ComPlexUs
ORIGINAL RESEARCH PAPER
Complexus 2004–05;2:61–70 DOI: 10.1159/000089983
Published online: November 25, 2005

Genome Diversity: A Grammar of Microbial Genomes

Antoine Danchin

Genetics of Bacterial Genomes, Institut Pasteur, Paris, France

Key Words
Phylogeny · Dissymmetry · Gene transfer, horizontal · Translation · Replication

Abstract
The metaphor of the genetic program is ubiquitously used. Does it bring about a deeper insight into what cells are? In brief, computers do not make computers, but cells make cells. Pursuing the analogy in its deepest consequences we explore bacterial genome diversity as organized around core programs meant to couple the expression of the program with the architecture of the cell. At first sight genomes appear to evolve fast and exhibit no rule of organization. However, when the huge number of generations separating various organisms is taken into account, diversity appears only as a trivial observation. In contrast, rules of gene organization are observed, for example in the separation between Archaea and Bacteria, in the composition of the leading and lagging strand of the chromosomes, in the distribution of genes along the DNA strands, or in the formation (and conservation) of operons and pathogenicity islands.

In contrast to the laws of physics, those of biology are full of exceptions: as soon as a rule is discovered, a further discovery shows an exception. This is even true of the most central features of what has been named the 'central dogma' of molecular biology (dogma indicated in a prominent way that this was more or less the place for belief, not for rational thought). The genetic code offers exceptions, reverse transcription is ubiquitous, and what about RNA editing, ribozymes and non-coding RNA's? Because of this situation there will always be biologists trying to discover universal rules, while others will fight against universality. The consequence is either that diversity will be perceived as the core of biological systems (and this culminates in genome studies in the ubiquitous fuzzy concept of 'genome plasticity' [1–3]), or that diversity will simply be seen as a collection of exceptions to the rule. In the present article I try to balance these views and explore whether – at least in prokaryotes – diversity is not simply a general process meant to propagate and stabilize in a selective way systems that are constructed along universal rules. The principle of selective stabilization is indeed at the bottom of all living processes, and is used over and over again, even at the level of the central nervous system of animals.

A Very Brief Summary of What Life Is

Placing life in context is necessary when we wish to understand what genomes are, and how they are constructed. All material systems submitted to the empedoclean/maupertuisian/malthusian/darwinian[1] triplet: variation/selection/amplification will evolve. In the course of their evolution they will create actions aiming at selective

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com
KARGER

1424–8492/05/0022–0061
$22.00/0
Accessible online at:
www.karger.com/cpu

Antoine Danchin, Genetics of Bacterial Genomes
URA 2171 CNRS, Institut Pasteur, 28 rue du Docteur Roux
FR–75724 Paris Cedex 15 (France)
Tel. +33 1 45 68 84 42, Fax +33 1 45 68 89 48,
E-Mail adanchin@pasteur.fr

[1] Selective theories are often attributed to Darwin. However, they were pervading the thoughts of philosophers interested in biology. Aristotle quotes Empedocles for such visions and tries to ridicule him. More recently much discussion of the topic can be found in Maupertuis, and even in the famous diaries of Samuel Pepys.

stabilization of the system, or rather (there is of course no prescribed goal to the evolution of material systems) we will only witness the existence of those systems that have been selectively stabilized into forms that stay long enough in existence. This is akin to learning with concomitant imbedding in the system of an image of the environment (that which created the selective pressure to which the system has been submitted) [4]. To this purpose, evolving systems will usually recruit preexisting structures rather than come up with a de novo magic construct that would fulfill the needs of the action required by the stabilization goal. This accounts for the 'tinkering' aspect of most biological constructs and is the context in which life – and genomes – have to be considered. Hence, contrary to what is often thought, the function will not tell the structure, except when it derives by divergent evolution from a sufficiently close kin. Among such evolving systems some are endowed with life.

Briefly, three major processes are required to make a living entity: (1) metabolism (ongoing chemical processes that transform molecules into other molecules), (2) compartmentalization (the cell with its inside and its outside is the atom of life), and (3) information transfer (this is the place where the central dogma of molecular biology operates).

Understanding genomes has to be placed in this particular context: an isolated genome will not construct another genome. This is what viruses do – and we shall not comment on that – but viruses do not exist in the absence of living cells. Genome diversity is reflecting these constraints, both those of metabolism and those of compartmentalization. As a matter of fact there is a broad difference between the genomes of prokaryotes and those of eukaryotes. The sequence of the former looks, at first sight, highly random, while that of the latter looks highly repeated [5]. We shall not go on further with this first observation, and only remark that this indicates a first link between the architecture of the genome and the architecture of the cell. In what follows we shall mostly deal with prokaryotes, as models for individual cells.

Data, Programs and Machines

The metaphor of the genetic program was created as a convenient way to describe how cells live and develop. Something stable had to be transmitted from generation to generation, but the preformist view of the whole organism as the transmitted entity was rapidly discarded as it was contradicted by the impossibility to segment objects ad infinitum. Hence, what had to be transmitted in the course of generations was not the final organism, but, rather, a recipe to make it (duplicating recipes is much less difficult to conceive, even though the question of errors during duplications must be included in the picture). This paved the way for the concept of a genetic program that became almost self-evident when the structure of DNA was discovered. It was, however, understood early on that some strange organisms (were they living or not alive?), the viruses, behaved as individual pieces of programs, using the cell as the machine needed to make them multiply and subsequently propagate (often by destroying the machine): no virus can survive without a living host cell. Later on, when computer programming took off on a very large scale, pieces of programs were found to behave formally as do biological viruses, and were named 'viruses' accordingly. This was a further indication that there was perhaps a deeper meaning than the surface meaning of the 'program' metaphor of what life is. Furthermore, when it became possible to manipulate DNA in vitro, the analogy appeared even deeper: working in silico on a string of symbols was enough to allow scientists to construct in vitro, and then in vivo, experiments that corresponded to the very concrete action of reprogramming the cell's fate.

