement, 1995). At the same time, the action of the new Ministry of Education (universities) and Research reduced the autonomy and closed the GREG, 2 years before its expected closure time (Slonimski, 1996). The end of the GREG initiated a period when scientists and administrators began to organize their individual criteria as to how they thought genomics should develop in France. Most of this negative action was carried out by those who had been most reluctant to support genomics at its start. Committees were set up at the EPSTs to discuss the future of this discipline, and an idea crept in to build on the success of The Institute for Genome Research in the United States, as well as on the Wellcome Trust initiative, which created the Sanger Centre. France had to have its own project, under an obvious code name, the TGS, similar to the TGV (the French fast train), a 'centre de Très Grand Séquençage'. Its natural head would be Weissenbach, who had, discretely but very efficiently, built up a fine map of the human genome. However, little of his advice was sought when it was decided to set up the centre in Evry, where one could identify both Généthon and Genset as possible partners but where no university infrastructure in biology existed—although there was sufficient room in Paris (at the University Paris 5). Furthermore, it was very difficult to set up the organization from an administrative point of view: the fall of the GREG had precluded the use of this body, and so another GIP had to be set up (which was notoriously unwieldy). Finally, the Centre National de Séquençage was set up in Evry (1 January 1997) in a building where everything had to be reconstructed, under the direct control of the Ministry of Research and the CNRS, which had already created its own genome programme. The initial mission of the CNS, headed by Weissenbach, was to perform large sequencing programmes, half for the community in France and half for its own programme that had to be decided in dialogue with its grant agencies. The overall structure was to be coordinated by three different structures: an administrative committee, an orientation committee (where most of the partners might, at some point, be interested in genomes), and a scientific committee—clearly a very awkward kind of organization.

At the same time, bioinformatics had to be reorganized since it had been pulled out of the GREG. In 1994, Vaysseix proposed a programme, Genespace, that was intended to develop a network of bioinformatics services in France, taking into account the existence of local servers. This initiative was stopped by the ministry at its very beginning. With the help of the AFM, the Ministry of Research decided instead to create a new organization (a 'Groupement d'Intérêt Scientifique'), and thought that it should correspond to the service BISANCE, then at the CITI2. On 1 January 1995, BISANCE was transferred to Villejuif with Dessen at its head, and the GIS-INFOBIOGEN ('Informatique appliquée à l'Etude des Biomolécules et des Génomes') was established, headed by Vaysseix. The partners of the GIS were four EPSTs, four Paris universities, Généthon, and the Ministry. Unfortunately, in spite of the clear directive from the Ministry, there was little financial support for it. This organization was meant to be a service, not a research centre, showing that no financial support for research in bioinformatics had been organized.

Sadly, the stage was no longer set for France to be a leading country in the genomics era, in spite of its early significant participation in genome sequencing. The yeast genome sequence was completed in 1996 (Goffeau *et al.*, 1996), and *Bacillus subtilis* was well on its way at that time. However, as in a theatre play, political circumstances entirely changed the deal. France's President, Jacques Chirac, decided to dissolve the National Assembly and the election result led to a complete change in government. Jospin became the Prime Minister, and Allègre his Minister for Education and Research. The people who had been in charge of the DRED, and had supported genome research were now major players.

Epilogue: towards the future

It is not sufficient for a minister to decide a political line for it to be effective. We have seen how it is possible to slacken the pace of reform, even when it is welcome, and this is occurring in France. However, a new pace is set, with some happy results: the genome sequence of *B. subtilis* was published in November 1997 (Kunst *et al.*, 1997), immediately followed by the sequence of *Mycobacterium tuberculosis* (Cole *et al.*, 1998) and *M. lepreae* is expected to come out soon. A new committee, headed by Nicole Le Douarin, is supposed to help with the coordination of EPSTs and other organizations active

in research in biology and support of 45 million euros has been provided for a programme of, genomics headed by Jacques Demaille, (including 12 millions euros for the CNS and 8 million euros for the Centre National de Génotypage). A Génopole is created in Evry (this means that the general orientation of Evry's university has to shift to biology), where INFOBIOGEN, with Vaysseix at its head, will develop a new structure, with significant support from the Ministry. Many genome programmes are underway—a large programme, Génoplante, funds and coordinates genomics of *Arabidopsis thaliana* and plants important in agronomy, with effective support from INRA (Michel Caboche), a new laboratory, devoted to the genomics of pathogens, has been set up at the Institut Pasteur (Kunst and Glaser), and the CNS is supposed to complete the sequencing of chromosome 14 before June 2000 (Weissenbach). In addition to the centres which develop community services (see above), we can today identify three major research goals in genome bioinformatics: a network centred on the phylogenetical analysis of sequences: Lyon, Grenoble, Marseilles and Montpellier; a joint effort between laboratories in Strasbourg and Toulouse working specifically on RNA structure; and a network of laboratories around Paris (Institut Pasteur, Université Paris 6, Université de Versailles-Saint-Quentin). In addition several Genopoles will be created. It is still difficult to foresee the future of genomics and bioinformatics in France, but we may hope at last that more time will be devoted to research than trying to extract the grains of sand that have been put everywhere in the machine by those who were unhappy with the genome initiatives!

