Curr. Issues Mol. Biol. (2003) 5: 37-42.

Genomes and Evolution

Antoine Danchin*

Génétique des Génomes Bactériens, Institut Pasteur,
28 rue du Docteur Roux, 75724 Paris Cedex 15, France
and HKU Pasteur Research Centre, 8 Sassoon Road,
Pokfulam, Hong Kong

Abstract

Genomics today involves the study of many genes at a time in order to gain an integrated picture of the cell or organism as a whole. This review considers the architecture and evolution of bacterial genomes. The many facets of large-scale functional investigation in a variety of bacteria and the search to find common rules in their dynamic and structural organization are discussed. Such rules could aid the understanding of common properties and essential differences corresponding to elusive functions, or of still unknown bacterial biotopes.

Introduction

Genomics today holds the fashionable position that molecular genetics used to occupy. Rather than work with individual genes, we now prefer to consider many genes at a time, and try to get an integrated picture of the cell or of the organism as a whole. We try to follow this state of integration by considering many facets of large-scale functional investigation in a variety of bacteria, with the explicit goal to find, in the end, common rules in their dynamic and structural organization, rules that would aid the understanding of common properties and essential differences corresponding to elusive functions, or of still unknown bacterial biotopes (Danchin 2002; 2003).

It seems indispensible to first stress the scientific reasons for genome sequencing (at the time of this article more than 700 public genome programs are under way). Scientific reasons have not always prevailed and we hope to provide here a strong conceptual justification to the sceptics who will then be enticed to know more by reading further articles on the subject (for further reading see Danchin 2002; 2003).

Naturally, the beginning of the Human Genome Programme is the case in point. It is important to note that this project was born of a political initiative. Begun in the mid-nineties as a full blown programme (and certainly much earlier, if we trace its origins) it very soon became inseparable from the commercial issues which surrounded it. This is so strong that, on March 14th 2000, Tony Blair and Bill Clinton published a joint declaration in which they "applaud the decision by scientists working on the Human Genome Project to release raw fundamental information about the human DNA sequence and its variants rapidly into the public domain". The declaration ended with an enigmatic phrase in which Blair and Clinton "commend other scientists around the world to adopt this policy" of rapid publication. It goes without saying that it is unusual for heads of state to intervene in scientists' decisions to publish. Incongruous as this is, the declaration reminded us that the Human Genome Project is based on a political initiative, not a scientific one (Kelly, 1990; Sinsheimer, 1990; Cook-Deegan, 1995). Is this the case with microbial genome projects? I hope to convince the reader that there is much more, and that genomics is in fact a revolution in the way we consider life.

This very political backgound contained an important technological insight. Genome studies are inseparable from the technique which led to DNA sequencing. Furthermore, genome sequencing has added a new dimension to the usual experimental palette we use to study life: in addition to work *in vivo* and *in vitro*, we have to work "*in silico*", to use computers (Danchin *et al.*, 1991). The enormous progress we continuously witness in genome sequencing would not have been possible without developments in computer memory and calculating speed, in parallel with improvement in sequencing techniques. As early as 1978, it had become clear that computer support was necessary to allow the scientific community to build the sequences into a continuous text which they could then interpret. A study undertaken by the Rockefeller University and the European Molecular Biology Laboratory (EMBL) at Heidelberg led to the idea of the creation of a databank for gene sequences. In fact it had even been proposed that the programme for the creation of EMBL should be the sequencing of the *Escherichia coli* genome. It became clear very early on that the possession of this information was of vital importance, with political implications. Frequent discussions, often heated, took place between Europe and the USA, to decide where these databanks would be, and how they would be structured. Who would be responsible for sequence quality – its producer or the database? Who would produce the annotations? This is clearly no small matter – a wrong annotation is tantamount to disinformation. And it is unfortunately now clear that major annotation errors are spreading via data banks through the entire scientific community. Two banks were established, in competition but also in touch with each other – one at Heidelberg, the other, the first GenBank, at one of the DOE's laboratories, the Los Alamos National Laboratory (LANL) and later at the National Center for Biotechnology Information (NCBI) in Bethesda. This was a somewhat unstable situation, but the informal association between GenBank and its European and Japanese counterparts, which had existed since 1990 and which later became official, brought stability. On the European side was the EMBL data library, first at Heidelberg, then at its outstation at Hinxton Hall, south of Cambridge, the European Bioinformatics Institute (EBI), and on the Japanese side the DNA Data Bank of Japan (DDBJ) at the National