The discovery of the processes setting up regulation of gene expression, followed by that of the genetic code, spread the representation of life as the result of the expression of a program that could be seen as a linear string of symbols [6]. This concept was already well understood at a time when the first computers had been shown to operate as predicted by Turing [7], von Neumann [8] and the many theoreticians and scientists who had uncovered the link between the arithmetics of whole numbers and logic. In his famous metaphor, Turing proposed that all computations involving integers as well as all operations of logic could be performed by a simple machine reading and modifying a tape carrying a linear sequence of symbols, the universal Turing machine. He was able to show that this required only the physical separation between the string of symbols (visualized as a tape) handled by the machine and the machine itself. More precisely, he showed that the tape was carrying the data that allowed the machine to proceed. In terms of their anthropocentric meaning, the data could be split into two types, a program that embedded the 'meaning' of the logical sequence recognized by the machine, and the pure data that were needed for the program to operate (in a way, they provided the context). This immediately suggested the following metaphor (or research program [9]): would it be possible to consider the cell as a Turing machine, and if so, what are the implications in terms of biological objects needed to make it run. As early as 1972 Woese [10] tried to associate the downstream process of translation with the tape reading metaphor, linking it with the creation of complexity during evolution. However, the core of Turing's description is that of the physical separation between 'data + program' and 'machine' and this could not be explored conceptually and experimentally before it became possible to manipulate DNA and reprogram cells, as performed in the process known as genetic engineering. The genetic pro-

gram is carried out by the DNA molecule; can we consider it as a separate entity, and if so, to what extent? Genetic engineering rests on the manipulation of DNA molecules (real or artificially constructed ones) and expression in foreign cells: this is a first proof of concept. Many bacteria today produce human proteins. However, this represents a small part of the genetic program; is it possible to extend this analogy? The identification of widely spread horizontal gene transfer [11], and subsequently nuclear cloning [12], perfected the analogy of the cell as a Turing machine to a point where it can be considered as highly revealing, if not (of course) explaining life in totality. We shall, therefore, explore bacterial genome diversity, using this metaphoric analogy as a research program, to find out the nature of processes that must be imbedded into the genetic programs and allow one to understand both their universal nature and their diversity. In this brief review we cannot extend our exploration to the counterpart of the exploration of similarities between cells and computers, which explores the similarity between computers and cells in biomimetic studies, such as the 'bioinspired' creations developed by Mange and his colleagues [13, 14] at the Ecole Polytechnique Fédérale de Lausanne (Switzerland). This in itself would warrant a full study.

Essential Genes, Housekeeping Genes and the Minimal Genome: What Type of Diversity?

The Turing machine is an abstraction. To make it concrete, i.e. to make a computer, von Neumann [8] tried to classify the various processes that needed to be implemented in any program that a Turing machine would handle. This made him propose the concept of what we now refer to as the operating system (OS), a particular piece of the program that is indispensable for the machine to operate [15]. The OS is defined by a series of functions creating, within the program, some sort of an image of the processes required by the machine to perform its role. As a first step, one needs to distinguish between the machine and its 'users'. Thus, a description of a 'virtual machine' is conceived within the program that hides from the users all the engineering details of the computer as a physical entity. At the core of the OS there must be a 'resource manager' providing efficient and effective sharing of the needed physical and abstract routines among users of the machine (each one using and creating data while running programs). Naturally, users are usually not human users but they can be other machines as well (printers, screens, memory storage devices, all kinds of peripherals) and some are even programs (this is why software engineering is so important for the creation of large software pieces). Among those, we find several important classes of programs such as system's programs (compilers, editors, loaders), application support programs (database management systems, networking systems) and finally, the programs that correspond to the goals of the machine, application programs.

Because this is very abstract, there is no reason, when these concepts are transformed into real lines of code, why there should exist only one type of OS. As a matter of fact, in the computer industry, many exist (and are in competition with each other, since no computer would work without an OS). OSs are not fixed in time, and they certainly evolve as we can see in the present computer market. In short, an OS plays the role of a 'housekeeping' program. Do we find similar properties in living organisms? If we analyze the number of articles using the expression 'housekeeping genes' we find several hundreds of articles, suggesting that there is some consensus on the nature of the processes needed to be present in all cells. At first sight, living cells display overall features similar to one another, and the (almost) universal rule of the genetic code as well as of the DNA replication machineries would argue for universality. However, the process of cell division is remarkably different between the eukaryotes and the prokaryotes, for example. Compartmentalization is also very different in these organisms, with the former having a well-formed nucleus. In the class of prokaryotes, Woese et al. [16] revolutionized the community of microbiologists when they uncovered a remarkable discrepancy between two classes of cells, separated by the very core of their housekeeping machinery (translation first, but also transcription, replication and compartmentalization), the Archaea and the (previously recognized) Bacteria. Even Bacteria are not homogeneous [see the debate about the origin and nature of prokaryotes, 17–18].

This exploration of the OS metaphor provides us with the first level of diversity in prokaryotic genomes, located at a very deep level and probably originating very early on in the evolution of life: despite similarities, there are large differences in the housekeeping genes controlling replication, transcription and translation in cells. The question of their origin and evolution is still open. Even in Bacteria, there are at least two classes of core DNA polymerase III: most use only one housekeeping DNA polymerase for the management of both DNA strands, while the A + T-rich 'monoderm' [Bacteria with a single membrane; 17] Firmicutes use two such enzymes (DnaE and PolC), perhaps for a different management of the leading and lagging strands [20]. Symmetrically, the Firmicutes use only one SpoT/RelA protein both for synthesis and degradation of the universal regulator pppGpp, while gammaproteobacteria (to which *Escherichia coli*, the best-known organism belongs) have two such enzymes, SpoT and RelA [21]. All this points to the idea of a common functional class, that of the analog of an OS [22], which would differ in different types of cells.

