THE ROYAL SOCIETY

Mapping the bacterial cell architecture into the chromosome

Antoine Danchin[1]*, Pascale Guerdoux-Jamet[2], Ivan Moszer[1] and Patrick Nitschké[3]

[1]*Régulation de l'Expression Génétique, Institut Pasteur, 28 Rue du Docteur Roux, 75724 Paris Cedex 15, France*
[2]*INSERM U49 Hôpital Pontchaillou, 35033 Rennes Cedex, France*
[3]*Université de Versailles Saint Quentin, Laboratoire Génome et Informatique, 45 Avenue des États Unis, 78000 Versailles, France*

A genome is not a simple collection of genes. We propose here that it can be viewed as being organized as a 'celluloculus' similar to the homunculus of preformists, but pertaining to the category of programmes (or algorithms) rather than to that of architectures or structures: a significant correlation exists between the distribution of genes along the chromosome and the physical architecture of the cell. We review here data supporting this observation, stressing physical constraints operating on the cell's architecture and dynamics, and their consequences in terms of gene and genome structure. If such a correlation exists, it derives from some selection pressure: simple and general physical principles acting at the level of the cell structure are discussed. As a first case in point we see the piling up of planar modules as a stable, entropy-driven, architectural principle that could be at the root of the coupling between the architecture of the cell and the location of genes at specific places in the chromosome. We propose that the specific organization of certain genes whose products have a general tendency to form easily planar modules is a general motor for architectural organization in the bacterial cell. A second mechanism, operating at the transcription level, is described that could account for the efficient building up of complex structures. As an organizing principle we suggest that exploration by biological polymers of the vast space of possible conformation states is constrained by anchoring points. In particular, we suggest that transcription does not always allow the 5′-end of the transcript to go free and explore the many conformations available, but that, in many cases, it remains linked to the transcribing RNA polymerase complex in such a way that loops of RNA, rather than threads with a free end, explore the surrounding medium. In bacteria, extension of the loops throughout the cytoplasm would therefore be mediated by the *de novo* synthesis of ribosomes in growing cells. Termination of transcription and mRNA turnover would accordingly be expected to be controlled by sequence features at both the 3′- and 5′-ends of the molecule. These concepts are discussed taking into account *in vitro* analysis of genome sequences and experimental data about cell compartmentalization, mRNA folding and turnover, as well as known structural features of protein and membrane complexes.

Keywords: mesoscopic scale; neighbourhood; genome style; pathogenicity islands; codon usage; translation

1. INTRODUCTION

Three thousand years ago the Pythia in Delphi answered enigmas asked by visitors and predicted their fate. She had the habit of answering them by asking questions in her turn. One of her questions used the following paradox: 'I have a boat made of oak planks. As I keep using the boat, its planks rot one after the other. At some time no original plank is still in the boat: is it the same boat?' The owner will undoubtedly answer, yes. And everybody will accept that he is right. This is quite puzzling, however: although a material thing, the boat cannot be reduced to the matter of the boat. It is something else, much more interesting than a heap of planks, that orders the matter of the planks: the relationships between the planks make the map of the boat. In a much similar way, and in contrast to the habits given to biologists by the domination of (bio)chemistry, the study of life should never be restricted to the study of objects, but must, preferably, study their relationships (Danchin 1998).

Because they are the blueprints of life, genomes cannot be considered as simple collections of genes. They are much more. How can we have access to the relevant features of their organization? Despite the very primitive and sketchy way in which genomes are annotated—in general, lists of Blast and Fasta searches with truncated annotations—one can isolate from the current flow of genome sequences two contrasted images. At first sight, genes appear to be distributed randomly along the chromosome; in sharp contrast, their organization into operons or pathogenicity islands suggests that related functions share physical proximity. For example, the rRNA genes are clustered near the origin of replication in all fast-growing bacteria, and it has been proposed that organization of these genes in the genome is linked to the bacterial niche colonization (Ginard *et al*. 1997). In the same way several operons, such as ribosomal protein operons or the operon directing synthesis of the

*Author for correspondence (adanchin@pasteur.fr).

Received 18 March 1999 *Accepted* 25 March 1999

membrane ATP synthase, are conserved in very distant bacteria. In order to try and understand genome organization, we must therefore explore the distribution of genes along the chromosome. A way to do this is to generalize the concept of gene neighbourhood to many more types of vicinities than their succession in the genome. Using this simple mode of inductive reasoning we shall discuss data and processes that strongly suggest that bacterial genomes are organized in a way directly correlated to the organization of the cell's architecture and dynamics.

2. A FIRST METHODOLOGICAL PRINCIPLE: THE CONCEPT OF NEIGHBOURHOOD AS A HELP FOR INDUCTIVE REASONING

The main idea underlying our approach is that the biological objects making a cell alive cannot be isolated from each other: biology must be described more as a science of relationships between objects than as a science describing objects. To study these relationships, we used the concept of 'neighbourhood' to organize the knowledge we have on model bacteria, *Escherichia coli* and *Bacillus subtilis*. This concept can be visualized as making reference in a broad sense to all the items, of all possible kinds, that can be related to a given item. Because we study the genomic text, we choose genes as the core items. For a given gene we constructed lists of 'neighbours' based on links of several possible categories (Nitschké *et al*. 1998). This concept of neighbourhood is very wide: it pertains to the category of notions that John Myhill named 'prospective characters' (Myhill 1952). As a matter of fact a discovery is often made when one establishes the connection between two pieces of data that are not obviously connected by some causality. Inductive exploration will consist of finding all neighbours of each gene. Here, 'neighbour' has the largest possible meaning. This is not simply a geometrical or structural notion. In the present context, each gene's neighbourhood is meant to illuminate specifically the context of a gene, looking for its function as bringing together the objects of the neighbourhood. A first natural neighbourhood is proximity in the chromosome: operons or pathogenicity islands show that genes neighbouring each other are often functionally related. We shall explore below why two genes may be neighbours because they use the genetic code in the same way: one can study all genes that belong to the same neighbourhood in the cloud of points describing the codon usage of all the genes of the organism. From the methodological standpoint this requires construction of neighbourhood files (conveniently available to scientists in databases: a field of choice for bioinformatics). We shall use here the neighbourhood data collected on the server Indigo (http://indigo.genetique.uvsq.fr).

3. A SECOND METHODOLOGICAL PRINCIPLE: THE CONCEPT OF MODEL

A genome is a text written with an alphabet of four letters. Of course one can study the text by analysing the presence of different words (oligonucleotides) or more complex letter motifs, and try to find out whether these words have some meaning. As is usual in linguistics, where syntax and semantics are separated from each other, meaning (another 'prospective' character of the Myhill type) is an important concept, related to the typically biological concept of function. The meaning of a word has to be placed in a functional perspective. In the absence of actual knowledge about all the functions in a cell the only way to investigate meaning is to start by assuming simply a random distribution of the bases along the chromosome. However, because this concerns finite sets, nothing will tell whether a highly biased word (either extremely rare or extremely frequent) is actually meaningful. In order to have a better insight one must compare the real case of a chromosome with a model, where all the existing functional knowledge has been incorporated. In general, using models is a very efficient way to elaborate on previously existing knowledge, much more general than the usually used concept of consensus sequence (Hénaut *et al*. 1996; Sagot 1994; Sagot & Myers 1998). A simple approach consists of comparing real sequences to model sequences built using the pre-existing knowledge accumulated about them. New signals are identified by comparing the real sequence with that of the model generated by an algorithm taking into account all the initial knowledge on that sequence. For example, in order to identify abnormal distributions of tetranucleotide motifs, one must devise means to permit one to subtract the contribution of the previous knowledge about the chromosome from the complete knowledge provided by the real sequence, so that a process of 'deconvolution' (i.e. unmixing of the various constraints that operate on coding sequences, for instance) may reveal new prominent features in the distribution of motifs (any kind of motif or distribution organization of motifs can be considered, taking into account the biological imagination of the investigator).