*For correspondence. Email adanchin@pasteur.fr

Institute of Genetics (NIG) at Mishima. Effectively, there is now one single International Nucleotide Sequence Database (INSD), with three entry points at the NCBI, the EBI and the NIG.

After the first meeting on the sequencing of micro-organisms organised by David Galas at The Institute for Genomic Research, it clearly appeared that one needed a powerful computer infrastructure to develop genome programmes. David Hopwood, present at the meeting, naturally defended the idea of sequencing a *Streptomyces* genome. But at the time this was premature, in particular because it was then really difficult to sequence G+C-rich DNA. Several other important technical issues had also to be solved, and it is the "shotgun" technique, advocated by Craig Venter which demonstrated that it was possible, in a short time, to entirely sequence bacterial genomes, at least those which were neither A+T nor G+C-rich (for different reasons: the AT-rich ones lead to a large proportion of gaps because their DNA is toxic in *Escherichia coli*, while for GC-rich DNA sequencing numerous artifacts caused problems in the early days; Frangeul *et al.*, 1999). Discussions were therefore mostly about the techniques needed for the achievement of the projects, not the reasons to undertake them.

The Human Genome Programme had been explicitly conceived as a gigantic enterprise meant to construct a technological empire in the context of competition between the USA and Asia, Japan in particular. In this context, it is easy to understand the scepticism of the majority of scientists thoughout the world, althouth this is of course a mistake of scientific judgement due to the scarcity of genome programmes with a purely scientific aim. Indeed, why sequence genomes? What we shall see is that not only is it a reasonable entreprise, but it is also a revolutionary turning point in our understanding of Life (Danchin 2002; 2003). Comparing genomes with one another shows in an illuminating way that the genetic programme, which specifies the RNA and the proteins, also contains an architectural programme which specifies the organisation of the cell (and of the organism). In the same way as Molecular Biology was a conceptual revolution, genomics is an even deeper revolution (why this unfortunately fashionable oxymoron "post-genomics"? "Genomics" is enough to summarize the integrative science which analyses the nature of living organisms from their genome sequence…). The history of genomes is the history of Life itself since no organism can be reduced to a single gene, nor summarized as a single family, be it expressed as a control for the cell cycle or for the secretion machinery.

In a nutshell, Life possesses a central property of self-consistency, and experience shows that this is reflected as an integration which is directly visible in the nature of the underlying genetic programme that drives the development of all cells or organisms. Various books on Gram positive bacteria amply demonstrate the integrated properties uncovered by large scale approaches (Danchin 2002; 2003).

The Alphabetic Metaphor

The central concept often named (wrongly, because there is no religion here!) the Central Dogma of Molecular Biology, is that 'information transfer', always associated with two processes essential for life, metabolism and compartimentalization, rests on two laws. The first one is *complementarity*, which to a sequence of primordial motifs associates a complementary sequence of motifs (chain of letters) that results in exact and complete specification of the former, and the second law is *coding*, which uses a cipher allowing the rewriting of the first chain symbolized by a four letter alphabet into a second chain symbolized by a twenty letter alphabet.

Molecular biology used to rely on the alphabetic description of the genes. It is now further exploring the metaphor, in which the genome is understood as providing the complete text of the recipe allowing the construction, the development, the survival and the evolution of any living organism. Let us note here that this would be a deep mistake ("*this is not a pipe*", painted the surrealist Magritte) to mix up the book of recipe with the meal, as one does when one tries to make the public believe that the knowledge of the Human Genome sequence will allow one to cure all diseases!