It is important to notice that, when the core housekeeping gene products differ, this will most probably have consequences in many other gene products, resulting in an in-built diversity that fits with the large classes (domains or kingdoms) of organisms as we now classify them. This opens the question (discussed below) of the 'self' of each species, and whether it is somehow labelled in the genome. This diversity is also at the core of the colonization of the Earth by species that tend to limit exchanges of genetic material within one domain or kingdom [18].

In the course of specification of this diversity, two general strategies are at work: either organisms are single-cell organisms, or they tend to multiply membrane and skins [5, 18]. Single cells would need an OS similar to that of personal computer OSs, with some time-sharing properties. In some instances the OS could degenerate to that of a simple batched OS or, more often, to multiprogrammed batched OSs. For more complex organisms, one would obviously need parallel and/or distributed systems. In general, because the analog of the OS must manage many nanomachines, it would probably be more of the object-oriented type (i.e. managing resources inside data files). As a consequence, while it is important for each organism to identify its housekeeping genes, there is no compelling reason that would state that these genes should have exact counterparts in all organisms. The only good reason for universality would be historical: if it is difficult to create this or that function, it is likely that once it has appeared somewhere it will spread everywhere. This implies divergent evolution (but horizontal transfer as well). In contrast, for functions that are more straightforward, one could witness diversity and/or convergent evolution. The major housekeeping functions are the replication machinery, the transcription machinery, and the translation machinery. The cell membrane also has to be constructed, and to allow for import and export of metabolites and proteins. Finally there must be a set of genes required for accuracy and maintenance of the major housekeeping processes. Various methods made it possible to compute the minimum number of genes needed to perform these functions, and both by reasoning (this evaluation was at the root of the creation of genome programs in the European Union in the mid 1980s [5, 23] and was discussed at many meetings meant to support the idea of sequencing genomes, see also [24]) and by experimental evidence [25], the minimum gene set is limited to about 300 genes or so. As expected, this set is strongly correlated with genes considered as essential for growth on media supplemented with most, if not all, indispensable basic metabolites [25]. It is interesting to see that, even in Bacteria, there is some variation in the set of essential genes [26, 27], in line with the idea of variations in the OS driving the corresponding Turing machine.

A first tentative conclusion that may be drawn from this observation is that, from an initial population of cells exchanging genetic material at a high frequency, a family of organisms began to differentiate in such a way that they would create isolates progressively more resistant in terms of invasion by foreign genes. These organisms could then colonize specific environmental niches.

Regularities in Bacterial Genomes: The Link between the Architecture of the Cell and the Architecture of the Genome

In a crucial reflection, von Neumann [8] remarked that machines do not make similar machines. Very simple automata, such as crystals, can do so, but as soon as they are complicated enough, this apparently becomes impossible. There is only one exception, living organisms. What would be the constraints if we had to think of a computer making a computer? The answer, according to von Neumann, is that, within the computer, there should be some type of an image of the machine that would also be passed from generation to generation. This requires both a hereditary component and a structural component. Because, in living organisms, the most obvious hereditary component is the chromosome, it is interesting to explore if, and how, some image of the cell could be built in the chromosome organization [5, 28]. In order to do so, we first analyze the way in which DNA is handled by the various machineries in bacteria, explore the diversity of the corresponding processes and then try to see whether, despite this diversity, some common features emerge.

Periods, Motifs and Repeats: A Highly Diverse World of Genomes

Starting with the plain genomic DNA sequence, several types of regularities have long been observed. The most prominent one is the result of the selection pressure caused by the nature of the genetic code: a period of 3 is prominent in all bacterial genomes, somewhat correlated to the RNY self-complementary motif [29, 30]. Interestingly, a second period around the value of 10 is also visible once the period of 3 has been subtracted [31]. Its explicit biological meaning has not yet been explored. Many more precise motifs have been analyzed in genomes. However, there are not many universally conserved motifs, except perhaps some biases in the TA versus AT content or CG versus GC content [which is nevertheless quite variable, but may be due to some structural property of the DNA molecule rather than to cytosine methylation, as often proposed, 32, 33]. When considering tetranucleotides in bacteria, the rarity of CTAG aside [33, 34], the frequency of tetranucleotides as well as of longer motifs is highly dependent on the genome considered [35], often correlated with the presence of repeated sequences such as the BIME or palindromic units in *E. coli* [36]. Other markers, such as the GATC motif

used for labelling the nascent strands as compared to their parents, is present with similar roles in many gammaproteobacteria, but is far from being ubiquitous [34, 37].

Analysis of the number of repeats in bacterial genomes gave a puzzling result: there is no correlation between the number of repeats in a genome and its length. Furthermore, for a given genome's length, some genomes harbor a large number of repeats (e.g. *E. coli*), while others appear to restrict considerably the phenomenon of repeated sequences (e.g. *Bacillus subtilis*) [38]. This latter situation is reminiscent of the repeat-induced point mutation process found in fungi [39], but it has not been further characterized (and it is all the more puzzling because rDNA clusters are present in such genomes, and cannot be submitted to a strong mutational process). All this pleads for the recognition that DNA management in genomes is highly variable, even in apparently related species. This large diversity will obviously force evolution, and tend to compartmentalize organisms within clusters of 'DNA management' types. This implies that, were we to find regularities in genomes, they would need to be selected by processes and constraints of universal importance.

