Molecular Microbiology (1995) **18**(2), 371–376

MicroOpinion

Why sequence genomes? The *Escherichia coli* imbroglio

Sir,

A little more than a decade ago, many people, including highly-respected molecular biologists, advocated the launch of a huge molecular biology programme, aimed at sequencing the whole human genome. This was perceived mostly as a technical feat rather than as a scientific programme. Indeed, genome sequencing was (and often still is) presented as an aim in itself, not as an approach to solve biological questions. This is certainly why sequencing programmes are often treated dismissively by many members of the scientific community (but, in contrast, not by the media where what is big and difficult to understand is always beautiful!). There are many good scientific reasons for sequencing genomes, and programmes for sequencing the genomes of several different organisms besides humans were soon proposed and developed. Now that two complete bacterial genomes have been sequenced (*Haemophilus influenzae* (Fleischmann *et al.*, 1995, *Science* **269:** 496–512) and *Mycoplasma genitalium* (Fraser *et al.*, 1995, *Science* **270:** 397–403)), and that about twenty bacterial genomes have already been half sequenced, nobody would argue that this is not of interest.

In the mid-seventies, a group of scientists proposed to sequence the *Escherichia coli* K-12 genome by an integrated programme which could justify the construction and development of the EMBL laboratory by its mother organisation, the EMBO. Curiously enough, the scientific side – i.e. sequencing as a means to answer specific questions concerning this, the most widely studied bacterium – was initially given little attention, perhaps because the project was considered as secondary to the human genome initiative, which was always placed in the limelight. Many molecular biologists like myself still have enormous numbers of questions concerning the molecular basis of life, and are convinced that knowledge of the total chemical definition of a living cell will be an important step in our quest to answer many of these questions. Like many of my colleagues, I remain convinced that *E. coli* is an organism admirably suited to such analysis, so why is it taking so long to sequence the K-12 genome, and what can we learn from the problems encountered?

It is not possible to discuss the situation of programmes aimed at the sequencing of complete genomes, and in particular that of *E. coli*, without alluding to both epistemological and political issues. Indeed, much more has been written on the subject that will never come to public attention, mostly in the huge number of grant applications or government papers concerning sequencing projects that remain confidential, even years after they were written. Sequencing projects also make the headlines in the 'popular' press at frequent intervals, and it is often instructive to compare statements and opinions made here with those in 'serious' articles published in scientific journals.

Why sequence genomes?

All major domains of science have evolved from a starting point where a given field of the physical world has been submitted to categorization, taxonomy or systematics. For example, Mendeleieff constructed a catalogue of the atoms present in the universe, a catalogue of stars in the sky has been established, and a catalogue of plants and animals has been (and still is being) constructed, organizing our knowledge of the complete living world. In the same way, before trying to understand the complete molecular basis of life, we need to have a complete chemical description of a cell. This requires identification of all metabolites in that cell. Most small molecules have already been identified in model organisms, but we are far from having identified all macromolecules; sequencing an entire genome is a major step towards this end because, using the genetic code, one has access not only to the chromosome sequence but also to the proteins it encodes, thus providing us with a list of all materials needed to make a living cell. This is certainly very far from understanding life, but is still much more than a simple collection of objects, because the DNA molecule is also the blueprint for the construction and the functioning of the cell. Knowing the sequence of entire genomes opens up a new field of research akin to sciences aiming at deciphering texts in unknown languages or devising and 'breaking' codes: biology 'in silico'. We are still at a very early stage but the elements of understanding that we have already (and the genetic code is not a trivial matter in this respect) permit us to characterize much of the function and regulation of gene expression, which is the major aim for the majority of molecular geneticists.

In order to study the function of a given gene (product), it is usually essential to identify and sequence it. Until recently, it was very often extremely costly in terms of human involvement to isolate individual genes in order to study their function; many years of hard work were often needed to isolate a single gene. This was because of a variety of reasons: lack of appropriate phenotype or mutants, toxicity of the gene when cloned as an isolated entity, functional redundancy etc. The sequencing of complete genomes results in the identification of all genes, and can help solve most of the individual problems that arise during the cloning of individual genes. Even considering only this simple task, genome programmes are actually very cost effective.