Knowing the genome texts, nevertheless, is truly revolutionary because this opens up the possibility of understanding the nature of life and of its evolution through the exploration at a symbolic level (in the same way as the letters which compose the recipe are symbolic with respect to its realisation, but have a very precise link with it). What this metaphor, which comprizes an alphabetic text and a coding process, accounts for is the possibility of a true creation (i.e. sudden appearance of an entity which cannot be predicted from what was before, but only accounted for in the chain of evolution *a posteriori*), as repeatedly witnessed when analyzing the evolution of species. Comparing genomes allows the identification of those places where creation operates: the Darwinian trio Variation / Selection / Amplification makes material systems evolve, evolution *creates* new functions, which, in order to come to being, *recruit* preexisting structures — hence the "tinkering" features stressed by François Jacob. And hence the impossibility of predicting a function from the structure alone (Danchin, 1999). Naturally, because any creation is unpredictable, one will only be able to explain *a posteriori* how this or that "capture" allowed the development of that particular function. It is there that we begin to construct the means to explore, from what we know of the past, what might be conditions for new creation to come, as well as the places where it will be possible (this is a central question, for example, when one wishes to understand how and where new diseases can emerge).

The alphabetic metaphor of the genetic programme is so appropriate that one can represent the cell and its programme as a computer and its programme. That particular computer would have the particularity, in the programme it deciphers and modifies, to provide the appropriate instructions not only for its own duplication, but for the duplication of the machine itself. The most elementary machine of this type, as proposed by Alan Turing, is made of a read/write head, of a mobile "band" which passes through the head and which carries only a linear sequence of symbols, and of a mechanical device

which allows the band to go forward, go backward, stop, or read the symbols in the band, as a function of the previous readings by the head. This machine is essentially defined by the fact that it allows formal separation between the *machine* proper (the read/write head and the mechanics needed to make the band move), the *data* which set the conditions under which the programme is executed, and the *programme* itself. With this view, data and programme are in fact playing the same role: they are sequences of states which are scanned by the read/write head during the functioning of the machine.

The major conceptual question asked by this metaphor is whether one is allowed to separate the programme from the machine. As I argued elsewhere (Danchin, 1998; 2002; 2003), the experiments which are at the root of the molecular techniques of biotechnology provide us with a first indication that this separation is effective. It is demonstrated in an illuminating way in these experiments of cell cloning that have been so fashionable since 1997. Thus, the metaphor raises the question of the duality that exists between the concreteness of material systems such as cells, and the symbolic nature of the systems of Number Theory (arithmetics). This led Turing for example to invent the machine we just described. For a living organism the genetic programme leads to the real construction of the machine: in this case, everything goes as if one could smoke Magritte's pipe. This indicates that we have more to understand in a genome than what we understand when we simply analyze the genes transcription and translation. We need to identify collective, global properties. Hence the need for large-scale approaches such as transcriptome and proteome analyses.

The logic needed for an effective complementarity between matter and its symbolic representation has long been discussed by von Neumann in the mid-sixties (von Neumann, 1958). It requires some "jump" from matter to abstractness. If the programme is a model for a universal construction rule that plans the duplication of both its own description and that of the machine, then the duplication of the programme together with that of the machine becomes logically possible. But this requires that the machine and its symbolic representation (programme in the programme) be somehow separated. Thus there is a need, in addition to the genetic programme, for the existence, somewhere, of an *image of the machine.* The implication, quite strong, of this line of reasoning, is that there must be, associated with the programme itself, a representation of its environnement (*i.e.* of the cell, and of the world where it lives). This is precisely what is allowed for by the fact that the programme has a physical support, the DNA molecule. We have indeed observed, from the analysis of the bacterial genomes text, that there is a definite correlation between the order of the genes in the genome, and the architecture of the cell (Danchin *et al.*, 2000).