Origin versus Terminus: Circular and Linear Chromosomes, a Highly Constrained World of Genomes

One such universal constraint is replication. DNA replication is asymmetrical. While this does not pose major problems at the origin of replication – there is just a need to melt the DNA structure, which is easily achieved by local A + T enrichment – the problem of the terminus is far more challenging. The alternative is either to have a circular chromosome (but then the cell must resolve the knotted structure present at the terminus), or to end up as linear sequences (but then the cell must find structures able to start replication for the lagging strand, without losing in length: this is the origin of the wide variety of telomeric structures). In eukaryotes, except for a few plasmids, linear chromosomes with telomeres are the rule. In bacteria the circular chromosome is the rule, but quite a few bacteria have, nevertheless, linear chromosomes with telomeres (e.g. *Borrelia burgdorferi* or *Streptomyces coelicolor*). Origins of replication are difficult to identify in eukaryotes in general, while they are most often well characterized in bacteria, where the leading versus the lagging strand replication difference gives rise to a GC skew (there is more G in the leading strand than in the lagging strand, making it possible to identify where the origin and terminus of replication are located, just by analyzing the GC content of the DNA molecule) noted after the first complete genome sequences had been published [40]. Before completion of the sequencing of genomes it was not really possible to investigate their global structure, and despite many studies indicating that the differences in mutation rates between the leading and the lagging strand might result in a different base composition in each strand no clear picture could be obtained, besides that presented by some single-stranded DNA viruses [41]. Remarkably, besides the GC-skew bias, further studies observed that the bias is of universal nature, with G (and T) enriched in the leading strand as compared to the lagging strand, and extreme consequences in terms of amino acid composition of the proteins coded by each strand [42]. Thus, in many genomes, the bias introduced by the presence of a well-defined origin of replication results in strong constraints that bias not only the genetic code usage, but also the amino acid sequence of the proteins. This usually occurs with conservation of synteny in the relevant parts of genomes (fig. 1). It is not yet known whether it has important consequences in terms of evolution, but it might have favored, at least in eukaryotes, the variation in the origin of replication, in order to average out the mutational bias on both strands of DNA replication.

Because there is such a large difference between the leading and the lagging strand of DNA replication, it was interesting to explore another consequence of this dissymmetry: replication and transcription do not proceed at the same rate, on the one hand (with replication much faster than transcription), and, on the other hand, replication/transcription conflicts might result in synthesis of truncated transcripts. While this may be of limited importance in eukaryotes, in particular because the need for splicing into mature transcripts into authentic messenger RNAs creates an error screening-out process, it would be important in bacteria since a truncated mRNA might direct the synthesis of a truncated protein (with a frequent negative dominant phenotype). Analysis of the distribution of essential genes in bacteria with a well-defined origin of replication supported the existence of this selective constraint since most essential genes, regardless of their relative level of expression, sit on the leading replication strand [27, 43]. It will, therefore, be interesting to explore the general organization of bacterial cells that do not have a genome with a well-defined origin of replication.

Codon Usage Bias

In bacteria, most of the DNA sequence is used to code for proteins. If there were to be a link between the organization of the genome and protein production, it would be expected that some bias might exist in the way in which proteins are synthesized. In the early 1980s, Grantham et al. [44], remarking that the genetic code redundancy might create a codon usage bias, analyzed the sequence of the genes known at that time and observed that there indeed existed a significant bias in the highly expressed genes of the translation and transcription machinery as compared to that of the bulk of the proteins. Later on, more subtle dif-

ferences were discovered [11]. It appeared quite remarkable that the general codon usage bias, in bacteria as diverse as *E. coli* and *B. subtilis*, despite the large difference in codon frequencies, was apparently correlated with the function of the corresponding proteins [28] (fig. 2). Several hypotheses could be proposed to account for this observation, but all require some type of compartmentalization, either of the tRNAs or of the messenger RNAs in the cell, and this should be somehow correlated to the organization of the genome [5, 28]. The most common explanation rests on the idea that there is a strong correlation between tRNA abundance, tRNA charge by its cognate amino acid and codon usage bias [45, 46]. A more detailed analysis of the relationship existing between cell compartments and codon usage bias has found further correlations [47]. This observation has been used with success to explore the neighborhoods of gene products in order to predict gene function [48]. However, naturally, correlation is not cause, and the question of the driving force that leads to different codon usage biases remains open.

Horizontal Gene Transfer

At this point we observe that, despite large differences in the biochemical processes underlying DNA management, there are rules that pervade the bacterial world, with the majority of bacteria having a circular chromosome with well-defined ori-

