

# Implication of gene distribution in the bacterial chromosome for the bacterial cell factory

Eduardo P.C. Rocha <sup>a,b</sup>, Pascale Guerdoux-Jamet <sup>c</sup>, Ivan Moszer <sup>a</sup>,  
Alain Viari <sup>b</sup>, Antoine Danchin <sup>a,\*</sup>

<sup>a</sup> *Régulation de l'Expression Génétique, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France*

<sup>b</sup> *Atelier de Bio-informatique, Université Paris 6, 12 rue Cuvier, 75005 Paris, France*

<sup>c</sup> *INSERM U 49 Hôpital Pontchaillou, 35033 Rennes Cedex, France*

Received 2 February 1999; accepted 23 July 1999

## Abstract

As bacterial genome sequences accumulate, more and more pieces of data suggest that there is a significant correlation between the distribution of genes along the chromosome and the physical architecture of the cell, suggesting that the map of the cell is in the chromosome. Considering sequences and experimental data indicative of cell compartmentalisation, mRNA folding and turnover, as well as known structural features of protein and membrane complexes, we show that preliminary *in silico* analysis of whole genome sequences strongly substantiates this hypothesis. If there is a correlation between the genome sequence and the cell architecture, it must derive from some selection pressure in the organisms growing in the wild. As a consequence, the underlying constraints should be optimised in genetically modified organisms if one is to expect high product yields. Consequences in terms of gene expression for biotechnology are straightforward: knocking genes out and in genomes should not be randomly performed, but should follow the rules of chromosome organisation. © 2000 Elsevier Science B.V. All rights reserved.

*Keywords:* Genomics; Functional analysis; Transcription complexes; Neighbours; Discovery; Optimisation

## 1. Introduction

The study of life should not be restricted to the study of biochemical objects, but must, preferably, investigate their relationships. Because genomes are the blueprints of life, they cannot be considered as simple collections of genes. In the

same way, cells are not tiny test tubes. They organise complexes of nucleic acids, proteins and other molecules as well. As much as we can tell, this organisation is reflected in enzyme high order structures: hardly any enzyme is really isolated as a single subunit type, and not as part of a complex with other proteins or RNAs (Vanzo et al., 1998). Of course, this is immediately relevant to the processes developed in biotechnology, where global properties of organisms are at work, for example during cultivation processes (see e.g.

\* Corresponding author. Tel.: +33-1-45688441; fax: +33-1-45688948.

*E-mail address:* adanchin@pasteur.fr (A. Danchin)

Iuchi and Weiner, 1996). How can we have access to the relevant features of genome organisation? In general, genome sequences are presented as Blast or Fasta annotated lists of genes, sometimes linked together in a more or less elaborate fashion. The global properties of genomes, if they exist, are absent from such primitive modes of annotation. This lack of consideration of an overall integration of genes into genomes gives a superficial view of what is a gene. As a result we are often witnessing, at genome conferences, declarations telling that a genome is a mere collection of genes distributed randomly along the chromosome. Indeed, one still finds rear guard views that see genomics as a pure fashion and interest for plain technology, without real signification. However, even when a genome is presented as ‘fluid’, its organisation reflects very strong constraints in the gene order (Liu and Sanderson, 1996).

In fact, the organisation of genes into operons, complex regulons (Collado-Vides, 1991) or pathogenicity islands (Finlay and Falkow, 1997) suggests that, often, related functions share physical proximity. Let us for example consider the ribosomal RNA genes: in all fast growing bacteria, they are clustered near the origin of replication (Honeycutt et al., 1993; Johansen et al., 1996; Cornillot et al., 1997), and they are generally transcribed from the continuously replicated (leading) strand. In the same way, several operons are conserved in very distant bacteria, such as ribosomal proteins operons, or the operon directing synthesis of the membrane ATP synthase. In order to understand genome organisation, we must therefore explore the distribution of genes along the chromosome. A way to do this is to explore the proximity of genes, extending this exploration to many more types of vicinities than their simple succession in the genomic text. Using this approach of inductive reasoning in the following text, we discuss data and processes that strongly suggest that bacterial genome organisation is directly correlated to the organisation of both the cell’s architecture and its dynamics.