Bacterial Genomes and Their Architecture

It is frequently heard that the distribution of the genes in genomes is more or less random, that genomes are "fluid" entities (Stibitz and Yang, 1997). In fact, a superficial observation of the position of orthologous genes in various genomes gives the impression that they can be located anywhere, and certainly not always at the same position in different genomes (this is difficult to assess since one would need for such an investigation to locate genes with respect to appropriate reference points, but are there "centromeres" in bacteria, for example?). However, it is also clear that genes have a real tendency to form clusters: genes coding for related functions, such as those coding for the subunits of an enzyme for example, are most often located side by side, in the same transcription unit. This was the very basis of the concept of *operon*. This grouping is certainly not random. This may even sometimes give clues for the identification of an unknown function (Nitschké *et al.*, 1998). Here is an example that is rarely discussed despite its reference to the paradigm of transcription expression and control. The full meaning of the lactose operon, formed of three genes, *lacZYA*, respectively coding for beta-galactosidase, lactose permease, and lactose transacetylase long remained elusive: what is the function of transacetylase? This is particularly puzzling, knowing that the affinity of this enzyme for lactose is so poor (K_M of about 1 M, Zabin and Fowler, 1984). A logical analysis of the operon shows that if the permease is very efficient (and it is!), then the concentration of lactose in the cell may reach such a high level that it will become toxic, be it only because lactose influx will increase osmolarity to an unbearable point. One needs therefore to think of a mechanism of excretion (Liu *et al.*, 1999). But one also needs to understand that the cell has to avoid a futile cycle, so that the excreted molecule cannot be the one which enters the cell: the formation of acetyl-lactose is thus likely to be the much needed security valve!

What do we see when we pay heed to the situation? In bacteria, in general, the chromosome is replicated from a unique replication origin. This origin can thus be used as a reference point. Furthermore, in these organisms, many genes are co-transcribed in the form of operons. When one takes the origin as the reference, and the existence of operons as markers of the gene distribution, then the fluid picture of the genomes changes considerably. Indeed, one first remarks that there are almost always more genes transcribed in the direction of the movement of the replication fork than in the opposite direction (this varies much from genome to genome: organisms such as mycoplasma or A+T-rich Gram positive bacteria have a very strong transcription bias, whereas this is much less true in *Escherichia coli* for example). Second, one remarks that, quite often, the nature of the base composition in each of the chromosome strands is very different: it is G+T-rich in the direct strand and A+C-rich in the complementary strand (this composition bias, when it exists, is the same for all bacteria) (Frank and Lobry, 1999; Rocha *et al.*, 1999). If the mutation rate had time to equilibrate (which would be the case if genes really moved frequently; Rocha and Danchin, 2001) then, as shown by Sueoka, one should have within each strand an equal number of A and T, and of G and C (the second Chargaff's law), as is observed between the strands (Sueoka, 1995).

This demonstrates that genes do not move that frequently, or, more precisely that most gene moves are

incompatible with the long term survival of the species. What we see as a move represents only what remains, keeping compatibility with long term survival, keeping mainly a privileged orientation (and maybe privileged distances) with respect to the origin of replication. As a consequence, if one looks carefully, one observes in bacteria a strong constraint in the distribution of the genes with respect to the origin and terminus of replication. Furthermore, the analysis of the codon usage bias indicates that there exist many additional constraints. Indeed, we have shown in 1991 that a significant part of the *E. coli* genome corresponded to genes with an unusual codon usage bias. These genes were most probably involved in horizontal gene transfers, and they were many, forming one sixth of the genome in *E. coli* K12 (Médigue *et al.*, 1991). Since then, this observation has been made over and over again in almost every bacteria, sometimes for a considerable proportion of the genome, as is the case for example in some *Bacillus cereus* species which may thus be considered as large reservoirs for genes of unknown functions (not visible in the phenotype of the colonies on a plate, nor in the classical tests for systematic identification; Carlson and Kolsto, 1994).