To achieve this goal, Hénaut and co-workers used features of coding sequences, their codon usage and the bias introduced by the amino-acid composition of the proteins, and constructed a model of the chromosome for *E. coli* and *B. subtilis*. Deconvolution was achieved by comparison of the real chromosome with the model counterpart. To solve the problem raised by the presence of unusually biased words (such as AGCT in *B. subtilis*), the coding sequences were simulated using a Markov process preserving the doublet codon frequency and accordingly preserving the frequency of oligonucleotides strictly comprising two consecutive codons (i.e. the mono-, di-, tri- and tetranucleotide frequencies). Intergenic regions were represented by random sequences in which the dinucleotide frequency of the corresponding regions found in the real chromosome was preserved (Karlin *et al*. 1997). This permitted analysis of GATC and AGCT frequency and prediction of the presence of regulatory sites comprising GATC motifs that monitored transition from the absence to the presence of oxygen in *E. coli* (Hénaut *et al*. 1996). This approach also led to the surprising prediction that genes, orthologous in *E. coli* and *B. subtilis*, might be expressed in a differential way using a process of translation hopping, due to the sliding of the mRNA between pairs of AGCU sites in the ribosome (Hénaut *et al*. 1998). This leads us to now consider the physical constraints underlying gene expression in bacteria.

4. SOME PHYSICAL PREREQUISITES FOR THE CONSTRUCTION OF A CELL

(a) *Planar layers*

For a physical system, the simplest way to evolve is to follow the arrow of time, to increase its entropy. Increase in entropy is therefore considered as the driving force for exploration of all types of space and energy levels as they become available. This can lead either to order or to disorder, according to circumstances. Note that this sharply contrasts with the widely spread opinion that the world evolves spontaneously towards disorder (Danchin 1987). Life results from a selective process. It is the time stability (the survival) of living organisms that tells us whether entropy has resulted in order or in disorder. In water, entropy is the driving force for the construction of many a biological structure: this physical parameter is at the root of the universal formation of helices, it allows the folding of proteins and the formation of viral capsids, it organizes membranes into bilayer structures and yields other complex biological structures. This brings to our attention that the largest increase in entropy of a molecular complex in water is when the surface to volume ratio is the highest. This is especially the case of planar structures, and is minimized in spherical structures. Indeed, when a planar structure is formed it orders the water molecules on both its faces. As a consequence, if this plane meets another one, it will lose one layer of water molecules, and for this reason, stick there. Formation of planar layers should therefore be a very strong organizing principle at the cell level.

But, is it possible to find out, just knowing the genomic text, whether a gene product will form such layers, whether it simply forms hexagons, for example? Unfortunately not: this is indeed even more unlikely than to think that an amino-acid sequence could tell us exactly the fold of a protein, without knowing pre-existing folds. Curiously many have thought that this might be the case, taking a single example, pancreatic RNase, as the paradigm. However, whereas pancreatic RNase would indeed fold after being unfolded, because selection isolated it for that very process (it is secreted in bile salts), this should never have been accepted as the paradigm of protein folding: the plain translation of a genomic text cannot (and will not) give us the three-dimensional structure of most proteins. This unfortunate example advocates an input of biological knowledge in addition to the genome sequence in order to try to understand gene and genome functions. In what follows we shall endeavour to comply with this important constraint, and refrain from drawing more conclusions from the genomic text than can reasonably be done. However, we shall see that the organizing principles of genomes may help us couple the results of experimental biological data to the knowledge of the general organization of genes in genomes.

Of course this hypothesis requires support by experimental data. Unfortunately the physical scale at which it is significant, the mesoscopic scale, is just outside the usual physical means that can be used to explore living organisms. Light photon microscopy has a resolution limited to *ca.* 500 nm by the smallest wavelength that can be used. In contrast, electron microscopy can visualize objects in the nanometre range. However, this is at a cost: electron microscopy has to fix the objects, using various techniques, including freezing, which disrupt the real structure that should be analysed. This technique can nevertheless give precious information about cell structures. The specific case of an enzyme which is known to be a hexamer, uridylate kinase, has been investigated. It poses an interesting question about compartmentalization of cell metabolites. This enzyme has features typical of soluble cytoplasmic proteins. However, it makes UDP that can be recognized by ribonucleoside diphosphate reductase, and therefore could ultimately lead to incorporation of uracil into DNA instead of thymine, thus posing a challenging DNA proofreading problem to the cell (el-Hajj *et al.* 1988; Nilsen *et al.* 1995; Weiss & el-Hajj 1986). The interesting electron microscopy observations with this enzyme suggests that that it is localized under the cytoplasmic membrane, thus solving the compartmentalization problem (Landais *et al.* 1999). It remains to be seen whether this corresponds to the formation of planar structures, but the hexameric structure of the enzyme is certainly compatible with such a hypothesis.

(b) *A DNA network*

Many physical constraints other than formation of planar structures cooperate in the organization of the cell. Let us consider the physics of polymers. The typical *E. coli* cell is 0.5–1 μm in length. *B. subtilis* is slightly longer (4 μm). And these cells must accommodate a genome of *ca.* 4.5 million bp, i.e. if stretched out, a 1.5 mm-long molecule. What do we know about the organization of DNA in the cell? Polymer statistics show that polymers fold randomly, as a function of a reference length related to the chemical structure of the polymer, the persistence length, that tells the average length after which orientation of the initial more or less rigid orientation of the segment considered has been lost (Delrow *et al.* 1997). For charged polymers the persistence length varies very rapidly with the presence of screening charges. Thus, in the absence of ions, DNA is very rigid (Schlick *et al.* 1994), and in the presence of physiological concentration of ions its persistence length is of the order of 50 nm (150 bp) (it is somewhat longer for the more rigid double-stranded RNA molecules) (Kebbekus *et al.* 1995) but much shorter for single strands (Rivetti *et al.* 1998; Zacharias & Hagerman 1996).

A randomly coiled DNA molecule of the length of the *E. coli* genome, at physiological salt concentration, would have a diameter of *ca.* 10 μm, ten times more than the diameter of the cell. Superordered structures of DNA are therefore to be considered to account for the packaging of DNA in the cell. This includes supercoiling, domain structure, and attachment to specific sites. The question is to know whether these physical constraints are reflected in the genome sequence. If they are, then we expect that this should be marked as specific binding sites, probably regularly spaced, along the DNA, or that the replication machinery should have its normal function probably slightly altered at places where physical constraints operate. Unless of extremely negative impact on the fitness of the organism, this should not be linked to the meaning of DNA, and therefore should not be related to the functions encoded. The distribution of local anomalies in dinucleotide or trinucleotide frequency, in sliding

windows of 5000 bp, was investigated in yeast in order to explore whether there exist landmarks of such processes. A significant distribution bias was found but it was difficult to relate it to an architectural property of the chromosomes (Ollivier *et al*. 1995; Rivals *et al*. 1997). At this point, finally, it is important to remark that DNA is packed into a very small compartment (and this may be a reason for the existence of a nucleus in eukaryotic cells): this strongly limits the available entropy-driven states of the molecule. This means that the degrees of freedom offered to DNA increase when the compartment grows (the cell or the nucleus). As a consequence there is a spontaneous (entropy-driven) tendency of replicating DNA to occupy the new space offered by cell growth, creating a natural process for DNA segregation.

(c) *RNA threads and loops*

In contrast to the situation with double strands, the persistent length of single-stranded DNA or RNA is much shorter, permitting tighter packing (Kebbekus *et al*. 1995; Rivetti *et al*. 1998). But the folding problem of these long polymers is that they have so many possible states that any organization of the cell architecture would be precluded if they were freely able to diffuse. Freely moving long polymers would rapidly tangle into an unsortable bulk of knotted structures, even if they diffused through an organized lattice (such as the ribosome lattice). Providing anchoring points is a way to drastically lower the number of states, as it can be seen in the fact that hair, unless very long, does not form knots. What would be the situation when 1000 transcripts are synthesized simultaneously? A single anchoring point (as is assumed in the general models of transcription) would restrict the number of explored states, but it might not be sufficient to restrict drastically the number of states that transcripts would explore (see uncombed long hair), except for preventing the formation of knots. In contrast, two anchoring points instead of only one would limit exploration of possible states to a manageable number. How could this be achieved? The most simple answer is to consider that ribosomes are organized as a more or less fixed lattice, and that as nascent RNA comes off DNA it is pulled by a first ribosome, then by the next one, as in a threading machine. However, one has to consider the means that would organize the ribosome lattice itself. How could it be constructed?