However, a genome is much more. It is a structure recognized by the replication machinery, and it is likely to be organized over short and even perhaps long distances. Because genes are collectively expressed, functionally-related genes are likely to have characteristic signatures or sequences. A genome contains promoters and control regions, which must have features recognized by the transcription machinery (and which should be recognizable by appropriate analysis of the sequence). A genome is a collection of coding sequences which can be translated using the genetic code. The corresponding collection of gene products must be compartmentalized within the cell, and the information for this must be present in the DNA sequence. The gene products are also the result of evolution, and their kinship will provide information on their function and structure, as well as their origin. A genome has a style of its own permitting one to differentiate it from other genomes, and providing the cell with a means to discriminate self from non-self. Knowledge of this, and of its implications for the cell and genome expression are essential for the construction of heterologous systems which play such an important role in modern biotechnology.

Which genomes?

All sequencing programmes involve a considerable amount of work (Danchin, 1989, In *Sequencing the Yeast Genome, A Detailed Assessment.* Commission of the European Communities pp. 1–24; Anonymous, 1990, *Understanding Our Genetic Inheritance.* The US Human Genome Project. National Institutes of Health and Department of the Energy, USA; Barnhardt, 1990, *Human Genome 1989–90 Program Report.* US Department of Energy; Watson, 1990, *Science* **248:** 44–51; McLaren, 1991, *Human Genome Research: A Review of European and International Contributions.* Medical Research Council, U.K). It is therefore important to choose appropriate organisms at the start of such programmes. In particular, one must have some preconceived idea of the way in which the information content of the genomes will be analysed. This requires some insight into the nature of this information. The human genome contains three billion base pairs, spread through 46 chromosomes. However, the coding information (expressed genes) is actually much smaller. According to different authors, there are 50 000–100 000 human genes. This corresponds to approximately 10^8 base pairs: 3% of the total length of the genome. In complex genomes, such as that of *Homo sapiens*, many other specific features reflect the outcome of their evolutionary history, determined by constraints involving recombination and error correction. In addition, human genes and the messenger RNA they specify are not colinear because of the presence of introns: a 1 kb messenger is sometimes encoded by a DNA fragment encompassing tens, hundreds, or even one thousand kilobases. Therefore, short exons may be hidden in an ocean of intron sequences, and sequencing even very long DNA fragments could be insufficient to define the complete gene it specifies. In addition to difficulties inherent in this modular mode of construction, it is extremely difficult to characterize a human genome at the microscopic level. Because of the large variability in human populations, the fine structure of individual genomes is highly polymorphic. This means that the determined sequence will be a patchwork coming from many individuals, an assembly of sequences that might sometimes be incompatible with each other. Thus, at least 95% of the sequence of human genomes corresponds to 'archives', i.e. sequences that, because they are not a heavy load for replication or gene expression, are maintained from generation to generation. It will therefore be extremely difficult, in the absence of independent information (e.g., the presence of an obvious coding frame), to distinguish between polymorphism and unavoidable sequencing errors. Thus, what has not been an unbearable load for evolution will be intractable for information analysis. It therefore seems vital to start sequencing programmes with genomes that are as compact as possible and amenable to genetic techniques that permit independent evaluation of gene structures and functions, in particular genomes in which the internal consistency created by the long history of evolution is accessible to computer analysis. The genomes that have been considered in this light, and which may pave the way to understanding the human genome, are discussed below.

Mammalian genomes are quite similar to each other. In order to minimize the difficulties mentioned above (split genes, archives, polymorphism etc.) one should study an organism in which the genotype can be controlled as much as possible. The laboratory mouse, an inbred strain, is a case in point. Its genome, spread through 40 chromosomes, is not unlike the human genome. Using appropriate techniques, it is possible to sort out each specific chromosome and to make DNA libraries. A further justification for studying the laboratory mouse is that it permits true reverse genetics (using embryo stem cells), so that it is possible, in principle, to replace a copy of any gene by a modified version of it. Despite the tediousness of this procedure, it has already been used for the study of many important genes. Such a technique cannot be used with humans both for obvious ethical reasons and because of the 25-year generation time. However, to sequence the mouse model still remains an irrealisable goal with the techniques currently available.