At first sight, these genes are distributed randomly. One remarks however that they are more numerous near the replication terminus, which is therefore a hot spot for input of foreign DNA in a genome (the mechanics of recombination needed to resolve the knotted stucture which forms when replication terminates accounts for this phenomenon; Corre *et al.*, 1997). However laterally transferred genes appear to be present almost anywhere. There does not yet exist a careful study of their distribution, but one must remark that these “foreign” genes are not randomly placed. As a matter of fact, when one goes from one strain to another one (for example comparing *E. coli* K12 and *E. coli* O157H7), one remarks that the foreign genes differ, but that they are placed at the same position in the chromosome, often near tRNA genes (Hou, 1999; Karch *et al.*, 1999; Blum-Oehler *et al.*, 2000; Hacker and Kaper, 2000; Al-Hasani *et al.*, 2001). This, again, is compatible with the notion of a strong architectural constraint in the gene distribution. It is as if one witnessed a rigid organisation of the chromosome, interspersed with a certain number of hot spots where foreign genes can insert.

Careful study of the codon usage bias further points to the existence of strong constraints in the gene order. Indeed, when one analyzes the genes within an operon, one usually observes that the corresponding codon usage bias is the same in its different genes. However, and this is remarkable, when one compares two organisms where genes are clustered into an operon in one of them while the corresponding genes are dispersed in the other, one finds that the codon usage bias is preserved, even when the the genes are no longer clustered but located distantly (Rocha *et al.*, 2000a). How can we account for this observation? The usage of a particular codon assumes that a particular transfer RNA is used. It is therefore easy to see that repeating the same codon in a gene tends to maintain the cognate tRNA local concentration at a high level (in the vicinity of the ribosome that translates the messenger RNA). A large bias in the codon usage implies therefore a local enrichment in some tRNAs, while in the same local environment the corresponding ribosome is depleted in other tRNAs. A polycistronic operon is usually translated by the same ribosomes (Petersen *et al.*, 1978). One understands therefore that the genes it is made of have the same codon usage bias (otherwise this would introduce a pause in the translation process, needed for the appropriate tRNA to become available: a fact probably used for folding proteins; Thanaraj and Argos, 1996). But the fact that genes distant in the chromosome share the same bias, strongly suggests that the corresponding mRNAs are sitting next to each other, and are translated from the same or very closely packed ribosomes… This indicates that there is a relation between the position of some ribosomes in the cell, and the position of some genes in the chromosome. There is a map of the cell in the chromosome.

A complementary observation strengthens this hypothesis considerably. The analysis of the codon usage bias in orthologous genes in organisms as different as *E. coli* and *B. subtilis*, shows a similar deviation of the bias with respect to its average value (different in both bacteria) for genes with orthologous functions, indicative of the existence of a common selection pressure in these quite different organisms (Danchin *et al.*, 2000). Apart from an architectural selection pressure what might be the cause of such a strong correlation? Naturally, for selection pressure to operate, one needs a constant and repeated action from the environment, and this can only happen for genes that are highly expressed, in conditions where the progeny of the cell is large (as happens during the exponential growth phase), so as to ensure amplification of the tiny differences in selection. The relevant genes are therefore those which are expressed at high level during exponential growth. Many of these genes are those which constitute the core of life, being responsible for the transcription and the translation of the genetic programme or for the general management of energy. Indeed, these genes are usually clustered into well recognizable large operons. The first step in the analysis of all newly obtained genome sequence should therefore consist in the fine analysis of the conservation of the genes neighborhoods. It is highly probable that, in the future, *specific* laws of this organisation will be uncovered. This becomes therefore a well defined goal for genomics. The investigation of the role, position and architecture of pathogenicity islands (Dobrindt and Reidl, 2000; Hacker and Kaper, 2000) is a first step in this exploration.