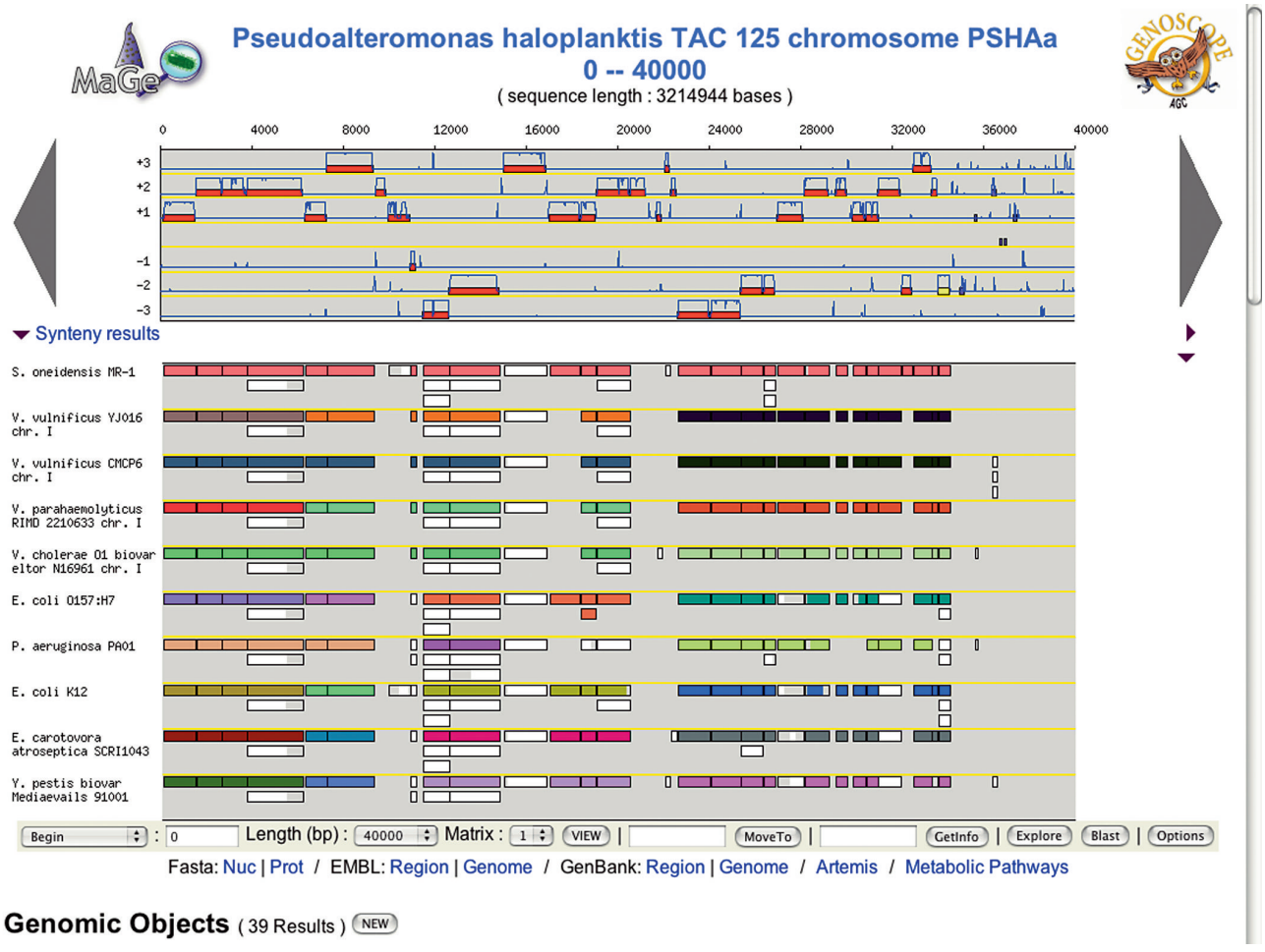

Fig. 1. Synteny near the origin of replication in gammaproteobacteria. The new genome of *Pseudoalteromonas haloplanktis* TAC125 is compared to counterparts in other bacteria of the same class, including marine bacteria. Conservation at the level of gene organization is prominent.

gins and termini of replication, and with their essential genes mostly distributed on the leading replication strand. Genes that are expressed at a high level are also often biased in their codon usage. A basic consensus, on which phylogenies rest, was for a long time that species were created by divergent evolution by dichotomy from common ancestors. With the molecular clock hypothesis, this allowed for construction of trees that some thought would lead back to an enigmatic progenote or last universal common ancestor. This very naïve view is still held by popular science, and in some circles interested in the problems of origins [26]. However, it is likely that many pieces of metabolism had to be put together to construct the first organisms, and that at the onset of life, what would become genes had to be widely shared and exchanged. The major discovery of the RNA world was that of coupling RNA to proteins with the creation of the translation machinery [49, 50]. DNA was subsequently discovered as a way to stabilize the memory process in the course of generations. What made the scenario get unified was the general sharing of the genetic code, with a common general translation machinery. With this view, horizontal gene transfer was initially the rule, rather than the exception. As we have seen in the previous paragraphs, it is, however, likely that the major housekeeping processes were selected early on creating different functional entities, presumably through geographic isolation of populations of cells with a large variety of genes, resulting in the three major kingdoms, Bacteria, Archaea and Eukarya. The main consequence of this separation was that gene exchanges between kingdoms are likely to have become more and more difficult. The analogy with the OS is revealing in this respect: software portability is much dependent on whether the software goes to computers with the same OS or with a different one. When analyzing the contribution of horizontal gene transfer in genomes, one should, therefore, distinguish between transfer within a given kingdom and transfer between kingdoms; the latter is expected to be much rarer than the former.

Lateral gene transfer was identified as early as the time when bacteriophage lysogeny was discovered. However, this was assumed to be of limited importance associated with the phenomenon of local transduction and/or prophage insertion in the genome [51]. The genome sequences presently available, while substantiating the phage hypothesis, show that there is also a very significant proportion of genes that may come from horizontal gene transfer. The first observation that a significant portion of genomes was associated with gene transfer was the result of the finding that the codon usage bias of some 15% of the reference *E. coli* genome was totally different from that of the majority of the genes [11]. Interestingly, this study suggested that the antimutator genes could be among the laterally transferred genes, with the suggestion that many bacterial species stay in the environment as mutator populations which, when they encounter a stable environment, gain stability in producing less mutations [11]. Since then, data about horizontal gene transfer accumulated, with some indication that, in some species, genes from outside could outnumber