## 2. Making parallels

Rather than follow the usual hypothetico-deductive method, where one combines the facts and concepts obtained through experiments to derive appropriate conclusions, one can use an inductive approach as a useful heuristics, trying to bring together observations that are otherwise unrelated, through the identification of neighbour families (Nitschké et al., 1998). As a matter of fact, a discovery (or a patentable idea) is often produced when one establishes the connection between two pieces of data or concepts that are not obviously related to each other. We are well aware of the fact that this ‘associationist’ approach is quite primitive, and that approaches using *rules* should be developed and at some point combined with it. But let us explore some of its preliminary consequences. Inductive exploration of genomes will consist in finding all neighbours of relevant sequences (e.g. genes). Of course ‘neighbour’ is meant here in the broadest possible sense. It means that objects that are neighbours share some common property. This includes not only similarity in structure or dynamics, but simply the existence of links they may have in common with other objects. The biological objects making a cell alive must not (and often they indeed cannot) be isolated from each other. As a matter of fact, when they start to purify an enzyme, biochemists have a hard time to get rid of what they used to name ‘contaminants’. Moreover, it often happens, as a protein is purified, that its recovered activity becomes lower and lower, and that the protein highest specific activity is expressed in physico-chemical conditions far from what is expected to be found in the cell. But, are not such proteins the manifestation of the existence of protein complexes having a relevant biological meaning? Is not this fact the indication that a living organism cannot be summarised as the collection of its genes and gene products? Knowledge of whole genome sequences is a unique opportunity to study the corresponding relationships at the cellular level.

Finding neighbours of a given class sheds a specific light on a gene, because this allows the description of its functions as a means to bring

together the various objects of the neighbourhood. Naturally, proximity in the chromosome is already known to be significant. Operons or pathogenicity islands show that genes located close to each other are often functionally related. There are many cases, however, where this physical proximity is not yet understood. As a case in point we do not see yet why *secY*, *adk* and *map* are clustered together in Gram positive organisms (they could be associated to specialised ribosomes, of course, as suggested by in vivo experiments, Matsumoto, 1997; Powers and Walter, 1997). But the fact is that they do, and persist to do so through billions of years of evolution.

Another neighbourhood, often explored, is the sequence similarity existing between genes or gene products. This is reflected by the creation of families of paralogues (with the recent ‘popular’ meaning: ‘genes that are significantly similar inside a given genome’), or orthologues (‘genes of different genomes that are significantly similar to each other’). The Darwinian triplet: variation/selection/amplification, which makes organisms evolve by creating new functions (for which they have to capture structures), results in a highly significant class of gene proximity. Paralogous or orthologous genes constitute ‘phylogenetic neighbours’. In the same way, the isoelectric point of a gene product may often be linked to an architecture: it often gives a first idea of its compartmentalisation: ‘isoelectric neighbours’ sometimes have functional or architectural features in common (Moszer et al., 1995).

More complex neighbourhoods give particularly revealing results. 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 clouds of points describing the codon usage bias of all the genes of the organism. Codon usage neighbours have been found to be biologically significant (Médigue et al., 1991). Finally, genes have been studied by scientists in laboratories all over the world, and this lead investigators to link them to other genes, for whatever reason: in this context ‘literature neighbours’ of a gene will be the genes found together with it in the same article. Here, we have used the neighbourhood data collected in the experimental

server Indigo (<http://indigo.genetique.uvsq.fr>; Nitschké et al., 1998).

### 3. Some principles of cell architecture

#### 3.1. Membranes and plane layers

A distinctive feature of living organisms is the ubiquitous presence of membrane structures. In fact a general ‘strategy’ of evolution has been either to compartmentalise the cell with a single — albeit sometimes very complex — envelope, comprising a lipid bilayer, or to multiply membranes and skins. This corresponds to an a posteriori efficient way to use the natural tendency of things to increase their entropy for the following reason. Liquid water is a highly organised fluid. As a built in principle, the natural tendency for water molecules is to occupy as many energy and spatial states as possible. This leads to systematic demixing of molecules that are in contact with water molecules, unless they provide energy-favoured interactions (e.g. ionic or dipolar interactions). As a consequence, the increase in entropy is the driving force for the construction of many biological structures: this physical parameter is at the root of the universal formation of helices, it drives the folding of proteins and the formation of viral capsids, it organises membranes into bilayers and creates higher complex biological structures.

This raises an important question for genomics: is it possible to find out, just knowing the genome text, whether a gene product will form a protein complex? This is of course even more unlikely than that an amino acid sequence could tell us exactly the fold of a protein, without knowing pre-existing folds. Pancreatic RNase would fold indeed, because selection isolated it with this behaviour (it is secreted in bile salts), but this should never have been accepted, as it did, as the paradigm of protein folding (Noppert et al., 1998). Alignment with model proteins of known structure will certainly help predicting a structure, because one has to take into account the selective forces that have, during phylogeny, led to the actual fold found in the models. This ‘threading’

approach is often used (Torda, 1997), but it should be extended to the study of protein complexes, by taking into consideration the intersubunit contacts in the models.

The largest increase in entropy of a molecular complex in water is when its surface/volume ratio is the highest. This is the case especially of plane layers, and this is minimised in spherical structures. As a consequence, when a plane meets another one, it can lose a layer of water molecules and stick there, piling up. Many geometrical structures can lead to plane structures, but regular hexagons have only two ways to interact: either they pile upon each other, forming tubes, or they form planes. Indeed many membrane structures are made of plane layers, and in particular of pieces of hexagonal paving (Peters et al., 1989; Ebisu et al., 1990; Lucken et al., 1990; Phipps et al., 1991; Meckenstock et al., 1994; Sara et al., 1996).