(i) *First model: trapping complexes*

There is no reason for the physical organization of the cell to follow the genetic information flow, going from DNA to RNA and from RNA to proteins. This purely conceptual view, although spread in university textbooks, is utterly unrealistic. In contrast, it is most probably the structure of the ribosome network that organizes the mechanics of gene expression: there we find most of the cell's inertia, and there is consumed the major part of the cell's energy. One should therefore consider this network as fixed, and orchestrating transcription by 'pulling' mRNAs off their DNA template, for which only initiation is purely controlled at the DNA level. Because each mRNA molecule is translated 10–20 times (see, for example, Kryzek & Rogers 1976), translation is the main mechanical engine for gene expression as well as for cell construction. That this is so is substantiated by the fact that, associated to the ribosome, and in addition to the energy-rich bond of the aminoacyl-tRNA, there exists a protein (elongation factor EF-G) which uses the energy-rich bond of GTP in order to elongate the polypeptide chain during protein synthesis, resulting in most of the energy spent in the cell being consumed in the process of translation. In fact, the energy which is locally consumed as GTP hydrolysis is so high that it could hardly operate at a pace faster than four or five codons per amino acid per second, without resulting in a requirement for a substantial increase in the local thermal energy dissipation. As soon as a mRNA primer comes off the DNA surface where it is synthesized by RNA polymerase, it is captured by a ribosome that scans for the translation initiation codon, and further uncoils the mRNA, itself unfolded by RNA polymerase as it is copied on a DNA strand. This process makes the DNA move and brings to its surface further genes ready for transcription. The mRNA passes from one ribosome to the next one, controlling synthesis at each ribosome of the protein it specifies (this allows an even distribution of the proteins in the cell, without requiring a three-dimensional diffusion pattern—a very slow process—but using instead linear diffusion of the mRNA—a much faster process). Finally, as an appropriate signal reaches the ribosome (this could be the leader sequence of another mRNA) at the same time as the translating messenger, it triggers degradation of the first one, thus ending its expression.

There is a strong argument in favour of this scenario. Because it rests on the assumption that messengers that are translated by the ribosomes play a major role in the cell dynamics (through translation, which is an energy-consuming process), it predicts that they are submitted to a mechanical tension, uncoupling translation from transcription. This leads to a very simple prediction: it must happen that the mRNA thread is not completed, because the RNA-polymerase–mRNA complex dissociates from its template. If this is true, what happens next? A truncated mRNA captured by a ribosome starts to be translated and a polypeptide chain elongates as the corresponding codons file one after the other. However, the situation at the break point is very different from what is normally occurring, since the ribosome does not find a termination codon to be recognized. If this situation really happens with a significant frequency, there must exist an appropriate mechanism permitting the ribosome to tackle the problem. This is the more necessary because many proteins are part of multiprotein or RNA–protein complexes, so that fragments of proteins—exactly what would be synthesized by a truncated mRNA—would take the place of the intact protein in the complex, and lead to a loss of function (in particular the truncated one would lead to a negative dominant behaviour) (Akiyama & Ito 1995; Charlier *et al*. 1995; Turner *et al*. 1997; Ueguchi *et al*. 1997; Williams *et al*. 1996). Indeed, it has been discovered in both *E. coli* and *B. subtilis* (and the cognate gene is present in all known bacterial genomes), that a specific RNA molecule, tmRNA (formerly 10Sa RNA), acts as a tRNA molecule charged with an alanine residue, and subsequently takes the place of the missing mRNA, putting in its place a set of ten codons, followed by a translation termination codon

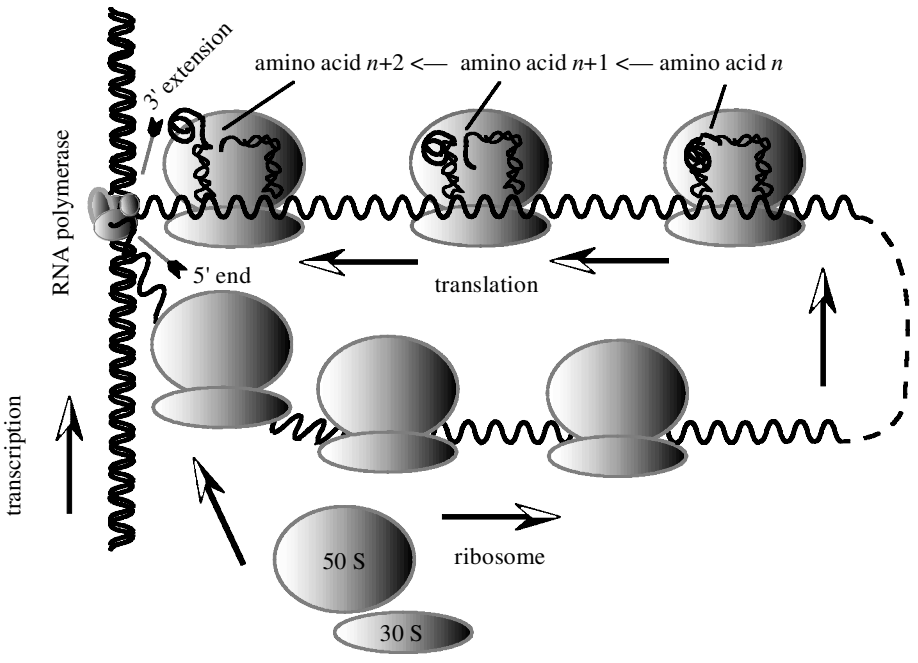

Figure 1. Overall view of the formation of a transcription loop, with translating ribosomes. This assumes that the 5′-terminus of the mRNA binds specifically to a subunit of RNA polymerase. Note that the 5′-nucleotide is a triphosphate, permitting selection of a specific binding site that would not recognize RNA molecules submitted to endonucleolytic cleavage.

(Himeno *et al*. 1997*a*,*b*; Keiler *et al*. 1996; Muto *et al*. 1996, 1998; Watanabe *et al*. 1998). This results in adding a carboxy-terminal tag to the protein, made of 11 residues, always the same. These tagged proteins are further directed to a proteolytic complex, comprising the ClpA and ClpX protein, where they are degraded (Gottesman *et al*. 1998; Herman *et al*. 1998).

(ii) *Second model: 5′-binding of nascent RNA to RNA polymerase*

Transcription starts when RNA polymerase recognizes a promoter sequence, and, after liberating its sigma subunit, begins to elongate an RNA molecule. As the nascent RNA chain gets off polymerase, it interacts with molecules in the cytoplasm, including itself, forming sometimes a stem and loops or other complex structures such as pseudoknots. In general textbooks, it is assumed that nascent mRNA molecules enter ribosomes and immediately start being translated. But is this substantiated by experiments? In fact, the first likely interaction of nascent RNA is with the most proximate object, RNA polymerase itself. The presently available experimental work is dominated by the analysis *in vitro* of initiation and elongation termination complexes of RNA polymerase. In the corresponding studies it was shown that the enzyme behaves in a very unusual way, in that it seems to contract as an inchworm would, as it transcribes a portion of its DNA template (Uptain *et al*. 1997). This actual mechanism has been recently challenged (Komissarova & Kashlev 1997; Nudler *et al*. 1997), but the fact remains that RNA polymerase has an unusual behaviour with respect to transcription elongation and termination (Nudler *et al*. 1998; Uptain & Chamberlin 1997). As research progressed, more and more factors were found to be involved in transcription termination (Nus proteins) (Burns *et al*. 1998; Court *et al*. 1995; Huenges *et al*. 1998; Van Gilst & Von Hippel 1997; Vogel & Jensen 1997) and in elongation (in particular proteins of the GreA/GreB family). Many experiments also permitted investigators to demonstrate that the 5′-end of mRNAs sometimes had an influence on transcription termination far downstream. This is in particular noticeable for the antitermination factor N of bacteriophage lambda (Friedman *et al*. 1990; Whalen & Das 1990). Finally, the stringent coupling of stable RNA synthesis of a family of mRNAs was also found to be linked to the elongation process, associated to synthesis of the alarmone ppGpp (Krohn & Wagner 1996; Vogel & Jensen 1994). However, no clear-cut picture of the control events of these processes is yet available.