Other differentiated organisms are much simpler and have already been used as models to study embryo determination and cell differentiation. This is the case with *Drosophila melanogaster*, the paragon of genetically-studied organisms. Its genome is 20-times shorter than a mammalian genome and 30-times longer than a bacterial genome. Plants and animals are so different that the former almost certainly require independent analyses, especially in view of their unique features such as their singular defense systems, photosynthesis and nucleus–chloroplast interactions (sequence analysis of the *c.* 100 kb chloroplast genomes of the latter has already revealed several interesting features). The crucifer *Arabidopsis thaliana* genome, which is around 100 Mb, is more amenable to thorough analysis than a mammalian genome. Plans to obtain a precise map and then to sequence this genome are already being drawn up.

While the general 'blueprint' of mammals, plants or even insects is relatively well conserved, its fine details vary considerably. For this reason, Sydney Brenner proposed, more than twenty years ago, to study a nematode worm, *Coenorhabditis elegans.* This nematode could be an excellent model because its developmental blueprint is absolutely fixed both spatially and temporally (including programmed cell death). The total genome sequence is approximately 100 Mb. A sequencing programme has already been started, using an ordered library of clones from different vector systems. It is the most advanced sequencing programme for differentiated organisms, and a consortium of two institutes (the laboratories of J. Sulston in Cambridge, UK, and R. Waterston in Saint Louis, USA) now generates more than 6 Mb of sequence every year.

Finally, the analysis of the genome of the baker's yeast, *Saccharomyces cerevisiae*, has been exemplary. The programme for sequencing its genome was organized by the European Union under the efficient direction of André Goffeau who, since 1986, has tried to convince his colleagues that it is an interesting, important and valid scientific challenge (Goffeau and Boutry, 1989, *Sequencing the Yeast Genome: Why and How?* Commission of the European Communities). A measure of his success is visible in the first publication of the group, which brought together 147 scientists to sequence the 316 kb of chromosome III. This first report generated tremendous interest, and the major part of the genome (15 Mb) has now been sequenced. Two main results have already emerged from this fascinating work. On the one hand, the genome is very compact (there is very little redundant DNA), while, on the other hand, half of the genes that are expressed code for products hitherto unidentified in any organism. This reflects the considerable depth of information which remains to be uncovered and analysed.

The *E. coli* paradigm

Apart from the organisms listed above, *E. coli* K-12, was the organism which, at first sight, appeared to be best suited to sequencing programmes (Roberts, 1989, *Science* **243:** 67–168; Roberts, 1989b, *Science* **246:** 439–440; Watson, 1990, *Science* **248:** 44–51). Its genome is compact, and we possess a vast amount of information about all features of its biology. In addition, it is of considerable interest regarding issues concerning the environment, industry and medicine. Finally, it can be easily manipulated, and it can have a very short generation time. *E. coli* K-12 has a 4700 kb chromosome. In 1989, more than 1000 loci had been described at the genetic level and more than 400 kb of the chromosome had been sequenced in a series of independent analyses by a large number of laboratories (Anderson, 1989, *Nature* **338:** 283; Médigue, *et al.*, 1990, *Mol Microbiol* **4:** 169–187). In addition, Kohara and co-workers created a collection of overlapping lambda clones which covered most of the chromosome and which permitted them to generate a restriction map using eight restriction enzymes (Kohara, *et al.*, 1987, *Cell* **50:** 495–508). It was natural, therefore, to consider this organism as a priority for genome sequencing programmes. Between 1987 and 1989, the most pessimistic estimate for completion of the *E. coli* genome sequencing was well before the middle of this decade (Lewin, 1987, *Science* **235:** 747–748; Anderson, 1989, ibid.).