The case of an hexameric enzyme having features typical of soluble cytoplasmic proteins, uridylylase kinase, has been investigated. This enzyme raises an interesting question about compartmentalisation of cell metabolites. Indeed, with UMP as its substrate, it makes UDP, which is recognised by ribonucleoside diphosphate reductase, and, therefore, could ultimately lead to massive incorporation of uracil into DNA instead of thymine. This would pose a challenging DNA proof-reading problem to the cell (Weiss and el-Hajj, 1986; el-Hajj et al., 1988; Nilsen et al., 1995). Electron microscopy demonstrated that this enzyme is compartmentalised under the cytoplasmic membrane (Landais et al., 1999). Whether this corresponds to the formation of small pieces of planes docked under the membrane is not yet known, but the hexameric structure of the enzyme is certainly compatible with such hypothesis.

In the same way it was recently found that *Bacillus subtilis* DNA polymerase (PolC) is stationary with respect to the bulk of the cytoplasm, predominantly at or near the centre of the cell, and that DNA is pulled through (Lemon and Grossman, 1998). The actual complex that is responsible for anchoring of polymerase to the membrane is not known, but it should be remarked that the DnaB (DnaC in *B. subtilis*) heli-

case makes hexagons (Bujalowski and Klonowska, 1993), so that it could well be a major partner in the constitution of the membrane associated polymerase complex. It seems also remarkable that another important helicase, the Rho protein, that controls separation of hybrid RNA–DNA duplexes formed during transcription is also an hexameric protein forming hexagonal structures (Geiselman et al., 1992a,b). For the same physical reason as that required for DNA duplex fusion, it seems necessary that this helicase be fixed with respect to its template. Whether this is the case is not yet known.

### 3.2. Linear polymers

If a DNA molecule of the length of the *Escherichia coli* genome were randomly coiled, standard polymer theory tells that it would fit a sphere with a diameter of ca 10  $\mu\text{m}$  at physiological salt concentration, ten times more than the diameter of the cell. Superordered structures of DNA must therefore account for the DNA packaging into the cell. They include supercoiling, organisation into domains, and attachment to specific sites (Trun and Marko, 1998). Are these physical constraints reflected in the genome sequence? Preliminary work with the genome of *Saccharomyces cerevisiae* suggested indeed that they exist, operating at the DNA level in this organism (Ollivier et al., 1995; Rivals et al., 1997). Packaging DNA into a small compartment is another reason explaining the nucleus in eucaryotic cells: this limits considerably the number of 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.

Another folding problem of long polymers such as DNA or RNA is that they have a great many number of possible states. If they were free to diffuse, this would be incompatible with an organisation of the cell architecture. In fact a set of freely moving long polymers would rapidly tangle

into an unsortable bulk of knotted structures, even if they diffused through an organised lattice (such as the ribosome lattice). Anchoring points provide a very efficient way to lower drastically the number of available states. A single anchoring point, as is assumed in the general models of transcription, would strongly restrict the number of explored states. This might not be enough to reduce sufficiently the number of states that transcripts would still be able to explore, but this would prevent the formation of knots. It is well established that two anchoring points instead of only one would limit drastically the exploration of possible states, reducing it to a manageable number. How could this be achieved in the cell?

A first answer considers that ribosomes are organised as a slowly moving lattice, and that nascent RNA coming off DNA is pulled by a first ribosome, then by the next one, as in a wire-drawing machine. Considering that most of the cell's inertia and that the major part of the cell's energy is in the translation machinery (one mRNA molecule is translated at least 20 times), it is the structure of the ribosome network that organises the mechanics of gene expression (a further energy cost is involved in proof-reading, Jakubowski, 1993, 1994). Translation/transcription coupling makes DNA move and brings at its surface new genes ready for transcription. The messenger RNA passes from a ribosome to the next one, controlling synthesis of the protein it specifies at each ribosome. Finally, as soon as an appropriate signal reaches the ribosome at the same time as the translated messenger, this triggers degradation of the mRNA from its 5'-end using a (yet unknown) degradation process (Bjornsson and Isaksson, 1996), thus ending its expression.

A second model, curiously never explored explicitly, does not assume that nascent mRNA molecules enter ribosomes and immediately start being translated. In contrast, it supposes that the 5'-triphosphate end of the message folds back, and remains linked to RNA polymerase until a specific signal, which can be located way downstream, tells it to detach (and to terminate transcription). Antitermination has been thoroughly investigated in the case of the antitermination factor N of bacteriophage  $\lambda$  (Friedman et al.,