Evolution of Genomes

If one understands which genes are submitted to this architectural selection pressure (one must add the genes of the core intermediary metabolism) it is necessary to attempt to identify the nature of the selection pressure that maintains them organized with respect to each other. To this aim we need to think that this probably already happened at the origin of the first cells, and persisted in present day organisms despite the very large time which elapsed since then.

Many scientists asked questions about the origin of life, and some understood that the hypothesis of a prebiotic soup (the very soup that Pasteur demonstrated to be incapable of spontaneous generation!) would act as a poisoned broth unable to allow life to be born (Danchin, 1989). It was necessary therefore to propose other scenarios. Steven Benner used to state that "*arguments that attempt to extrapolate from modern biochemistry back to the origin of life are futile*" (Benner *et al.*, 1987). However Granick first (Granick, 1957), followed by Wächtershäuser who advocated "*a reconstruction of precursor pathways by retrodiction from extant pathways*" (Wächtershäuser, 1988), have shown that the most plausible one was to considerer first extant metabolic processes (one must construct the machine and maintain it) and subsequently to make the hypothesis that, despite the many years of evolution, some characters reminiscent of the origin were still present today. Briefly, the most plausible is to think that a selective process, which kept mainly charged molecules at the surface of solids (hence the ubiquitous character of the phosphate and carboxylate groups in anabolic metabolites) was the initial driving force. Both authors, furthermore, also thought that, as electron transfers are central to the management of energy, iron and sulfur are at the core of the primaeval metabolism (Wächtershäuser, 1992), sulfur being also a remarkable element which allows the transfer of all kinds of chemical groups (Danchin, 1989; Sekowska *et al.*, 2000).

The current working hypothesis is that metabolism organized itself around nuclei of iron-sulfur clusters, and more generally around sulfur metabolism, at the time when the first nucleotides were invented, together with the first coenzymes, then the first membranes and the first nucleic acids. It is therefore necessary to explore first within the genomes (and in particular within those of prototrophic organisms, because they are unlikely to result from the degradation of richer ancestral genomes), those genes which make these essential metabolic pathways. In *E. coli*, *in silico* genome analysis shows that the genes of sulfur metabolism are not randomly distributed but form islands (Rocha *et al.*, 2000b). One does not know yet the homologous genes in other bacteria, but with the little information already in our possession it appears clearly that, here too, these genes form islands.

Where does the corresponding selective constraints originate from? A simple observation allows one to propose a conjecture to account for it. Most genomes contain a very large proportion of unknown genes, or of genes similar to genes of unknown function. This observation, made simultaneously by the consortium sequencing the yeast chromosome III and the scientists sequencing the first long fragment of the genome of *B. subtilis*, was quite unexpected (presented in 1991 at the European Union meeting organized in Elounda, Greece). Indeed, the existence of many mutant types with the identification of the cognate genes let everybody assume at this time that one knew all the major protein categories with their associated functions. What are these genes? Some thought initially that they corresponded to unknown regulatory systems, to unexpected cell compartments or to new functions, such as clocks or timers. However it appears that as we progress in uncovering new functions, many among these genes code for … enzymes! How then did they escape attention? The explanation, perhaps, is simple: the biochemical study of enzymes is readily available in the laboratory using pure products with straightforward ways to measure their concentration and availability. Two classes of molecules resist these experimental constraints, the gases and the radicals. Indeed many of these "new" genes code for functions involving gases and radicals. This leads to an important question: gases diffuse fast, and many are quite reactive: putting H_2S and O_2 together (with an appropriate catalyst) makes them react violently.

How does the cell cope with this situation? The answer provides us with the missing link between the architecture of the cell and the distribution of the genes in the genome: what led to the organized clustering of genes in genomes is the selection pressure due to the reactivity and diffusibility of gases and radicals. This selection pressure had a major impact, there must exist a link between the chromosome, and the location of the product of translation in the cell. While this is easily visualized in procaryotes (because transcription is not separated from translation, allowing an organized network of mRNAs to form in relation with the construction of protein complexes), this is much more difficult to visualize when there is a nucleus. One is thus led to ask whether the role of introns is not, at least in part, to allow the formation of such a functional relation (if splicing is performed not in a more or less random way in the nucleus, but at precise loci associated with nuclear pores well placed at the nuclear envelope). But this is another story: we are concerned here only with bacteria.