This pessimistic estimate turned out to be highly optimistic, however. Although more than 1800 loci have now been characterized genetically, and many genes have been sequenced (2800 kb, by mid 1995), the project is still far from complete. In my view, this is because of several factors, a major one being that underestimation of the difficulty of the programme induced competition between laboratories engrossed in sequencing, rather than collaborative effort. In addition, the *E. coli* project was only seen as a side-project to the human genome initiative. The situation was made even more frustrating by competition between countries (USA and Japan; see Swinbanks, 1987, *Nature* **328:** 195; Swinbanks, 1989, *Nature* **339:** 648; Swinbanks, 1989, *Nature* **342:** 463; Swinbanks, 1989c, *Nature* **342:** 724–725) and funding agencies (US Department of Energy and National Institutes of Health; see early reports of conflicts in Koshland, 1987, *Science* **236:** 505; Lewin, 1987, *Science* **235:** 1453; Roberts, 1987, *Science* **237:** 486–488; Ebbert, 1988, *Nature* **333:** 7; Palca, 1989, *Science* **245:** 131), which resulted in a lack of dispassionate evaluation required for the success of any genome sequencing programme.

Sequencing projects are only successful if one is able to evaluate and explore the technical processes which are needed to achieve the goal efficiently and quickly. Initially, the sequencing procedure was rather slow and tedious, and much emphasis was placed on sequencing *per se*, not on the need for sequencing an appropriate set of DNA fragments covering the complete genome. In fact, grant applications for sequencing projects estimated that sequencing itself would be the most time-consuming, rate-limiting step. An assumed level of redundancy permitted one to evaluate the amount of DNA which had to be sequenced, and this figure was compared to the output, in terms of sequences per day or per week, which could be obtained in a laboratory. For example, in 1988 it was proposed to sequence one cosmid insert (*c.* 40 kb) every week: this would have permitted a laboratory to obtain the whole sequence of the *E. coli* genome in two to three years. There were even enthusiastic suggestions that one could obtain 10 to 15 kb of sequence a day, and perhaps much more, using machines with fluorescence detection by laser or other physical techniques, which were still at the conceptual level (Roberts, 1987b, *Science* **238:** 271–273; Smith, 1993, *Science* **262:** 530–532).

Another approach, using the chemical sequencing technique of Maxam and Gilbert , was the Multiplex procedure advocated by George Church (the inventor of this astute technique). This technique would permit sequencing of the whole *E. coli* genome by direct shotgun cloning and sequencing on Multiplex gels where 20 different templates would be run together in each lane of a gel, dividing by 20 the number of runs (but not the number of hybridization steps). In 1987, the promoters of this technique were so optimistic that they thought they would be able to sequence 90% of the genome within one year (Lewin, 1987, *Science* **235:** 747–748).

Unfortunately, it was soon observed that the sequencing step was not the bottleneck. In fact, this should have been understood from the very start. Before starting to sequence, one must have a collection of DNA fragments covering the genome. From simple statistical considerations (Poisson distribution), a rule of thumb shows that chance is such that if one clones one equivalent length of a genome, about one third of the genome is not covered by the fragments. If one wishes to cover 90% of the genome one must have a collection of fragments spanning almost 10-times the genome length. It is immediately apparent that gap filling will therefore be a problem: one has to decide how one is going to fill these gaps at the start of a programme. Does one need to build up a library covering all of the genome? Of which type? This will determine the ease with which sequencing will be performed, and this is a step which must be added to the time needed to complete the programme. There are several major difficulties: some fragments will be more difficult to sequence than others, some will display compressions or sequencing artefacts, and the extent to which these will hinder the sequencing process depends on the technique used (multiplex, chemical, labelled or fluorescent dideoxy chain termination). In some cases, fragments simply cannot be cloned because they are lethal. Finally, reconstructing contigs may be difficult when regions are duplicated (e.g., regions containing insertion sequences, ribosomal RNA genes or any type of long duplication). These problems can be overcome, but this takes time: one needs to use synthetic primers to change the starting point of the sequence, or to use the polymerase chain reaction (PCR) to obtain a fragment which cannot be obtained otherwise, and one needs to use base analogues so that the anomalies in gel migration are displaced, etc. At the start of the *E. coli* sequencing programme, PCR was not a routine technique, and it is only recently that long PCR fragments can be easily obtained. In addition, PCR can and does introduce errors when copying templates; sequence verification therefore requires extra steps. Altogether, this meant that in 1988 (and even in the early nineties), one person could not sequence more than 50 kb a year without gaps and with an error rate lower than 1/2000. This is almost fifty times slower than what was proposed in most grant applications of the time!