1990; Whalen and Das, 1990), or, more recently in the case of response to nitrate, where protein NasR acts as a RNA binding antiterminator (Chai and Stewart, 1998). Antitermination is readily compatible with a scanning process permitting the 5'-end of the RNA to explore what happens in 3'-downstream sequences. As a case in point, a most interesting observation suggests that not only is there an effect on RNA of a protein, CspE, but that there is a link between the binding of this protein on the nascent RNA and chromosome partition in the daughter cells (Hanna and Liu, 1998). Finally, the stringent coupling of stable RNA synthesis as well as that of a family of mRNA was also found to be linked to the transcription elongation process, associated to synthesis of the alarmone ppGpp. This suggests that there is concerted transcription of at least some genes. It would therefore be interesting to see whether there are not cases where several individual RNA polymerases, associated by some specific binding protein, do not transcribe different regions of the chromosome simultaneously. This would provide a function for some of these puzzling unknown genes that seem to be spread in all newly sequenced genomes. However, no clear-cut picture of the control events of these processes are yet known (DiRusso and Nystrom, 1998; Loewen et al., 1998). Since it is certainly very difficult at this time to visualise in living cells the ongoing transcription process, it will be interesting to look for 5'-3' correlations in the nucleotide sequences of operons, as well as for factors that could make complexes comprising several RNA polymerases.

In summary one would expect two distinct fates for transcripts: either they would form loops, with the 5'-end scanning the 3'-end until it encounters some termination signal, or the 5'-end would fold and form an RNA-protein complex, with specific binding proteins, shifting away from the RNA polymerase transcribing complex, sometimes comprising several RNA polymerases transcribing DNA in parallel. This might be the case of ribosomal RNA, that associate 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).

#### 4. Distribution of families along the chromosome

Considering that the cytoplasm of bacteria is an ordered structure, we expect that the folding of its chromosome must somehow be constrained by the architectural components of the cell. This hypothesis implies also that time-dependent processes are organised with respect to architecture: indeed, during rapid exponential growth, a significant gene dosage effect can be expected between the origin and terminus of replication (there are more gene copies of the genes near the origin of replication, and therefore it is likely that there is more of the corresponding gene products). This phenomenon should be reflected in the gene order in the genome (Krawiec and Riley, 1990; Liu and Sanderson, 1996). But, because of the general folding of the chromosome, one cannot consider that we could identify explicitly most of this phenomenon, even if it exists, unless we have means to detect, in advance, a periodicity in the distribution of gene or sequence features. Do we see, nevertheless, global properties in genomes?

##### 4.1. Dissymmetry in the leading versus lagging strands

Replication is oriented, because it is continuous on one strand (leading strand) and discontinuous on the complement strand (lagging strand, Okasaki fragments (Fijalkowska et al., 1998)). It is interesting to investigate the relative orientation of transcription and translation of genes with respect to replication. 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 replication goes (55 vs 45% in the opposite orientation). The situation is very different in *B. subtilis* since 75% of the genes are transcribed in the same orientation as the replication fork movement. In this case, the 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). In general, there is a bias in oligonucle-

otide distribution as a function of the strand, but this has not yet been coupled to functional properties of the cell (Rocha et al., 1998).

Using factorial correspondence analysis (FCA) of the codon usage in *Borrelia burgdorferi*, McInerney found a very surprising pattern (McInerney, 1998). In this organism, genes transcribed from the lagging strand did not display the same bias in the genetic code usage as genes transcribed from the leading strand. The difference in codon usage was so strong that it could be used to predict, with codon usage bias as the sole piece of information, whether a gene is coded by the leading or the lagging strand. This prompted Viari and co-workers to make an in depth analysis of codon usage and related properties in all known bacterial genomes. Using FCA alone with codon usage bias as the labelling features of genes did not reveal prominent features of the organisms (except in *B. burgdorferi*, of course). However, because objects in the class have often hidden properties that are superimposed to each other, FCA does not always allow one to construct pertinent classes, unless some knowledge is obtained independently. For this reason Viari used another statistical technique, discriminant analysis, meant to assess whether a working hypothesis about classes is substantiated or not from the statistical point of view (Rocha et al., 1999).

Because the situation in *B. burgdorferi* revealed a difference of the nature of the codon usage of genes in both strands, it was interesting to start, as an hypothesis, with the assumption that a difference should be looked for in genes coming from each strand, in other bacteria. When the origin and terminus of replication were known (e.g. in *E. coli* and *B. subtilis*) this hypothesis was then submitted to discriminant analysis (Perrière et al., 1996). It was subsequently tested with all 16 available bacterial genomes. In most cases, a strong asymmetry between the genes lying on the leading versus lagging strand was observed at the level of nucleotides, codons and also, very surprisingly, amino acids. For several species (noticeably *B. burgdorferi* and *Chlamydia trachomatis*), the bias is so high that the sole knowledge of a protein sequence allows one to predict, with accuracy, whether the gene is transcribed from one

strand or from its complement. These findings have important consequences not only for our understanding of fundamental biological processes in bacteria such as replication fidelity, codon usage in genes and even amino acid usage in proteins, but also for genome engineering. In addition, taking into account the gene dosage effect near the origin of replication in rapidly growing cells, these observations suggest that, because the selection pressure for a bias is lower, the best locus for insertion of new DNA is the terminus of replication in circular chromosomes. Indeed, this feature is found in *E. coli* and *B. subtilis* (an example is given below displaying the distribution of the *B. subtilis* genes belonging to the codon usage class of horizontally transferred genes (Fig. 2b)), as well as in other bacteria such as *Bacillus cereus* (Carlson and Kolsto, 1994). The study of fluctuation between linear and circular chromosomes in *Streptomyces* sp. will reveal important features of the underlying constraints (Volf and Altenbuchner, 1998).