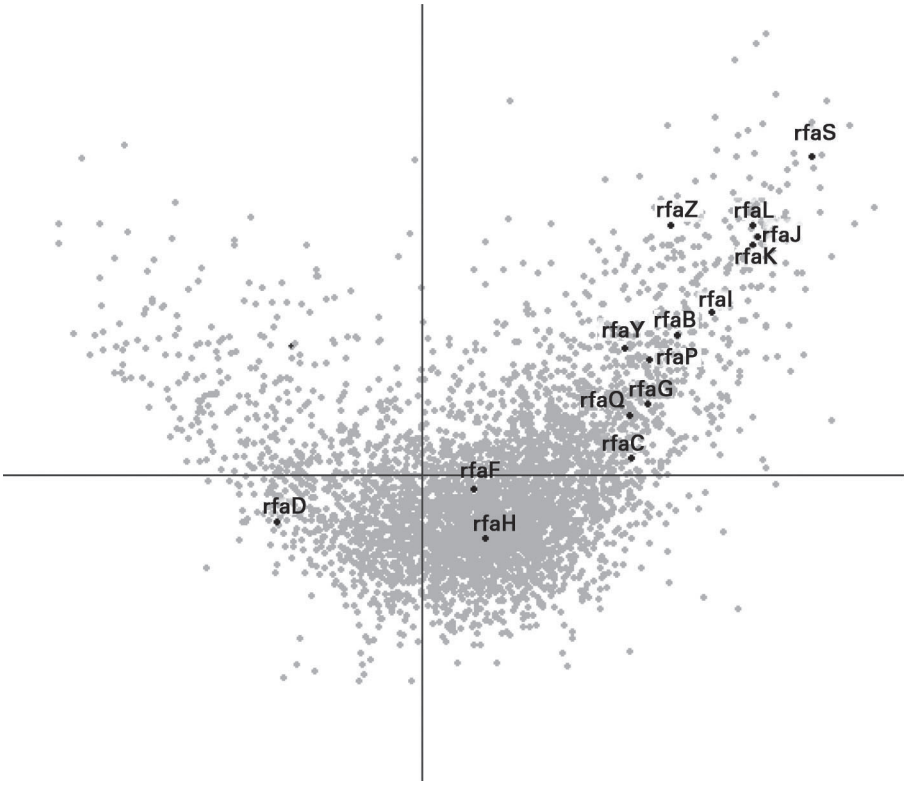

Fig. 2. Distribution of proteins from *E. coli* K12 according to their codon usage, in the first discriminating plane using Correspondence Analysis [11]. The cloud of points is clearly not homogeneous. It could be split into three classes: standard genes, genes expressed at a high level under exponential growth conditions (left ear of the 'rabbit' head) and genes coming from horizontal gene transfer (right ear of the 'rabbit' head). Correlations in the codon usage bias in genes involved in a common metabolism are illustrated in the case of histidine biosynthesis, where genes are located along a common line. This can be accounted for by compartmentalization of translation [5].

the core genes of the species [52]. The question of lateral gene transfer remains, however, one of the most controversial questions in genomics [53–56]. At present, the extent of horizontal gene transfer is not really known, since the corresponding knowledge mostly derives from genomes of a biased sample of species (in particular most are more or less readily cultivatable in the laboratory). For our purpose here, the main question that would have to be examined is whether transfer is random in the genome, or whether there exist hot spots for foreign gene insertions. Of course primary insertion could be (and probably is) random, but what we see as having been selectively stabilized may be of a different type. The study of pathogenesis indicates that the so-called 'pathogenicity islands' are not randomly distributed in genomes, but are often associated with tRNA coding loci [57–60]. In the same way, whole metabolic pathways can be transferred from species to species [61], and, generally speaking, the region of the terminus of replication is particularly rich in horizontally transferred genes. This is easily explained by the need of the cell to use a recombination system in this region to resolve the knotted structure created by chromosome circularity.

Common rules are necessary to allow gene transfer. In particular, there must exist rules for DNA packaging that would be common to organisms sharing genes. They are not yet known. Rules allowing for gene expression are needed, and, as we saw, the difference in the process of gene expression between Archaea, Bacteria and Eukarya makes a productive transfer difficult. In fact, it seems likely that DNA transfer is frequent [see for example the role of conjugation, 62]. Apparently however, gene transfer between these kingdoms is extremely rare. The only well-documented highly active process is between plants and bacteria, with the plant tumor promoting *Agrobacterium* species. The very creation of kingdoms of living organisms may have been selected through evolution as a means to screen out a number of horizontally propagating, possibly 'selfish' DNA, viruses for example.

Hence, while blurring the phylogenetic picture, and making it more difficult to discern rules of genome organization, it is likely that horizontal gene transfer has its own rules, and that, at some point, uncovering these rules will produce a clearer picture of the way in which genes move in genomes, whether in a random fashion or in an organized way.

Diversity of Genomes

Life has evolved for more than 3.5 billion years. The trio driving evolution, variation/selection/amplification, acts at the root of genome construction. The genome sequence evolves continuously: DNA polymerase cannot be totally exempt of errors, and mutations occur during the process of replication; the Earth is continuously submitted to the flow of cosmic radiation, natural radioactivity, and reactive chemical species (in particular reactive oxygen species, when oxygen, now almost ubiquitous, is available) and this modifies the chemical nature of DNA, also leading to mutations. In short, genomes must evolve, they have no way out. Diversity is, therefore, the rule. However, it is interesting to explore whether there are underlying principles that compound diversity. For example, sex is ubiquitous, and it is generally admitted that this is caused by a need to escape Muller's ratchet [63]. Are there, nevertheless, further universals that obey strong selective principles, those described above aside?

Selective Principles

What are the selective principles that tend to organize a genome? Let us start with one of the major constraints, often overlooked, that of metabolism. The first pressure exerted on genes and genomes is of a physicochemical nature. Temperature changes, for example, are impossible to escape (except for homeotherms that solved the problem at least in part). Temperature-dependent maintenance of the components of the cells is, therefore, ubiquitous. Associated with that is diffusion of gasses. In this respect, the contamination of the atmosphere by dioxygen must have played a major role in species distribution and evolution. Because of its sensitivity to electron transfers, sulfur metabolism is a case in point, and it is indeed organized in islands in bacterial genomes [64]. Because water is the basic medium for life, an increase in water entropy is a major driving force for folding structures and making planar structure important for compartmentalization [5, 28]. A genome is not an abstract entity. It is made of building blocks that require both specific features of the environment and genes. The corresponding metabolism has important consequences in terms of DNA and protein sequences [65, 66]. It will be interesting in particular to consider further the importance of metabolic cost in the evolution of proteins. Other selective principles operate at the level of gene expression [67] or gene essentiality [27, 43]. Finally, it is most likely that proteins are rarely isolated entities in the cell: they are rather grouped into complexes [48]. This is obviously reflected in the existence of operons that code for proteins that often are interacting in complexes. The extreme crowding of the cytoplasm makes that compartmentalization is essential, including for small molecules [68]. All these features result in the creation of a 'style' specific for each organism [5, 69].