Let us consider the situation where the 5′-end of mRNA stays on a site located on RNA polymerase, perhaps through interaction with appropriate factors (such as the Nus proteins, for example). During elongation a loop forms, and both pre-existing ribosomes and newly formed ribosomes may assemble near the 5′-end (i.e. near the RNA polymerase transcribing complex), pushing the whole structure of the translating–transcribing loop away into the cell cytoplasm (figure 1). With this model, because both extremities of the RNA molecule are located next to each other, a scanning process, permitting the 3′-end of the transcript to interact with the 5′-end at appropriate places, can operate. With this picture one could easily visualize a loop of mRNA, comprising 10–20 translating ribosomes, bulging out from the transcription complex. It is worth noting that this figure fits extremely well with the average length of bacterial genes (*ca*. 1000 bp). Upon transcription termination a loop of mRNA is liberated, and this might be the step triggering mRNA degradation: this might account for the still unexplained fact that mRNA seems to disappear from its 5′-end, while RNases act either as endonucleases or 3′-exonucleases.

No data directly support this model, but few have looked for such an interaction. Most experiments showing functional interaction between the 5′- and 3′-ends of the mRNA molecule did not take the hypothesis of loops as a possibility (Mackie 1998). In fact, in electron micrographs of translating ribosomes, precautions are taken to spread

out the mRNA molecules as much as possible. In contrast, the fluffy chromosomes of salivary glands of insects show a large number of RNA loops. In nucleoli, a special protein complex interacts with the 5′-end of the transcripts and prevents reassociation of this end to the transcription complex, thus preventing the formation of loops. Since it remains very difficult to visualize the ongoing transcription process in living cells, it will be interesting to look for 5′–3′ correlations in the nucleotide sequences of operons. In general one therefore expects two distinct fates for transcripts: either they form loops, with the 5′-end scanning the 3′-end until it encounters some termination signal, or the 5′-end folds and forms an RNA–protein complex, with specific binding proteins, shifting away from the RNA polymerase transcribing complex. This would be the case of the rRNA, that rapidly associates with ribosomal proteins, but also of complexes such as the 5′-terminal regulator of the transcription control of tRNA synthetase genes in *B. subtilis* (Henkin 1994, 1996). In eukaryotes it has indeed been found that the 5′-end of the rDNA moves away from its DNA template in an organized fashion (Lazdins *et al*. 1997).

(iii) *Spacers and timers*

Replication, transcription, and translation are time-dependent processes. It is usually assumed that the rate of replication is 50–100 times faster than the rate of transcription, which is thought, in bacteria, to match roughly with the rate of translation (i.e. three nucleotides are transcribed during a time when one amino-acid residue is incorporated into the elongating polypeptide chain). As a consequence no DNA sequence can be said to be absolutely neutral in its action: it acts both as a spacer, placing sequences apart from each other (or next to each other again, if this corresponds to an appropriate fold of the polymer), and as a timer, allowing events to be delayed with respect to each other, instead of being simultaneous. For this reason several sequences within or outside genes have a function not by the actual succession of nucleotides, but by their length. It therefore seems likely that in eukaryotes some introns are conserved in length but not in sequence in cognate genes between related organisms. In bacteria this would fit with a non-random distribution of intervening sequences (ISs). A case in point might be the observation that ISs and prophage-like elements cluster at the terminus of replication in both *E. coli* and *B. subtilis*, although these organisms have a completely different organization of repeats in their genomes (Rocha *et al*. 1999*b*).

5. DISTRIBUTION OF FAMILIES ALONG THE CHROMOSOME

If the cell is a highly organized structure, the folding of its chromosome must somehow be constrained by the architectural components of the cell. Time-dependent processes should also be organized with respect to architecture. It therefore becomes important to analyse comparable properties of genes in the genome and investigate whether they display some regularities.

Replication is orientated: how is the relative orientation of transcription and translation of genes poised with respect to the progression of the replication fork? Overall, in *E. coli*, there is not a large difference in both orientations, although there is a slight increase in the number of genes transcribed in the same direction as that of replication (55 versus 45% in the opposite orientation). The situation is very different in *B. subtilis* since almost three-quarters of the genes are transcribed in the same orientation as the replication fork's movement. In this case the general trend is probably significant, since the major discrepancies in the general pattern favouring transcription in the same orientation as the movement of the replicating fork are found in prophage elements, where no gene expression is expected unless the lytic cycle is induced (Kunst *et al*. 1997).

Viari and co-workers used discriminant analysis to assess whether there was a bias in the properties of genes and gene products coming from each strand in bacteria. When the origin and terminus of replication were known (e.g. in *E. coli* and *B. subtilis*) they found a large asymmetry between the genes lying on the leading versus lagging strand at the level of nucleotides, codons and also, very surprisingly, amino acids. For several species (noticeably *Borrelia burgdorferi* and *Chlamydia trachomatis*), the bias is so high that the sole knowledge of a protein sequence allowed them to predict, with high accuracy, whether the gene is transcribed from one strand or from its complement (Rocha *et al*. 1999*a*). These findings, that indicate a strong organization principle in the bacterial chromosome, will have important consequences not only for our understanding of fundamental biological processes such as replication fidelity, codon usage in genes and amino-acid usage in proteins, but also for phylogenetic studies.

In bacteria most genes are organized in co-transcribed entities, the operons. Usually an operon comprises genes of the same metabolic pathway, or genes coding for subunits of a heteromeric enzyme. There are, however, much more complicated operons, with genes apparently of unrelated functions clustered together. A case in point is the operon comprising the *cmk* gene (encoding cytidylate kinase) and the *rpsA* gene (ribosomal protein S1), that is conserved in *E. coli* and *B. subtilis*. The functional consistency of this operon has been found to be presumably due to the role of CDP in DNA synthesis, this molecule deriving mostly from mRNA turnover in bacteria (Danchin 1997). This suggests that there is a force driving genes to cluster together, emphasizing the importance of architecture in the chromosome.

Some operons have genes clustered in *E. coli*, but apparently randomly distributed in *B. subtilis*, and vice versa. In her doctorate work P. Guerdoux-Jamet has observed that there is a correlation between the codon usage of genes involved in multiprotein complexes, or in metabolic pathways that have in common the fact that they are expressed in similar conditions or simultaneously. As a case in point (figure 2) she observed that genes coding isoenzymes expressed under aerobic conditions or anaerobic conditions could be split according to their difference in codon usage (Guerdoux-Jamet 1997). As a matter of fact, the corresponding genes expressed in the absence of oxygen were AT-rich and those expressed in the presence of oxygen were GC-rich. Knowing that the biotope of *E. coli* is either cold and aerobic or warm and anaerobic, this fits with the idea that the selection