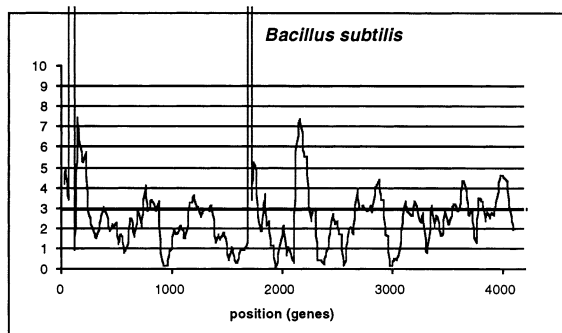


Fig. 1. Non-random distribution of gene orientation in *B. subtilis*. The figure displays the  $z$ -value (serial randomness) taken in windows of 100 genes along the chromosome (very high values indicate windows where all or nearly all genes are co-orientated, therefore revealing a high degree of aggregation). Serial randomness ( $z$ -value) is a serial robust test that indicates, given a succession of two classes if the string is random, i.e. if the elements are locally aggregated (e.g. AAAABBBBAA), or if the elements are uniformly distributed (e.g. ABABABAB). The  $z$ -values follow a central reduced Normal distribution (Gaussian law) for the hypothesis of null aggregation, therefore a  $z$ -value higher than 3.29 indicates significant aggregation at the 1% level.

#### 4.2. Operons and general bias in gene orientation

In bacteria most genes are organised in co-transcribed entities, the operons. In *E. coli* or *B. subtilis*, combining knowledge of translational coupling and Rho-independent transcription termination signals, it can be computed that the average transcript comprises three protein coding sequences (Fig. 1). Usually an operon comprises genes of a same metabolic pathway, or genes coding for subunits of an heteromeric enzyme. Often, the promoter distal genes are coding for the first steps of the pathway (e.g. *ilvGEDA* in *E. coli*; or *leuDCBA* in *E. coli* or *B. subtilis*). There are however much more complicated operons, with genes of apparently unrelated functions clustering together. An example 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).

Generally speaking, it is interesting to explore the local density of genes with the same orientation (orientation persistence) relative to the orientation of the replication fork movement. As shown in Table 1 the persistence of gene orientation for 13 complete genomes shows a significant bias in genes displaying locally the same orientation. The contrast (i.e. the relative number of genes in the leading strand as compared to the total) varies between an almost equal distribution between strands (*Methanococcus jannaschii*, *E. coli*, *C. trachomatis*) and an extreme bias in favour of the leading strand (the mycoplasmas and *B. subtilis*). The persistence of local orientation, however, is significant in all genomes, as shown in Table 1. This means that even if the genes are shuffled in equal amount between the two strands, there is a tendency in all these genomes towards the local aggregation of co-orientated genes.

If this clustering has something to do with the gene function, one expects that genes having similar functions in different genomes might have architectural properties in common. We therefore analysed orthologous genes (with the 'popular'

Table 1  
Polarity of genomes with predicted (or confirmed) Ori and Ter regions<sup>a</sup>

Species	Strand (+)	Strand (–)	Contrast (%)	Persistence	z-value
<i>B. subtilis</i>	3147	1071	75	8.88	20.5
<i>B. burgdorferi</i>	577	310	65	5.47	11.9
<i>C. trachomatis</i>	508	413	55	4.07	12.9
<i>E. coli</i>	2407	1998	55	4.06	27.4
<i>H. influenzae</i>	1007	783	56	4.35	19.1
<i>H. pylori</i>	945	688	58	5.31	22.3
<i>M. jannaschii</i>	909	814	53	4.18	16.9
<i>M. thermoautotrophicum</i>	1076	842	56	5.11	21.7
<i>M. genitalium</i>	402	102	80	9.76	10.6
<i>M. pneumoniae</i>	558	153	78	6.02	11.6
<i>M. tuberculosis</i>	2342	1630	59	3.53	20.2
<i>R. prowaseki</i>	532	338	61	4.44	14.0
<i>T. pallidum</i>	711	380	65	5.30	13.8