Orphan Genes

Another ubiquitous feature reflects the self of the organism. A remarkable observation stems from the multiplication of genome sequences now available. As new genomes are sequenced, the number of genes that do not have a counterpart in another genome does not really go to zero. In fact, it appears that some 10% or so of the genes of any genome do not look like anything

known. These orphan genes are not only not similar to other genes, but they do not have a clear function, since their inactivation does not yield a recognizable phenotype. While some of these genes are clearly of phage descent – they are clustered together with some genes with similarities to known phage proteins [70] – most cannot be formally linked to phages. Another suggestion has been that these genes are pseudogenes, resulting from a past lateral transfer, and in the course of disappearing [71]. Gene expression profiling, however, demonstrates that many are expressed and vary as a function of environmental conditions. What could be the function of these genes? A tentative suggestion came from the study of obligatory parasites: because the traces of horizontal transfer were absent, authors concluded that ancestral genes of orphan genes were present in the last common ancestor of gammaproteobacteria, and lost afterward in many lineages [72]. This sounds unlikely, in view of the fact that such genes – at least until now – have not been associated with phenotypes. It is more likely that cells can systematically create new genes, and that these genes are in fact optimizing existing functions. We should indeed remember that life is also a competition between individuals, and that stability is the most important trait selected during evolution: those organisms that are slightly more stable will have better chances to survive, and this conclusion, of course, also applies to protein complexes.

Tentative Conclusions

Bacterial genomes are diverse. However, there are rules governing this diversity. In the prokaryotic world the first dichotomy is between Archaea and Bacteria, and this separation results in a very limited possibility for gene exchange between both kingdoms. However, after this major diversity (reflected in distinct styles for different organisms) has been taken into account, rules of organization begin to emerge. Despite extreme differences in the way in which DNA is handled in bacteria, the way in which genes are distributed in the chromosome is far from random. Genes are grouped into operons, metabolic or pathogenicity islands, and these reflect the need to compartmentalize protein complexes inside the cell. Exploring phylogenies will benefit from better understanding the selective constraints that may link genome organization and the cell's architecture. The Hox gene complexes in multicellular Eukarya [73–75] are an intriguing observation that must be transposed to the cell's organization.

Acknowledgments

This work was supported by the BioSapiens program contract LSHG-CT-2003-503265.

References

1 Campo N, Dias MJ, Daveran-Mingot ML, Ritzenthaler P, Le Bourgeois P: Genome plasticity in *Lactococcus lactis*. Antonie Van Leeuwenhoek 2002; 82: 123–132.
2 Dobrindt U, Hentschel U, Kaper JB, Hacker J: Genome plasticity in pathogenic and nonpathogenic enterobacteria. Curr Top Microbiol Immunol 2002; 264: 157–175.
3 Dobrindt U, Hacker J: Whole genome plasticity in pathogenic bacteria. Curr Opin Microbiol 2001; 4: 550–557.
4 Changeux JP, Courrege P, Danchin A: A theory of the epigenesis of neuronal networks by selective stabilization of synapses. Proc Natl Acad Sci USA 1973; 70: 2974–2978.
5 Danchin A: The Delphic Boat. What Genomes Tell Us. Cambridge, Harvard University Press, 2003.
6 Yockey HP: Information Theory and Molecular Biology. Cambridge, Cambridge University Press, 1992.
7 Turing A, Wilkinson JH (ed. Copeland BJ): The Automatic Computing Engine, Lectures given at the Ministry of Supply, December 1946 and January 1947, in Machine Intelligence 15:381–444, K. Furukawa, D. Michie, S. Muggleton (eds.), Oxford University Press, 1999.
8 von Neumann J: The Computer and the Brain. New Haven, Yale University Press, 1958.
9 Lakatos I: The Methodology of Scientific Research Programmes. Cambridge, Cambridge University Press, 1980, vol 1: Philosophical Papers.
10 Woese CR: The evolution of cellular tape reading processes and macromolecular complexity. Brookhaven Symp Biol 1972; 23: 326–365.
11 Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A: Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 1991; 222: 851–856.
12 Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KH: Viable offspring derived from fetal and adult mammalian cells. Nature 1997; 385: 810–813.
13 Tempesti G, Mange D, Stauffer A: The Embryonics Project: a machine made of artificial cells. Riv Biol 1999; 92: 143–188.
14 Teuscher C, Mange D, Stauffer A, Tempesti G: Bio-inspired computing tissues: towards machines that evolve, grow, and learn. Biosystems 2003; 68: 235–244.
15 Silberschatz A, Galvin PB, Gagne G: Operating System Concepts, ed 6. New York, Wiley, 2001.
16 Woese CR, Magrum LJ, Fox GE: Archaebacteria. J Mol Evol 1978; 11: 245–251.
17 Gupta RS: Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev 1998; 62: 1435–1491.
18 Mayr E: Two empires or three? Proc Natl Acad Sci USA 1998; 95: 9720–9723.
19 Woese CR: Default taxonomy: Ernst Mayr's view of the microbial world. Proc Natl Acad Sci USA 1998; 95: 11043–11046.
20 Rocha E: Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 2002; 10: 393–395.
21 Hogg T, Mechold U, Malke H, Cashel M, Hilgenfeld R: Conformational antagonism between opposing active sites in a bifunctional RelA/SpoT homolog modulates (p)ppGpp metabolism during the stringent response. Cell 2004; 117: 57–68.
22 Bozinovski S, Jovancevski G, Bozinovska N: DNA as a Real Time, Database Operating System. International Conference on Information Systems, Analysis and Synthesis, La Habana, 2001.
23 Danchin A: Complete genome sequencing: future and prospects; in Goffeau A (ed): BAP 1988–1989. Brussels, Commission of the European Communities, 1988, pp 1–24.
24 Mushegian AR, Koonin EV: A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 1996; 93: 10268–10273.
25 Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Debarbouille M, Dervyn E, Deuerling E, Devine K, Devine SK, Dreesen O, Errington J, Fillinger S, Foster SJ, Fujita Y, Galizzi A, Gardan R, Eschevins C, Fukushima T, Haga K, Harwood CR, Hecker M, Hosoya D, Hullo MF, Kakeshita H, Karamata D, Kasahara Y, Kawamura F, Koga K, Koski P, Kuwana R, Imamura D, Ishimaru M, Ishikawa S, Ishio I, Le Coq D, Masson A, Mauel C, Meima R, Mellado RP, Moir A, Moriya S, Nagakawa E, Nanamiya H, Nakai S, Nygaard P, Ogura M, Ohanan T, O'Reilly M, O'Rourke M, Pragai Z, Pooley HM, Rapoport G, Rawlins JP, Rivas LA, Rivolta C, Sadaie A, Sadaie Y, Sarvas M, Sato T, Saxild HH, Scanlan E, Schumann W, Seegers JF, Sekiguchi J, Sekowska A, Seror SJ, Simon M, Stragier P, Studer R, Takamatsu H, Tanaka T, Takeuchi M, Thomaides HB, Vagner V, van Dijl JM, Watabe K, Wipat A, Yamamoto H, Yamamoto M,