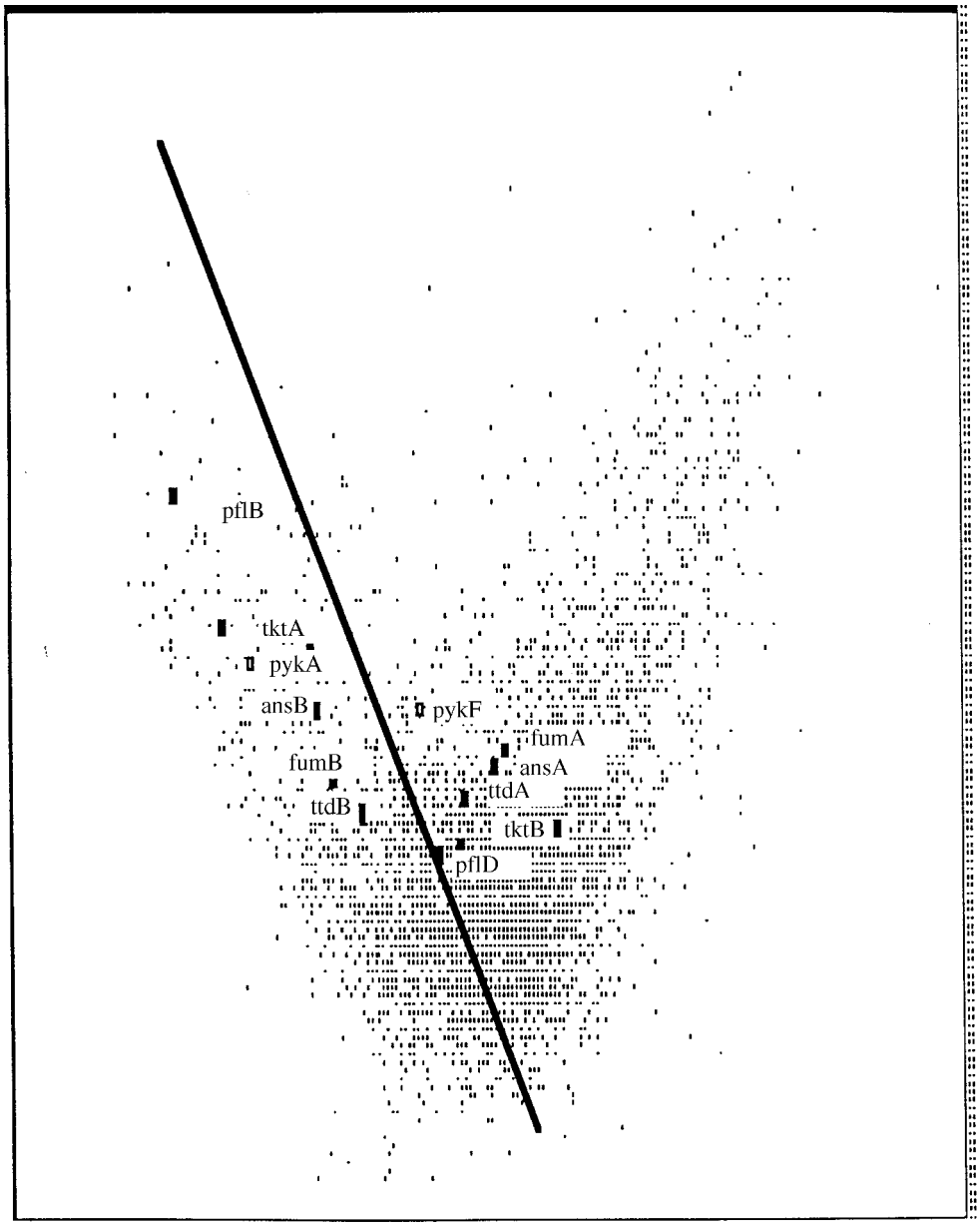

Figure 2. Distribution of isoenzymes expressed under aerobic and anaerobic growth conditions in the cloud of genes of *E. coli* distributed according to their codon usage. The axes are the first axes in factorial correspondence analysis (FCA) (maximum inertia).

pressure here is driven by the temperature. The question is therefore to try and understand the underlying principles of this selection. Distribution of the corresponding gene products did not appear to be random; however, because the number of the genes is small it remains difficult to see whether this is meaningful. We shall therefore explore now in more depth the reasons that might explain the creation of biases in codon usage.

6. CODON USAGE AND THE ORGANIZATION OF BACTERIAL GENOME SEQUENCES

As a science of relationships between objects, biology is extremely abstract. However, living beings are concrete entities. The question therefore arises to find out appropriate links that can build up concrete processes, starting from abstract structures and dynamics. The existence of the genetic code gives us a first hint of how to proceed: there is nothing in common between the chemistry of nucleotides and that of amino acids. However, there is a concrete link that creates the correspondence between nucleic acids and proteins, as initially proposed by Crick on a purely theoretical background. tRNA molecules act as adaptors that give flesh to the correspondence between codons and amino acids. As an average, an amino acid is encoded by three different codons. As a consequence any gene can be expressed using different codons: each gene has a specific codon usage. To study gene neighbours is to study clusters inside classes which comprise objects having a given common property. Many methods have been used to analyse such families, in particular principal components analysis, with two main types of distances, either identity or normalizing the distances by dividing each coordinate with the corresponding standard deviation for each dimension. FCA has been widely used for the exploration of the biological meaning of codon wage bias

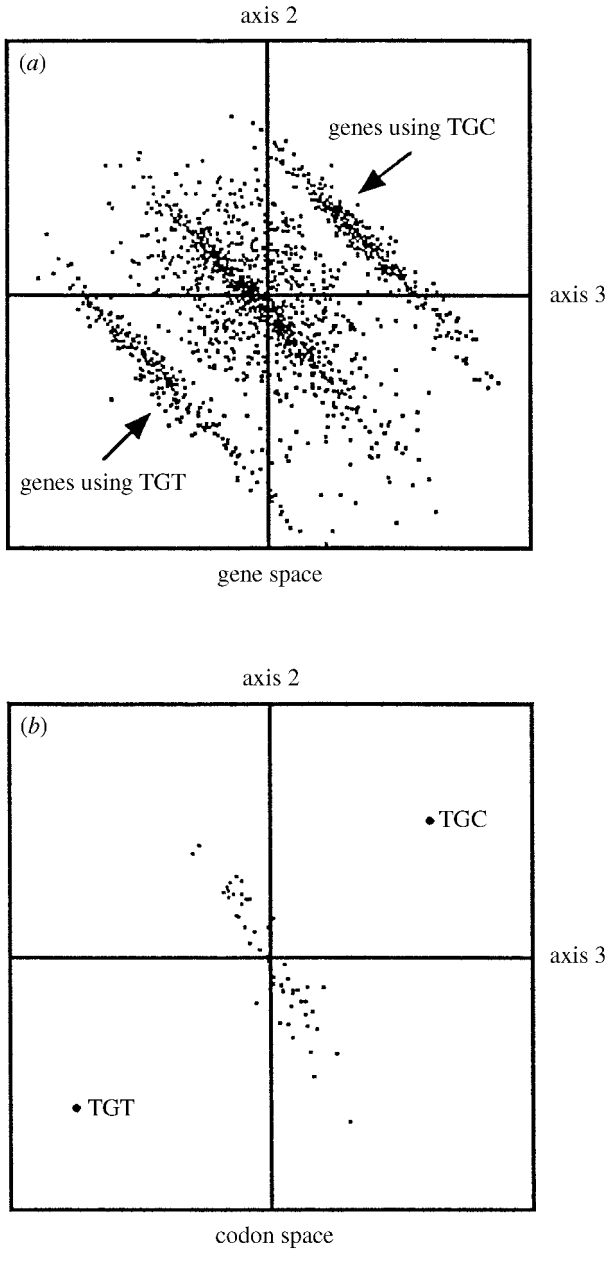

Figure 3. FCA analysis of the codon usage of the *B. subtilis* genes, along axes 2 and 3. It appears clearly that, with this view, genes form three classes. (*a*) represents the genes in the codon space, and (*b*) codons in the dual space (gene space). This allows one to identify the class in the upper (respectively lower) part of the gene cloud, as comprising genes with TGC (respectively TGT) cysteine codons.

(references in Médigue *et al.* (1991)), because it provides a means to evaluate the distance between objects with a weight (the χ^2-measure) that smooths out the differences in the number of objects making each class, and, above all, because this distance is meaningful for discontinuous sets of objects and associated properties that have always positive values (Lebart *et al.* 1984).

Once the cloud of points has been placed in the codon space it is possible to compute new orthogonal axes along which the cloud can be maximally spread out. With this construction, the first two axes display the largest differences between genes. The following axes show less and less difference between genes, as their rank increases. As shown by I. Moszer in his doctorate work, along the third axis, in both *E. coli* and *B. subtilis* genes cluster clearly as three classes of genes (Moszer 1996): a major, central one, and two well-separated classes (figure 3). The character that makes this clustering significant is the use of the cysteine codons, TGT and TGC. The bulk is made up of genes without cysteine codons, or with two or more cysteine codons. Putting aside these somewhat trivial classes, we are left with a distribution of genes that indicates a very large difference in their use of the genetic code and with no simple explanation for the corresponding behaviour.