<sup>a</sup> Entries *strand* (+) and *strand* (–) collect the number of genes in each strand (note that in the absence of experimental confirmation, we predicted *strand* (+) as the leading strand either because it has more genes, or because it is the one carrying the *dnaA* gene). The ‘contrast’ is computed as:  $strand+/(strand++strand-)$ . The ‘persistence’ is the number of immediately downstream genes expected to be found after a given gene displaying the same orientation. If  $\rightarrow$  denotes a gene, the genes  $i, j, k, l, m, n$ , in succession as:  $\rightarrow\rightarrow\rightarrow\rightarrow\leftarrow\leftarrow\leftarrow$  show individual persistence of:  $i = 3, j = 2, k = 1, l = 2, m = 1$ . The persistence of a gene represents how far one has to go from its position to find the first gene in the inverse orientation. In the table we present the average persistence for all genes. The definition of the z-value is given in the legend of Fig. 1. Note that in *B. subtilis* there are 75% of genes in the leading strand, therefore persistence is necessarily higher in this organism than in *E. coli* (55%). Genes are nevertheless significantly aggregated in both bacteria.

meaning given above) of *E. coli* and *B. subtilis* and studied their distribution in the leading and the lagging strand. As shown in Table 2 there is a very strong conservation of strand when we go from one genome to the other, despite the fact that the general bias of gene orientation is very different in the two genomes.

#### 4.3. Codon usage and architecture

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 tryptophane, which have a single codon, cysteine, that is rare in proteins, and normalising codon usage such as to give amino acid with two codons similar weight 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 (Médigue et al., 1991; Moszer, 1998).

The selection pressure maintaining this bias is linked to the organisation of the cell’s cytoplasm, which consists in a slowly moving ribosome network. In order to account for the existence of this bias, one may consider that ribosomes act as attractors of certain tRNA species, as a function of the local codon usage of the mRNA molecules they translate. This would adapt the codon usage of the gene corresponding to a given function to the position of its product in the cell. Two genes having a very different codon usage might not be

Table 2  
Relative distribution of orthologous genes in *E. coli* and *B. subtilis*<sup>a</sup>

	lag. Bs	lead. Bs	
lag. Ec	18% (7)	13% (24)	31%
lead. Ec	4% (15)	65% (54)	69%
	22%	78%	

<sup>a</sup> Experimental results (expected results in brackets).

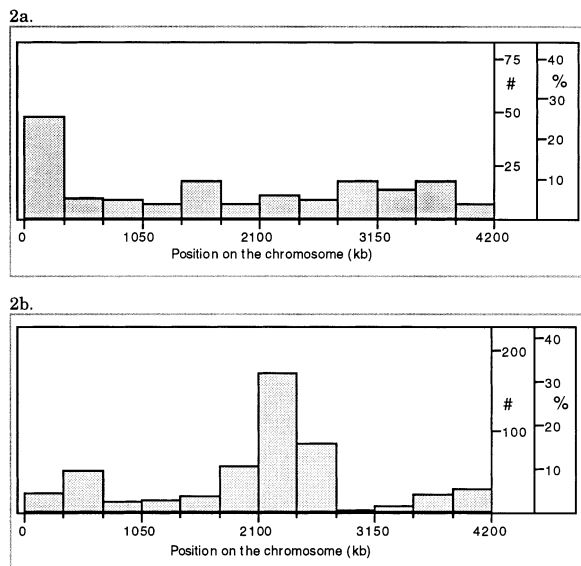


Fig. 2. Asymmetry in the distribution of genes in terms of FCA classes of codon usage bias in *B. subtilis* (a). There is a strong asymmetry in the distribution of genes in terms of FCA classes of codon usage bias. Class 2 genes cluster on one side of the origin of replication. (b) Class 3 genes cluster near the terminus of replication.

translated at the same place in the cell. As a consequence the global concentration of tRNA species only partially reflects the codon composition of highly expressed genes, as demonstrated by Kurland and co-workers (Dong et al., 1996). Organisation of the genes into polycistronic operons results in the fact that proteins having related functions are co-expressed locally, allowing compartmentalisation 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 would create a selection pressure producing a gradient in codon usage, as one goes further away from the most biased messengers and ribosomes. As a consequence if certain ribosomes were the cell's organisers, mRNAs from genes highly expressed under exponential growth conditions would be situated next to the centre of these organisers, whereas the others mRNAs would be translated as successive layers, up to the cytoplasmic membrane. A thorough analysis of the organisation of the genes in the

chromosome should therefore place in the lime-light regularities linked to this architecture, which could be further constrained by RNA polymerases transcribing in concert.

Substantiating this hypothesis, the distribution of the three classes of genes found in *B. subtilis* is far from random (Kunst et al., 1997). Remarkably, as shown in Fig. 2, the two highly biased classes (class 2, genes highly expressed under exponential growth conditions; class 3, genes deriving from horizontal transfer) are not distributed randomly along the chromosome as shown by several tests for randomness (data not shown). In fact class 2 genes are clustered mostly on one side, near the origin of replication (Fig. 2a). This may be related to the asymmetrical division process that takes place during sporulation. In contrast class 3 genes (Fig. 2b) cluster mostly near the terminus of replication, indicating that this region is a preferred place for the integration of new DNA.