Finally, automation was an important advance, although it required that the sequencing technique be minimally sensitive to variations in external conditions. In the late eighties and early nineties, this criterion was rarely met by techniques using fluorescence because they required extremely clean templates in order to be successful. The attempt made using the Multiplex technique failed because the gels were not clean enough to permit long readings. In addition, it would only be possible to read a large number of gels if this process could be automated: a non-trivial task that has not yet been properly solved. Furthermore, one had to generate the contigs using elaborate informatics, which does not permit one to escape the statistical presence of gaps. The result of these problems is that the fraction of the *E. coli* genome sequenced as a contig by Church's group is small. For all these reasons Fred Blattner, in Madison, decided to scale up the standard radioactive technique by using robots for sequencing reactions (Frank *et al.*, 1988, *Biotechnology* **6:** 1211–1213) and developing appropriate cloning procedures for the *E. coli* chromosome. This was certainly a reasonable bet at the time, but the actual throughput of the sequencing process was overestimated (Daniels, *et al.*, 1992, *Science* **257:** 771–778; Blattner, *et al.*, 1993, *Nucl Acids Res* **21:** 5408–5417; Burland, *et al.*, 1993, *Genomics* **16:** 551–561; Plunkett, *et al.*, 1993, *Nucl Acids Res* **21:** 3391–3398; Sofia, *et al.*, 1994, *Nucl Acids Res* **22:** 2576–2586). In Japan, the overoptimistic appreciation of the sequencing process, combined with the usual length of administrative procedures in this country, resulted in the production of a gap-filling sequence of the 0–4 min region of the chromosome which has had to be corrected many times since it was deposited in the data banks (Yura, *et al.*, 1992, *Nucl Acids Res* **20:** 3305–3308; Fujita, *et al.*, 1994, *Nucl Acids Res* **22:** 1637–1639). Finally, the sequencing of the terminus region, which was supposed to be performed by a team led by K. Isono in Kobe, did not get the appropriate support, perhaps because of the delays in obtaining the sequences that the Japanese groups had promised to their authorities (Kasai, *et al.*, 1992, *Nucl Acids Res* **20:** 6509–6515). The result of this is that we have knowledge of about 75% of the *E. coli* genome, gained by adding up all sequences produced in many laboratories (Colibri update III; Médigue, *et al.*, 1993, *Microbiol Rev* **57:** 623–654); a little less than 40% of the genome is in a single contig.

Sequencing techniques are slowly but steadily improving. In particular, the efficiency of long PCR cloning can significantly reduce the time needed to fill gaps, but this still requires appropriate controls in order to avoid PCR-induced mutations (including small deletions). It seems likely, however, that in a laboratory set up for large sequencing programmes (in particular, using fluorescence sequencing machines routinely) one person could sequence 200 kb per year. In this respect, it is interesting to note that the *H. influenzae* genome sequence (1.8 Mb) was obtained by Craig Venter's group in about one year, starting from direct shotgun fragmentation of the whole chromosome (Fleischmann, *et al.*, 1995, *Science* **269:** 496–512). This success results mainly from the use of excellent software for generating contigs and from the techniques used to fill gaps. It is, however, probably still too early to be sure that the corresponding sequence is devoid of frameshifts or of spurious duplications or translocations generated by the softwares.

Blattner's grant for sequencing the *E. coli* genome terminated in 1995, and a first evaluation of a new application that proposed the rapid completion of the sequencing process was rejected (Nowak, 1995, *Science* **267:** 172–174). However, a global reaction of the community resulted in a new peer-review process that, in the summer of 1995, awarded the project to finish sequencing the *E. coli* genome (with no annotations), to Blattner's *E. coli* Genome Center at the University of Wisconsin, Madison. The programme is to complete the sequence by going counter-clockwise, from 75 min. In parallel, a Japanese team, headed by Prof. T. Horiuchi, decided to carry on sequencing the genome, clockwise from 16 min. without official support, but with the collaboration of 10 laboratories, which use the collection of lambda clones generated by Y. Kohara. Data acquisition (approx. 20 kb per week is co-ordinated by H. Mori. Technological improvements made by R. Davies have allowed his group to clone and to begin the sequencing of over 1 Mb of the chromosome that was not sequenced by Blattner. Thus, the complete sequence should be known by the end of 1996.