If the use of codons were random one would expect a random distribution of codons in each gene, every gene being similar in this respect to all the others. This is not what is observed in bacteria such as *E. coli* and *B. subtilis*. If one plots the genes in the space of the 61 possible codons (in fact 57, putting aside methionine and tryptophan, which have a single codon, as well as the two cysteine codons, and normalizing codon usage such as to give the amino acids with two codons a weight similar to that of amino acids with more codons), one finds that in *E. coli*, as in *B. subtilis*, genes can be split into three classes according to the way they use the genetic code. The selection pressure maintaining this bias is linked to the organization of the cytoplasm (in a ribosome network) moving slowly with respect to local diffusion of the small molecules and macromolecules present in the cell. Ribosomes act as attractors of certain tRNA species, as a function of the local codon usage of the mRNA molecules they translate. This adapts codon usage of the gene corresponding to a given function to the position of its product. In particular, if two genes are biased very differently in the way they use codons this indicates that the mRNAs are not translated at the same place in the cell. Organization of the genes into polycistronic operons results in the fact that proteins having related functions are co-expressed locally, allowing compartmentalization of the corresponding substrates and products. As a consequence, if one goes from a very biased ribosome to a less biased one, the local concentration of the most biased tRNAs decreases. In turn this creates a selection pressure that produces a gradient in codon usage, as one goes further away from the most biased messengers and ribosomes. We have observed that this is related to the pattern of genes along the chromosome. If certain ribosomes are the cell's organizers, mRNAs from genes highly expressed under exponential growth conditions will be situated next to the centre of these organizers, whereas the other mRNAs will be translated as successive layers, up to the cytoplasmic membrane. The organization of the genes in the chromosome should therefore place in the limelight regularities linked to this architecture.

(a) *Codon usage: distribution of orthologues in* E. coli *and* B. subtilis

Making a collection of genes with identical function in *E. coli* and *B. subtilis*, it is possible to study their distribution as a function of the way they use the genetic code, using FCA as described above. As shown in figure 4, it is obvious that *E. coli* and *B. subtilis* use the code in a very different way. They have a style of their own: placing a gene in the right-hand side of the figure would predict that it is, most

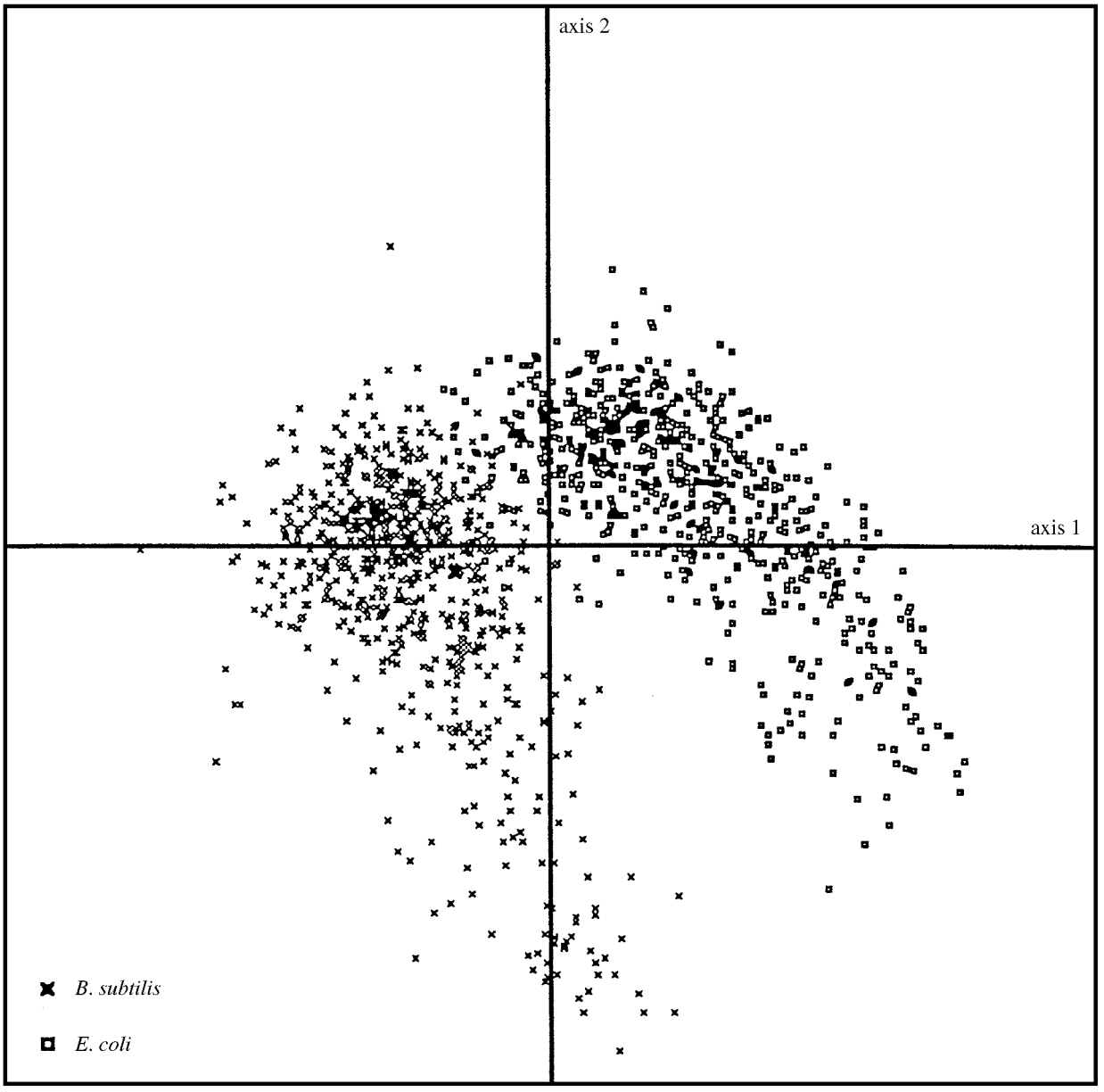

Figure 4. FCA analysis of *E. coli* and *B. subtilis* orthologues. Note that the 'style' of each genome is different, but that the overall shape of the cloud is similar.

probably, an *E. coli* gene. In contrast placing a gene in the left-hand side cloud of points predicts that it belongs to *B. subtilis*. However, there is a very surprising feature in this picture: the cloud of points of *E. coli* and *B. subtilis* genes have more or less the same shape. More precisely, if one considers the points corresponding to the same function in both organisms, one finds that, generally speaking, they deviate equally from the centre of gravity of the cloud of points. This indicates that a similar selection pressure is acting in both organisms, although they diverged probably some 1.5 billion years ago. There must be, therefore, a general physical constraint that behaves in a similar way in both organisms.

7. CONCLUSION: IS THE MAP OF THE CELL IN THE CHROMOSOME?

Can we propose a hypothesis to account for all these facts? It seems likely that if the bacterial chromosome behaves as a celluloculus, organizing the construction of the main functional components of the cell, then it is possible to understand the existence of a selection pressure acting on the text of the genes. The building up of the translation machinery is the driving force for cell growth. It organizes the chromosome separation, and it must be compartmentalized, if one has to explain the existence of large codon usage biases.

What are the underlying targeting principles that permit compartmentalization? We have suggested that the formation of planar structures may be a very general driving force. Indeed, a small proportion of mRNA with translated products having the tendency to pile up under the cytoplasmic membrane might suffice (as do drawing pins scaffolding the assembly of sheets on a board) to organize layers of gene products expressed from mRNA molecules transcribed in between these scaffolding complexes of mRNA with their stacked products. In addition, the creation of translated RNA transcription loops

might help structuration of the cytoplasm. Of course, other principles might also be at work, and we can expect that a thorough global analysis of the genomic texts will help us formulate new ones and make discoveries about the intricacies of the cell construction. In spite of the general agreement that there exists some compartmentalization in metabolism of small molecules in bacteria, not much is known about the cell's organization. Recently, many experiments using the green fluorescent protein from *Aequorea victoria* have revealed that many proteins are highly compartmentalized in bacteria, even when they do not have an intricate intracystoplasmic compartment system (such as in the case of cyanobacteria) (Glaser *et al*. 1997; Lemon & Grossman 1998; Lewis & Errington 1996; Wu *et al*. 1998).