## 5. Conclusion

In spite of much data suggesting the contrary, biochemists generally behave as if procaryotic cells were tiny test tubes. Furthermore, in biotechnological processes aiming at heterologous gene expression, one uses generally constructions that do not take this fact into consideration. Recently, many experiments using the green fluorescent protein from *Aequorea victoria* have revealed that many proteins are highly compartmentalised in bacteria, even when they do not have an intricate intracytoplasmic compartment system (such as in the case of cyanobacteria) (Lewis and Errington, 1996; Glaser et al., 1997; Lemon and Grossman, 1998; Wu et al., 1998). As summarised here, global analysis of genome sequences has independently shown that the gene order cannot be considered to be random.

It seems remarkable that this special distribution of genes having a different composition and/or codon usage bias, resulting in a particular bias in the amino-acid content of proteins gives a specific 'style' to a genome. As a consequence a genome must express some character of robust-

ness that might have important consequences in terms of stable invasion by foreign DNA, and this argues in favour of the existence of true species in bacteria, despite the known ubiquitous importance of lateral gene transfer.

Finally for biotechnology processes, considering expression of heterologous genes in an organism, these pieces of data suggest that the means used to optimally express a gene should consider its place in the chromosome (or in a plasmid) with respect to the origin of replication, its neighbours, and/or its codon usage bias.

## Acknowledgements

This work summarises the conclusion of work that benefited from discussion with numerous colleagues. It was supported by grants from the Groupement d'Étude et de Recherche sur les Génomes (GIP GREG), several EU Biotechnology grants, and was partially funded by the EU grant Biotech BIO4-CT96-0655. E.R. acknowledges the support of PRAXIS XXI, through grant BD/9394/96.

## References

Bjornsson, A., Isaksson, L.A., 1996. Accumulation of a mRNA decay intermediate by ribosomal pausing at a stop codon. *Nucleic Acids Res.* 24, 1753–1757.

Bujalowski, W., Klonowska, M.M., 1993. Negative cooperativity in the binding of nucleotides to *Escherichia coli* replicative helicase DnaB protein. Interactions with fluorescent nucleotide analogs. *Biochemistry* 32, 5888–5900.

Carlson, C.R., 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.

Chai, W., Stewart, V., 1998. NasR, a novel RNA-binding protein, mediates nitrate-responsive transcription antitermination of the *Klebsiella oxytoca* M5al nasF operon leader in vitro. *J. Mol. Biol.* 283, 339–351.

Collado-Vides, J., 1991. A syntactic representation of units of genetic information – a syntax of units of genetic information. *J. Theor. Biol.* 148, 401–429.

Cornillot, E., Croux, C., Soucaille, P., 1997. Physical and genetic map of the *Clostridium acetobutylicum* ATCC 824 chromosome. *J. Bacteriol.* 179, 7426–7434.

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.

DiRusso, C.C., Nystrom, T., 1998. The fates of *Escherichia coli* during infancy and old age: regulation by global regulators, alarmones and lipid intermediates. *Mol. Microbiol.* 27, 1–8.

Dong, H., Nilsson, L., Kurland, C.G., 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* 260, 649–663.

Ebisu, S., Tsuboi, A., Takagi, H., Naruse, Y., Yamagata, H., Tsukagoshi, N., Udaka, S., 1990. Conserved structures of cell wall protein genes among protein-producing *Bacillus brevis* strains. *J. Bacteriol.* 172, 1312–1320.

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

Fijalkowska, I.J., Jonczyk, P., Tkaczyk, M.M., Bialoskorska, M., Schaaper, R.M., 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *Proc. Natl Acad. Sci. USA* 95, 10020–10025.

Finlay, B.B., Falkow, S., 1997. Common themes in microbial pathogenicity revisited. *Microbiol. Mol. Biol. Rev.* 61, 136–169.

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.

Geiselman, J., Seifried, S.E., Yager, T.D., Liang, C., von Hippel, P.H., 1992a. Physical properties of the *Escherichia coli* transcription termination factor Rho. 2. Quaternary structure of the Rho hexamer. *Biochemistry* 31, 121–132.

Geiselman, J., Yager, T.D., Gill, S.C., Calmettes, P., von Hippel, P.H., 1992b. Physical properties of the *Escherichia coli* transcription termination factor Rho. 1. Association states and geometry of the Rho hexamer. *Biochemistry* 31, 111–121.

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.

Hanna, M.M., Liu, K., 1998. Nascent RNA in transcription complexes interacts with CspE, a small protein in *E. coli* implicated in chromatin condensation. *J. Mol. Biol.* 282, 227–239.

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

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

Honeycutt, R.J., McClelland, M., Sobral, B.W., 1993. Physical map of the genome of *Rhizobium meliloti* 1021. *J. Bacteriol.* 175, 6945–6952.