Acknowledgements

I wish to thank all those who participated in the GREG and GDR1029 effort to start an *in silico* biology of genomes in France, and in particular Jacques Haiech, Alain Hénaut, Claudine Médigue, Ivan Moszer, François Rechenmann and Alain Viari. I also wish to give my special thanks to André Goffeau and Piotr Slonimski, who were extremely supportive from the beginning.

References

Adams,J. (1998) Benchmarking international research. *Nature*, **396**, 615–618.

Allègre,C. (1998) Intervention du Ministre de l'Education Nationale de la Recherche et de la Technologie à l'installation du Comité de Coordination des Sciences du Vivant, 20 October.

Anonymous (1994) Le GIP GREG: origine, définition et missions. *La lettre du GREG*, **1**, 2–7.

Anonymous (1996) Itinéraire d'un projet financé par le GREG: le séquençage du génome de *Bacillus subtilis*. *La Lettre du GREG*, 12–14.

Ayala,F.J. (1986) On the virtues and pitfalls of the molecular evolutionary clock. *J. Hered.*, **77**, 226–235.

Azencott,R. and Ruget,G. (1977) Mélanges d'équations différentielles et grands écarts à la loi des grands nombres. *Zeitung Wahrscheinlichtkeitstheorie verw. Gebiete*, **38**, 1–54.

Balter,M. (1997) Key player: Daniel Cohen. *Science*, **275**, 773.

Barreto,G.R., Bauin,S., Gomez,I., Bordons,M., Cami,J., White,P.C. and Herskovic,S. (1997) Science by the country. *Science*, **276**, 882–885.

Benzécri,J.P. (ed.) (1984) *L'Analyse des Données. La Taxinomie*. Dunod, Paris.

Biochimie (1993) Genomes and informatics. *Biochimie*, **75**, 319–422.

Biochimie (1996) Genomes and informatics. *Biochimie*, **78**, 299–375.

Burian,R.M. (1990) La contribution française aux instruments de recherche dans le domaine de la génétique moléculaire. In Fischer,J.-L. and Schneider,W. (eds), *Histoire de la Génétique* A.R.P.E.M & Editions sciences en situation, Paris, pp. 247–269.

Burian,R.M. and Zallen,D.T. (1992) The non-interaction of regulatory and human cytogenetics in France, 1950–1975. In Dronamraju,K. (ed.), *History and Development of Human Genetics* World Scientific Publishing Co., London, pp. 91–101.

Burian,R.M., Gayon,J. and Zallen,D. (1988) The singular fate of genetics in the history of french biology, 1900–1940. *J. Hist. Biol.*, **21**, 357–402.

Changeux,J.P., Courrège,P. and Danchin,A. (1973) A theory of the epigenesis of neuronal networks by selective stabilisation of synapses. *Proc. Natl. Acad. Sci. USA*, **70**, 2974–2978.

Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E., 3rd, Tekaia,F., Badcock,K., Basham,D., Brown,D., Chillingworth,T., Connor,R., Davies,R., Devlin,K., Feltwell,T., Gentles,S., Hamlin,N., Holroyd,S., Hornsby,T., Jagels,K., Barrell,B.G. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.

collective (1984) Intelligence artificielle et analyse des séquences biologiques. *Points de Curie 1984*, Atelier de Bio-Informatique, Institut Curie.

collective (1995) Rapport d'activité du GDR 1029, juin 1995.

Danchin,A. (1974) Règles de réécriture en biologie moléculaire. *Change*, **19**, 10–24.

Danchin,A. (1979a) The generation of immune specificity: a general selective model. *Molec. Immunol.*, **16**, 515–526.

Danchin,A. (1979b) Is a metabolic control for the doubling of the macromolecule synthesizing machinery possible? *Biochimie*, **61**, 45–50.

Danchin,A. (1988) Complete genome sequencing: future and prospects. In Goffeau,A. (ed.), *BAP 1988–1989* Commission of the European Communities, Brussels, pp. 1–24.

Danchin,A. (1990) *Une Aurore de Pierres. Aux Origines de la Vie (Remerciements)*. Le Seuil, Paris.

Danchin,A. (1995) Why sequence genomes? The *Escherichia coli* imbroglio. *Molec. Microbiol.*, **18**, 371–376.

Danchin,A. (1998) *La Barque de Delphes, ce que Révèle le Texte des Génomes*. Éditions Odile Jacob, Paris.

Dessen,P., Fondrat,C., Valencien,C. and Mugnier,C. (1990) BISANCE: a French service for access to biomolecular sequence databases. *Comput. Appl. Biosci.*, **6**, 355–356.