Is this hypothesis general? In contrast to the situation in most bacteria, eukaryotic cells are already known to be highly organized. Many compartments are known to be important for the cell's architecture, in particular a network of membrane structures, together with a cytoskeleton made of microtubules, intermediary filaments and actin. It has been shown that the distribution of mRNA in the egg is not random, and that it is correlated to the future fate of the cells that make the embryo. But the distribution of mRNA has also been demonstrated to be highly organized in other cell types, such as neurons. In interphase cells, microtubules play fundamental roles in the intracellular distribution and movement of organelles and vesicles and thereby contribute to cellular polarization and differentiation. The organization of microtubules varies with the cell type and is presumably controlled by tissue-specific microtubule-associated proteins (MAPs). As a case in point, Chun and co-workers discovered that the squid giant axon contains a heterogeneous population of mRNAs that includes the transcripts for β-actin, β-tubulin, kinesin, neurofilament proteins, and, as in the case of the prokaryotic degradosome (Kaberdin *et al*. 1998), enolase (Chun *et al*. 1996). They quantified the levels of five mRNAs in the giant axon and compared it with the situation in the parental cell soma. In the latter, the number of transcripts for these mRNAs varied over a fourfold range, with β-tubulin being the most abundant species. The rank order of mRNA levels in the soma was β-tubulin > β-actin > kinesin > enolase > MAP H1. In contrast, the kinesin mRNA was the most abundant species in the axon (4.1×10^7 molecules per axon) with individual mRNA levels varying in a 15-fold range (with the order: kinesin > β-tubulin > MAP H1 > β-actin > enolase). In addition they found that the relative abundance of the mRNA species in the axon did not correlate with the size of the transcript. It was not directly related to their corresponding levels in the soma either. Taken together, these findings strongly suggest that specific mRNAs are differentially transported into the axon. The situation can probably be extended to other eukaryotic cells as well, where formation of planar structures together with continuous membrane synthesis acting as a conveyor belt might be a driving principle for the cell organization (Genty *et al*. 1994).

This work benefited from discussions with Cyprien Gay who pointed out to us the importance of the constraints underlying polymer statistics. We thank Alain Hénaut for his continuous interest. Work on databases was supported by the European Union BIOTECH programme BIO4-CT96–0655.

REFERENCES

Akiyama, Y. & Ito, K. 1995 A new *Escherichia coli* gene, *fdrA*, identified by suppression analysis of dominant negative FtsH mutations. *Mol. Gen. Genet.* **249**, 202–208.

Burns, C. M., Richardson, L. V. & Richardson, J. P. 1998 Combinatorial effects of NusA and NusG on transcription elongation and Rho-dependent termination in *Escherichia coli*. *J. Mol. Biol.* **278**, 307–316.

Charlier, D., Hassanzadeh, G., Kholti, A., Gigot, D., Pierard, A. & Glansdorff, N. 1995 *carP*, involved in pyrimidine regulation of the *Escherichia coli* carbamoylphosphate synthetase operon encodes a sequence-specific DNA-binding protein identical to XerB and PepA, also required for resolution of ColEI multimers. *J. Mol. Biol.* **250**, 392–406.

Chun, J. T., Gioio, A. E., Crispino, M., Giuditta, A. & Kaplan, B. B. 1996 Differential compartmentalization of mRNAs in squid giant axon. *J. Neurochem.* **67**, 1806–1812.

Court, D. L., Patterson, T. A., Baker, T., Costantino, N., Mao, X. & Friedman, D. I. 1995 Structural and functional analyses of the transcription-translation proteins NusB and NusE. *J. Bacteriol.* **177**, 2589–2591.

Danchin, A. 1987 Order and necessity. In *From enzyme adaptation to natural philosophy: heritage from J. Monod* (ed. E. Quagliariello, G. Bernardi & A. Ullmann), pp. 187–196. Amsterdam: Elsevier.

Danchin, A. 1997 Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP. *DNA Res.* **4**, 9–18.

Danchin, A. 1998 *La barque de Delphes. Ce que révéle le texte des génomes*. Paris: Odile Jacob.

Delrow, J., Gebe, J. & Schurr, J. 1997 Comparison of hard-cylinder and screened Coulomb interactions in the modeling of supercoiled DNAs. *Biopolymers* **42**, 455–470.

el-Hajj, H. H., Zhang, H. & Weiss, B. 1988 Lethality of a dut (deoxyuridine triphosphatase) mutation in *Escherichia coli*. *J. Bacteriol.* **170**, 1069–1075.

Friedman, D. I., Olson, E. R., Johnson, L. L., Alessi, D. & Craven, M. G. 1990 Transcription-dependent competition for a host factor: the function and optimal sequence of the phage lambda *boxA* transcription antitermination signal. *Genes Dev.* **4**, 2210–2222.

Genty, N., Paly, J., Edery, M., Kelly, P. A., Djiane, J. & Salesse, R. 1994 Endocytosis and degradation of prolactin and its receptor in Chinese hamster ovary cells stably transfected with prolactin receptor cDNA. *Mol. Cell. Endocrinol.* **99**, 221–228.

Ginard, M., Lalucat, J., Tummler, B. & Romling, U. 1997 Genome organization of *Pseudomonas stutzeri* and resulting taxonomic and evolutionary considerations. *Int. J. Syst. Bacteriol.* **47**, 132–143.

Glaser, P., Sharpe, M. E., Raether, B., Perego, M., Ohlsen, K. & Errington, J. 1997 Dynamic, mitotic-like behavior of a bacterial protein required for accurate chromosome partitioning. *Genes Dev.* **11**, 1160–1168.

Gottesman, S., Roche, E., Zhou, Y. & Sauer, R. T. 1998 The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes Dev.* **12**, 1338–1347.

Guerdoux-Jamet, P. 1997 Mise en oeuvre de logiciels de comparaison de séquences biologiques sur des machines spécialisées symboliques. Applications á l'analyse des génomes. PhD thesis, UFR de Biologie, University of Paris.

Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I. & Danchin, A. 1996 Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. *J. Mol. Biol.* **257**, 574–585.

Hénaut, A., Lisacek, F., Nitschké, P., Moszer, I. & Danchin, A. 1998 Global analysis of genomic texts: the distribution of AGCT tetranucleotides in the *Escherichia coli* and *Bacillus subtilis* genomes predicts translational frameshifting and ribosomal hopping in several genes. *Electrophoresis* **19**, 515–527.

Henkin, T. M. 1994 tRNA-directed transcription antitermination. *Mol. Microbiol.* **13**, 381–387.

Henkin, T. M. 1996 Control of transcription termination in prokaryotes. *A. Rev. Genet.* **30**, 35–57.

Herman, C., Thevenet, D., Bouloc, P., Walker, G. C. & D'Ari, R. 1998 Degradation of carboxy-terminal-tagged cytoplasmic proteins by the *Escherichia coli* protease HflB (FtsH). *Genes Dev.* **12**, 1348–1355.

Himeno, H., Nameki, N., Tadaki, T., Sato, M., Hanawa, K., Fukushima, M., Ishii, M., Ushida, C. & Muto, A. 1997*a* *Escherichia coli* tmRNA (10Sa RNA) in trans-translation. *Nucl. Acids Symp. Ser.* **37**, 185–186.

Himeno, H., Sato, M., Tadaki, T., Fukushima, M., Ushida, C. & Muto, A. 1997*b* *In vitro* trans translation mediated by alanine-charged 10Sa RNA. *J. Mol. Biol.* **268**, 803–808.

Huenges, M., Rolz, C., Gschwind, R., Peteranderl, R., Berglechner, F., Richter, G., Bacher, A., Kessler, H. & Gemmecker, G. 1998 Solution structure of the antitermination protein NusB of Escherichia coli: a novel all-helical fold for an RNA-binding protein. *EMBO J.* **17**, 4092–4100.

Kaberdin, V. R., Miczak, A., Jakobsen, J. S., Lin-Chao, S., McDowall, K. J. & Von Gabain, A. 1998 The endoribonucleolytic N-terminal half of *Escherichia coli* RNase E is evolutionarily conserved in *Synechocystis* sp. and other bacteria but not the C-terminal half, which is sufficient for degradosome assembly. *Proc. Natl Acad. Sci. USA* **95**, 11 637–11 642.

Karlin, S., Mrazek, J. & Campbell, A. M. 1997 Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179**, 3899–3913.