Iuchi, S., Weiner, L., 1996. Cellular and molecular physiology of *Escherichia coli* in the adaptation to aerobic environments. *J. Biochem. (Tokyo)* 120, 1055–1063.

Jakubowski, H., 1993. Energy cost of proofreading in vivo: the charging of methionine tRNAs in *Escherichia coli*. *FASEB J.* 7, 168–172.

- Jakubowski, H., 1994. Energy cost of translational proofreading in vivo. The aminoacylation of transfer RNA in *Escherichia coli*. Ann. NY Acad. Sci. 745, 4–20.
- Johansen, T., Carlson, C.R., Kolsto, A.B., 1996. Variable numbers of rRNA gene operons in *Bacillus cereus* strains. FEMS Microbiol. Lett. 136, 325–328.
- Krawiec, S., Riley, M., 1990. Organization of the bacterial chromosome. Microbiol. Rev. 54, 502–539.
- Kunst, F., Ogasawara, N., Moszer, I., et al., 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.
- Lemon, K.P., Grossman, A.D., 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.
- Liu, S.L., Sanderson, K.E., 1996. Highly plastic chromosomal organization in *Salmonella typhi*. Proc. Natl Acad. Sci. USA 93, 10303–10308.
- Loewen, P.C., Hu, B., Strutinsky, J., Sparling, R., 1998. Regulation in the rpoS regulon of *Escherichia coli*. Can J. Microbiol. 44, 707–717.
- Lucken, U., Gogol, E.P., Capaldi, R.A., 1990. Structure of the ATP synthase complex (ECF1F0) of *Escherichia coli* from cryoelectron microscopy. Biochemistry 29, 5339–5343.
- Matsumoto, K., 1997. Phosphatidylserine synthase from bacteria. Biochim. Biophys. Acta 1348, 214–227.
- McInerney, J.O., 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl Acad. Sci. USA 95, 10698–10703.
- Meckenstock, R.U., Krusche, K., Staehelin, L.A., Cyrklaff, M., Zuber, H., 1994. The six fold symmetry of the B880 light-harvesting complex and the structure of the photosynthetic membranes of *Rhodospseudomonas marina*. Biol. Chem. Hoppe Seyler 375, 429–438.
- 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., 1998. The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. FEBS Lett. 430, 28–36.
- Moszer, I., Glaser, P., Danchin, A., 1995. SubtiList: a relational database for the *Bacillus subtilis* genome. Microbiology 141, 261–268.
- 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.
- Noppert, A., Gast, K., Zirwer, D., Damaschun, G., 1998. Initial hydrophobic collapse is not necessary for folding RNase A. Fold Des. 3, 213–221.
- 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.
- Perrière, G., Lobry, J.R., Thioulouse, J., 1996. Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. Comput. Appl. Biosci. 12, 519–524.
- Peters, J., Peters, M., Lottspeich, F., Baumeister, W., 1989. S-layer protein gene of *Acetogenium kivui*: cloning and expression in *Escherichia coli* and determination of the nucleotide sequence. J. Bacteriol. 171, 6307–6315.
- Phipps, B.M., Huber, R., Baumeister, W., 1991. The cell envelope of the hyperthermophilic archaeobacterium *Pyrobaculum organotrophum* consists of two regularly arrayed protein layers: three-dimensional structure of the outer layer. Mol. Microbiol. 5, 253–265.
- Powers, T., Walter, P., 1997. Co-translational protein targeting catalyzed by the *Escherichia coli* signal recognition particle and its receptor. EMBO J. 16, 4880–4886.
- 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.
- Rocha, E.P.C., Danchin, A., Viari, A., 1999. Universal replication biases in bacteria. Mol. Microbiol. 32, 11–16.
- Rocha, E.P.C., Viari, A., Danchin, A., 1998. Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. Nucleic Acids Res. 26, 2971–2980.
- Sara, M., Kuen, B., Mayer, H.F., Mandl, F., Schuster, K.C., Sleytr, U.B., 1996. Dynamics in oxygen-induced changes in S-layer protein synthesis from *Bacillus stearothermophilus* PV72 and the S-layer-deficient variant T5 in continuous culture and studies of the cell wall composition. J. Bacteriol. 178, 2108–2117.
- Torda, A.E., 1997. Perspectives in protein-fold recognition. Curr. Opin. Struct. Biol. 7, 200–205.
- Trun, N., Marko, J., 1998. Architecture of the bacterial chromosome. ASM News 64, 276–283.
- Vanzo, N.F., Li, Y.S., Py, B., Blum, E., Higgins, C.F., Raynal, L.C., Krisch, H.M., Carpousis, A.J., 1998. Ribonuclease E organizes the protein interactions in the *Escherichia coli* RNA degradosome. Genes Dev 12, 2770–2781.
- Volf, J.N., Altenbuchner, J., 1998. Genetic instability of the *Streptomyces chromosome*. Mol. Microbiol. 27, 239–246.
- 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.
- 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.