Conclusion

As just discussed, we now face a new picture of the organisation of bacterial genomes. Of course this picture should be completed by studies in the domain of other bacteria such as the Archea, or the bacteria filled with complex membrane structures such as the cyanobacteria. The situation of G+C rich Gram positives is particularly interesting since it comprises bacteria growing very slowly such as *Mycobacterium tuberculosis* or *M. leprae*, as well as bacteria which differentiate into complex structures such as the *Streptomyces* sp. At the date when this introduction was written Nikos Kyrpides (http://wit.integratedgenomics.com/GOLD/) identifies 717 genome sequecing programmes (among which 139 are already complete and published). We are just beginning to investigate the surface of the information they contain. Recently published reviews (Danchin, 2002) present focused views on some of the topics mentioned above. Let us hope that this will be used by our colleagues as a starting point for exploring further the deep meaning of genome sequences and above all couple it to biological meaning.

References

Al-Hasani, K., Adler, B., Rajakumar, K., and Sakellaris, H. 2001. Distribution and structural variation of the she pathogenicity island in enteric bacterial pathogens. J.

Med. Microbiol. 50: 780-786.
Benner, S.A., Allemann, R.K., Ellington, A.D., Ge, L., Glasfeld, A., Leanz, G.F., Krauch, T., MacPherson, L.J., Moroney, S., Piccirilli, J.A., and et al. 1987. Natural selection, protein engineering, and the last riboorganism: rational model building in biochemistry. Cold Spring Harb. Symp. Quant. Biol. 52: 53-63.
Blum-Oehler, G., Dobrindt, U., Janke, B., Nagy, G., Piechaczek, K., and Hacker, J. 2000. Pathogenicity islands of uropathogenic E. coli and evolution of virulence. Adv. Exp. Med. Biol. 485: 25-32.
Carlson, C.R., and Kolsto, A.B. 1994. A small (2.4 Mb) Bacillus cereus chromosome corresponds to a conserved region of a larger (5.3 Mb) Bacillus cereus chromosome. Mol. Microbiol. 13: 161-169.
Cook-Deegan, R.M. 1995. Origins of the Human Genome Project. WWW Article: http: //www.fplc.edu/risk/vol5/spring/cookdeeg.htm.
Corre, J., Cornet, F., Patte, J., and Louarn, J.M. 1997. Unraveling a region-specific hyper-recombination phenomenon: genetic control and modalities of terminal recombination in Escherichia coli. Genetics 147: 979-989.
Danchin, A. 1989. Homeotopic transformation and the origin of translation. Prog. Biophys. Mol. Biol. 54: 81-86.
Danchin, A. 1998. La Barque de Delphes. Odile Jacob, Paris. (*Translation:* The Delphic Boat. Harvard University Press, 2003.)
Danchin, A. 1999. From protein sequence to function. Curr. Opin. Struct. Biol. 9: 363-367.
Danchin, A. 2002. Genomics of GC-Rich Gram-Positive Bacteria. Caister Academic Press, Wymondham, UK.
Danchin, A. 2003. The Delphic Boat. What Genomes Tell Us. Harvard University Press, Cambridge, USA.
Danchin, A., Guerdoux-Jamet, P., Moszer, I., and Nitschké, P. 2000. Mapping the bacterial cell architecture into the chromosome. Philos. Trans. R. Soc. Lond. B Biol. Sci. 355: 179-190.
Danchin, A., Médigue, C., Gascuel, O., Soldano, H., and Hénaut, A. 1991. From data banks to data bases. Res. Microbiol. 142: 913-916.
Dobrindt, U., and Reidl, J. 2000. Pathogenicity islands and phage conversion: evolutionary aspects of bacterial pathogenesis. Int. J. Med. Microbiol. 290: 519-527.