Annotating the sequence

There remains a very important bottleneck in all genome programmes: annotation. In my view, this step accounts for most of the delays in Blattner's *E. coli* programme. Unfortunately, at the start of genome programmes there was (and still is) an enormous under-evaluation of the cost in terms of time and effort needed for proper annotation. Some say that annotation is not important because those who are interested will do the work, and indeed the new grant obtained by Blattner is only to sequence the remainder of the *E. coli* genome, without annotation. I feel very strongly that this is not a sensible idea; indeed it is probably dangerous and uneconomical in the long-term. If one is trying to sequence a genome, one does not wish to stay at the purely technical level required to obtain the 'naked' sequence; one wishes to know something about it. Furthermore, a naked sequence is much more error-prone than an annotated one; indeed experience proves that unannotated sequences have a much higher error rate than annotated sequences (Médigue, *et al.*, 1995, *Cooperative Computer System for Genome Sequence Analysis.* In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, in press). For example, the regions that are unannotated or poorly annotated in the *E. coli* sequence reveal a very high error rate, sometimes of the order of 1%. The reason for this is that many signals, and in particular the coding frames, can be detected through annotation, so that likely frameshifts (the major source of error, in addition to the frequent GC inversions) can be seen when annotating a fragment, leading to feedback control on data acquisition. Annotation is also very important when controlling the validity of the contigs generated automatically. Confirmation of the overall structure of the *H. influenzae* genome sequence will be a case in point. Finally, most geneticists do not possess easy means to identify regions of interest for them in a naked DNA sequence. Most of us rely on annotation by others to identify regions which are relevant to our research. This indicates that annotation by the co-ordinators of a sequencing programme is a very important contribution for the whole scientific community. The remarkable contribution, unfortunately yet unpublished, of Kenn Rudd at the NCBI (as well as the help given by Amos Bairoch), should be stressed at this point.

At present one must still annotate sequences manually. This means that one has to take segments, scan libraries with chunks of *c.* 2–10 kb with BLASTN, then BLASTX, check for possible frameshifts, and then try to identify start codons and ribosome-binding sites. In the case of *E. coli*, this permits one to annotate about 50% of the sequence relatively easily. Things are much more complex for the remaining 50%. One must identify the putative coding sequences inside open reading frames (ORFs) using various rules (such as the RNY rule or the GENEMARK program). One must then try to relate unique coding sequences to known counterparts. However, because the similarities were not found using BLAST, one must use combinations of other programs such as FASTA and BLITZ. One must then study the literature in order to check whether low similarities are significant. Further annotation permits one to identify likely promoters and transcription terminators, as well as repeated units or other regularities in the genome.

This first annotation step takes a very long time, and yet is still very incomplete. It is important, for instance, to predict protein compartmentalization on the basis of the presence of signal peptides or other known targeting and sorting signals, and to predict protein functions more accurately from multi-alignments with counterparts present in data libraries such as PIR or SWISSPROT. In the same way, it is important to identify the nature of promoters, and whether they are controlled by the usual sigma 70-containing RNA polymerase or by other sigma factors. It is also important to describe other control regions such as operators and other *cis*-acting DNA structures. In the same way, the mRNA leader sequences can often be organized in a 2-D (3-D) structure. In addition, introns must be considered. Clearly, there is no definite limit to annotation. It seems, therefore, to be of the utmost importance for the future not only of the *E. coli* genome sequencing programme, but also of all other genome sequencing programmes, to develop integrated informatics to manage not only the data and their annotations, but the methods required to make the annotations in an automatic or semi-automatic fashion (Médigue, *et al.*, 1995, *Gene COMBIS* (World Wide Web :http://www.elsevier.nl/journals/gene combis/) *Gene* **165:** GC37–GC51).

Antoine Danchin
Régulation de l'Expression Génétique, Institut Pasteur,
28 rue du Docteur Roux, 75724 Paris Cedex 15, France.
E-mail adanchin@pasteur.fr; Tel. (1) 45 68 84 41;
Fax (1) 45 68 89 48.

Received 14 August, 1995; revised 2 October, 1995;
accepted 4 October, 1995.