Dulbecco,R. (1986) A turning point in cancer research: sequencing the human genome. *Science*, **231**, 1055–1056.

Gascuel,O. (1985) Structural descriptions. Discrimination and learning of these descriptions. *Biochimie*, **67**, 499–507.

Gascuel,O. and Danchin,A. (1986) Protein export in prokaryotes and eukaryotes: indications for a difference in the mechanism of exportation. *J. Mol. Evol.*, **24**, 130–142.

Glaser,P., Ladant,D., Sezer,O., Pichot,F., Ullmann,A. and Danchin,A. (1988) The calmodulin-sensitive adenylate cyclase of *Bordetella pertussis*: cloning and expression in *Escherichia coli*. *Molec. Microbiol.*, **2**, 19–30.

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. and Oliver,S.G. (1996) Life with 6000 genes. *Science*, **274**, 546–547, 563.

Goodman,M., Pechère,J.F., Haiech,J. and Demaille,J.G. (1979) Evolutionary diversification of structure and function in the family of intracellular calcium-binding proteins. *J. Mol. Evol.*, **13**, 331–352.

Gouy,M., Milleret,F., Mugnier,C., Jacobzone,M. and Gautier,C. (1984) ACNUC: a nucleic acid sequence data base and analysis system. *Nucleic Acids Res.*, **12**, 121–127.

Hill,M.O. (1974) Correspondence analysis: a neglected multivariate method. *Appl. Stat.*, **23**, 340–353.

Kahane,B. (1994) Synthèse de l'audit du programme IMABIO. CNRS, Délégation aux études et Audits, Paris.

Kotiaho,J.S. (1998) Papers vanish in *mis*-citation black hole. *Nature*, **398**, 19.

Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessiéres,P., Yoshikawa,H., Danchin,A. *et al.* (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.

Marck,C. (1988) 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucleic Acids Res.*, **16**, 1829–1836.

May,R.M. (1997a) Science by the country. *Science*, **276**, 884–885.

May,R.M. (1997b) The scientific wealth of nations. *Science*, **275**, 793–796.

Médigue,C., Gascuel,O., Soldano,H., Hénaut,A. and Danchin,A. (1991) From data banks to data bases. *Res. Microbiol.*, **142**, 913–916.

Michener,C.D. and Sokal,R.R. (1957) A quantitative approach to classification. *Evolution*, **11**, 130–162.

Oliver,S.G., van der Aart,Q.J., Agostoni-Carbone,M.L., Aigle,M., Alberghina,L., Alexandraki,D., Antoine,G., Anwar,R., Ballesta,J.P., Benit,P. *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38–46.

Paris,G., DeLeo,G., Menozzi,P. and Gatto,M. (1998) Region-based citation bias in science. *Nature*, **396**, 210.

Perriere,G., Moszer,I. and Gojobori,T. (1997) The NRSub database: update 1997. *Nucl. Acids Res.*, **25**, 53–56.

Price,N.C. (1998) What's in a name (or a number or a date)? *Nature*, **395**, 538.

Proceedings (1994) Computer science and molecular biology. In Sankoff,D., Cedergren,R., Gautier,C. and Rechenmann,F. (eds), *Computer Science and Molecular Biology* CNRS/INRIA/IMAG, Lyon.

Rambach,A. and Tiollais,P. (1974) Bacteriophage lambda having EcoRI endonuclease sites only in the nonessential region of the genome. *Proc. Natl. Acad. Sci. USA*, **71**, 3927–3930.

Rodier,F. and Sallantin,J. (1985) Localization of the initiation of translation in messenger RNAs of prokaryotes by learning techniques. *Biochimie*, **67**, 533–539.

Romains,J. (1925–1934) *Les Hommes de Bonne Volonté*. Flammarion, Paris.

Rosenstiehl,P. and Petitot,J. (1974) Automate asocial et systèmes acentrés. *Communications*, **22**, 45–62.

Sallantin,J., Haiech,J. and Rodier,F. (1985) Search for promoter sites of prokaryotic DNA using learning techniques. *Biochimie*, **67**, 549–553.

Schoonbaert,D. and Roelants,G. (1998) Impact takes precedence over interest. *Nature*, **391**, 222.

Slonimski,P. (1996) Editorial. *La lettre du GREG*, **8**, 1–2.

Sneath,P.H. A. (1957) The application of computers to taxonomy. *The J. gen. Microbiol.*, **17**, 201–226.

Soldano,H. and Moisy,J.L. (1985) Statistico-syntactic learning techniques. *Biochimie*, **67**, 493–498.

Thiellement,H. (1995) Editorial. *La lettre du GREG*, **5**, 1.

Weissenbach,J., Gyapay,G., Dib,C., Vignal,A., Morissette,J., Millasseau,P., Vaysseix,G. and Lathrop,M. (1992) A second-generation linkage map of the human genome. *Nature*, **359**, 794–801.