Kebbekus, P., Draper, D. & Hagerman, P. 1995 Persistence length of RNA. *Biochemistry* **34**, 4354–4357.

Keiler, K. C., Waller, P. R. & Sauer, R. T. 1996 Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science* **271**, 990–993.

Komissarova, N. & Kashlev, M. 1997 Transcriptional arrest: *Escherichia coli* RNA polymerase translocates backward, leaving the 3′ end of the RNA intact and extruded. *Proc. Natl Acad. Sci. USA* **94**, 1755–1760.

Krohn, M. & Wagner, R. 1996 Transcriptional pausing of RNA polymerase in the presence of guanosine tetraphosphate depends on the promoter and gene sequence. *J. Biol. Chem.* **271**, 23 884–23 894.

Kryzek, R. A. & Rogers, P. 1976 Dual regulation by arginine of the expression of the *Escherichia coli argECBH* operon. *J. Bacteriol.* **126**, 348–364.

Kunst, F. (and 150 others) 1997 The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Landais, S., Gounon, P., Laurent-Winter, C., Mazié, J. C., Danchin, A., Barzu, O. & Sakamoto, H. 1999 Immunochemical analysis of UMP kinase from *Escherichia coli*. *J. Bacteriol.* **181**, 833–840.

Lazdins, I. B., Delannoy, M. & Sollner-Webb, B. 1997 Analysis of nucleolar transcription and processing domains and pre-rRNA movements by *in situ* hybridization. *Chromosoma* **105**, 481–495.

Lebart, L., Morineau, A. & Warwick, K. A. 1984 *Multivariate descriptive statistical analysis*. New York: Wiley.

Lemon, K. & Grossman, A. 1998 Localization of bacterial DNA polymerase: evidence for a factory model of replication. *Science* **282**, 1516–1519.

Lewis, P. J. & Errington, J. 1996 Use of green fluorescent protein for detection of cell-specific gene expression and subcellular protein localization during sporulation in *Bacillus subtilis*. *Microbiology* **142**, 733–740.

Mackie, G. A. 1998 Ribonuclease E is a 5′-end-dependent endonuclease. *Nature* **395**, 720–722.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. & Danchin, A. 1991 Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**, 851–856.

Moszer, I. 1996 Représentation et analyse des génomes: application au projet de séquençage du génome de *Bacillus subtilis*. PhD thesis, UFR de Biologie, University of Paris.

Moszer, I., Glaser, P. & Danchin, A. 1996 SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiology* **141**, 261–268.

Muto, A., Sato, M., Tadaki, T., Fukushima, M., Ushida, C. & Himeno, H. 1996 Structure and function of 10Sa RNA: trans-translation system. *Biochimie* **78**, 985–991.

Muto, A., Ushida, C. & Himeno, H. 1998 A bacterial RNA that functions as both a tRNA and an mRNA. *Trends Biochem. Sci.* **23**, 25–29.

Myhill, J. 1952 Some philosophical implications of mathematical logic. I. Three classes of ideas. *Rev. Metaphys.* **6**, 165–198.

Nilsen, H., Yazdankhah, S. P., Eftedal, I. & Krokan, H. E. 1995 Sequence specificity for removal of uracil from U.A pairs and U.G mismatches by uracil-DNA glycosylase from *Escherichia coli*, and correlation with mutational hotspots. *FEBS Lett.* **362**, 205–209.

Nitschké, P., Guerdoux-Jamet, P., Chiapello, H., Faroux, G., Hénaut, C., Hénaut, A. & Danchin, A. 1998 Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol. Rev.* **22**, 207–227.

Nudler, E., Mustaev, A., Lukhtanov, E. & Goldfarb, A. 1997 The RNA–DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell* **89**, 33–41.

Nudler, E., Gusarov, I., Avetissova, E., Kozlov, M. & Goldfarb, A. 1998 Spatial organization of transcription elongation complex in *Escherichia coli*. *Science* **281**, 424–428.

Ollivier, E., Delorme, M. O. & Hénaut, A. 1995 DosDNA occurs along yeast chromosomes, regardless of functional significance of the sequence. *C. R. Acad. Sci. III* **318**, 599–608.

Rivals, E., Delgrange, O., Delahaye, J. P., Dauchet, M., Delorme, M. O., Hénaut, A. & Ollivier, E. 1997 Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequences. *Comput. Appl. Biosci.* **13**, 131–136.

Rivetti, C., Walker, C. & Bustamante, C. 1998 Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. *J. Mol. Biol.* **280**, 41–59.

Rocha, E. P. C., Danchin, A. & Viari, A. 1999*a* Universal replication biases in bacteria. *Mol. Microbiol.* **32**, 11–16.

Rocha, E. P. C., Viari, A. & Danchin, A. 1999*b* Long repeats reveal alternative evolutionary mechanisms in competent prokaryotes. *J. Mol. Evol.* (In the press.)

Sagot, M.-F. 1994 *Formalisation et approches combinatoires du problème de ressemblance lexicale et structurale entre macromolécules biologiques*. Marne-la Vallée, France: Université Gaspar Monge.

Sagot, M. F. & Myers, E. W. 1998 Identifying satellites and periodic repetitions in biological sequences. *J. Comput. Biol.* **5**, 539–553.

Schlick, T., Li, B. & Olson, W. 1994 The influence of salt on the structure and energetics of supercoiled DNA. *Biophys. J.* **67**, 2146–2166.

Turner, L. R., Olson, J. W. & Lory, S. 1997 The XcpR protein of *Pseudomonas aeruginosa* dimerizes via its N-terminus. *Mol. Microbiol.* **26**, 877–887.

Ueguchi, C., Seto, C., Suzuki, T. & Mizuno, T. 1997 Clarification of the dimerization domain and its functional significance for the *Escherichia coli* nucleoid protein H-NS. *J. Mol. Biol.* **274**, 145–151.

Uptain, S. M. & Chamberlin, M. J. 1997 *Escherichia coli* RNA polymerase terminates transcription efficiently at rho-independent terminators on single-stranded DNA templates. *Proc. Natl Acad. Sci. USA* **94**, 13 548–13 553.

Uptain, S. M., Kane, C. M. & Chamberlin, M. J. 1997 Basic mechanisms of transcript elongation and its regulation. *A. Rev. Biochem.* **66**, 117–172.

Van Gilst, M. R. & Von Hippel, P. H. 1997 Assembly of the N-dependent antitermination complex of phage lambda: NusA and RNA bind independently to different unfolded domains of the N protein. *J. Mol. Biol.* **274**, 160–173.

Vogel, U. & Jensen, K. F. 1994 Effects of guanosine 3′,5′-bisdiphosphate (ppGpp) on rate of transcription elongation in isoleucine-starved *Escherichia coli*. *J. Biol. Chem.* **269**, 16 236–16 241.

Vogel, U. & Jensen, K. F. 1997 NusA is required for ribosomal antitermination and for modulation of the transcription elongation rate of both antiterminated RNA and mRNA. *J. Biol. Chem.* **272**, 12 265–12 271.

Watanabe, T., Sugita, M. & Sugiura, M. 1998 Identification of 10Sa RNA (tmRNA) homologues from the cyanobacterium *Synechococcus* sp. strain PCC6301 and related organisms. *Biochim. Biophys. Acta* **1396**, 97–104.

Weiss, B. & el-Hajj, H. H. 1986 The repair of uracil-containing DNA. *Basic Life Sci.* **38**, 349–356.

Whalen, W. A. & Das, A. 1990 Action of an RNA site at a distance: role of the nut genetic signal in transcription antitermination by phage-lambda N gene product. *New Biol.* **2**, 975–991.

Williams, R. M., Rimsky, S. & Buc, H. 1996 Probing the structure, function, and interactions of the *Escherichia coli* H-NS and StpA proteins by using dominant negative derivatives. *J. Bacteriol.* **178**, 4335–4343.

Wu, L. J., Feucht, A. & Errington, J. 1998 Prespore-specific gene expression in *Bacillus subtilis* is driven by sequestration of SpoIIE phosphatase to the prespore side of the asymmetric septum. *Genes Dev.* **12**, 1371–1380.

Zacharias, M. & Hagerman, P. 1996 The influence of symmetric internal loops on the flexibility of RNA. *J. Mol. Biol.* **257**, 276–289.