Frangeul, L., Nelson, K.E., Buchrieser, C., Danchin, A., Glaser, P., and Kunst, F. 1999. Cloning and assembly strategies in microbial genome projects. Microbiology 145: 2625-2634.
Frank, A.C., and Lobry, J.R. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238: 65-77.
Granick, S. 1957. Speculations on the origin and evolution of photosynthesis. Annals New York Acad. Sci. 69: 292-308.
Hacker, J., and Kaper, J.B. 2000. Pathogenicity islands and the evolution of microbes. Annu. Rev. Microbiol. 54: 641-679.
Hou, Y.M. 1999. Transfer RNAs and pathogenicity islands. Trends Biochem. Sci. 24: 295-298.
Karch, H., Schubert, S., Zhang, D., Zhang, W., Schmidt, H., Olschlager, T., and Hacker, J. 1999. A genomic island, termed high-pathogenicity island, is present in certain non-O157 Shiga toxin-producing Escherichia coli clonal lineages. Infect. Immun. 67: 5994-6001.
Kelly, E.H. 1990. The Human Genome Initiative: a different type of research. Faseb J. 4: 1423-1424.
Liu, J.Y., Miller, P.F., Willard, J., and Olson, E.R. 1999. Functional and biochemical characterization of Escherichia coli sugar efflux transporters. J. Biol. Chem. 274: 22977-22984.
Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., and Danchin, A. 1991. Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol 222: 851-856.
Nitschké, P., Guerdoux-Jamet, P., Chiapello, H., Faroux, G., Hénaut, C., Hénaut, A., and Danchin, A. 1998. Indigo: a World-Wide-Web review of genomes and gene functions. FEMS Microbiol. Rev. 22: 207-227.
Petersen, H.U., Joseph, E., Ullmann, A., and Danchin, A. 1978. Formylation of initiator tRNA methionine in procaryotic protein synthesis: in vivo polarity in lactose operon expression. J. Bacteriol. 135: 453-459.
Rocha, E.P., and Danchin, A. 2001. Ongoing evolution of strand composition in bacterial genomes. Mol. Biol. Evol. 18: 1789-1799.
Rocha, E.P., Danchin, A., and Viari, A. 1999. Universal replication biases in bacteria. Mol Microbiol 32: 11-16.
Rocha, E.P., Guerdoux-Jamet, P., Moszer, I., Viari, A., and Danchin, A. 2000a. Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. J. Biotechnol. 78: 209-219.
Rocha, E.P., Sekowska, A., and Danchin, A. 2000b. Sulphur islands in the Escherichia coli genome: markers of the cell's architecture? FEBS Lett. 476: 8-11.
Sekowska, A., Kung, H.F., and Danchin, A. 2000. Sulfur metabolism in Escherichia coli and related bacteria: facts and fiction. J Mol Microbiol Biotechnol 2: 145-177.
Sinsheimer, R.L. 1990. Human genome initiative. Science 249: 1359.
Stibitz, S., and Yang, M.S. 1997. Genomic fluidity of Bordetella pertussis assessed by a new method for chromosomal mapping. J. Bacteriol. 179: 5820-5826.
Sueoka, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40: 318-325.
Thanaraj, T.A., and Argos, P. 1996. Ribosome-mediated translational pause and protein domain organization. Protein Sci. 5: 1594-1612.
von Neumann, J. 1958 (reed. 1979). The Computer and the Brain. Yale University Press, New Haven.
Wächtershäuser, G. 1988. Before enzymes and templates: theory of surface metabolism. Microbiol Rev 52: 452-484.
Wächtershäuser, G. 1992. Groundworks for an evolutionary biochemistry: the iron-sulphur world. Prog. Biophys. Mol. Biol. 58: 85-201.
Zabin, I., and Fowler, A.V. 1984. Purification of thiogalactoside transacetylase by affinity chromatography. Anal. Biochem. 136: 493-496.