Yamamoto Y, Yamane K, Yata K, Yoshida K, Yoshikawa H, Zuber U, Ogasawara N: Essential *Bacillus subtilis* genes. Proc Natl Acad Sci USA 2003; 100: 4678–4683.
26 Koonin EV: Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 2003; 1: 127–136.
27 Rocha EP, Danchin A: Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res 2003; 31: 6570–6577.
28 Danchin A, Guerdoux-Jamet P, Moszer I, Nitschke P: Mapping the bacterial cell architecture into the chromosome. Philos Trans R Soc Lond B Biol Sci 2000; 355: 179–190.
29 Jukes TH: On the prevalence of certain codons ('RNY') in genes for proteins. J Mol Evol 1996; 42: 377–381.
30 Shepherd JC: Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. J Mol Evol 1981; 17: 94–102.
31 Fukushima A, Ikemura T, Kinouchi M, Oshima T, Kudo Y, Mori H, Kanaya S: Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. Gene 2002; 300: 203–211.
32 Wang Y, Rocha EPC, Leung FCC, Danchin A: Cytosine methylation is not the major factor inducing CpG dinucleotide deficiency in bacterial genomes. J Mol Evol 2004; 58: 692–700.
33 Karlin S, Mrazek J, Campbell AM: Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 1997; 179: 3899–3913.
34 Medigue C, Viari A, Henaut A, Danchin A: *Escherichia coli* molecular genetic map (1500 kbp): update II. Mol Microbiol 1991; 5: 2629–2640.
35 Rocha EP, Viari A, Danchin A: Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. Nucleic Acids Res 1998; 26: 2971–2980.
36 Gilson E, Saurin W, Perrin D, Bachellier S, Hofnung M: The BIME family of bacterial highly repetitive sequences. Res Microbiol 1991; 142: 217–222.
37 Henaut A, Rouxel T, Gleizes A, Moszer I, Danchin A: Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. J Mol Biol 1996; 257: 574–585.
38 Rocha EP, Danchin A, Viari A: Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol Biol Evol 1999; 16: 1219–1230.
39 Borkovich KA, Alex LA, Yarden O, Freitag M, Turner GE, Read ND, Seiler S, Bell-Pedersen D, Paietta J, Plesofsky N, Plamann M, Goodrich-Tanrikulu M, Schulte U, Mannhaupt G, Nargang FE, Radford A, Selitrennikoff C, Galagan JE, Dunlap JC, Loros JJ, Catcheside D, Inoue H, Aramayo R, Polymenis M, Selker EU, Sachs MS, Marzluf GA, Paulsen I, Davis R, Ebbole DJ, Zelter A, Kalkman ER, O'Rourke R, Bowring F, Yeadon J, Ishii C, Suzuki K, Sakai W, Pratt R: Lessons from the genome sequence of *Neurospora crassa*: tracing the path from genomic blueprint to multicellular organism. Microbiol Mol Biol Rev 2004; 68: 1–108.
40 Lobry JR: Origin of replication of *Mycoplasma genitalium*. Science 1996; 272: 745–746.
41 Sueoka N: Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 1995; 40: 318–325.
42 Rocha EP, Danchin A, Viari A: Universal replication biases in bacteria. Mol Microbiol 1999; 32: 11–16.
43 Rocha EP, Danchin A: Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 2003; 34: 377–378.
44 Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R: Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 1981; 9: r43–74.
45 Ikemura T: Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 1985; 2: 13–34.
46 Elf J, Nilsson D, Tenson T, Ehrenberg M: Selective charging of tRNA isoacceptors explains patterns of codon usage. Science 2003; 300: 1718–1722.
47 Guerdoux-Jamet P, Henaut A, Nitschke P, Risler JL, Danchin A: Using codon usage to predict genes origin: is the *Escherichia coli* outer membrane a patchwork of products from different genomes? DNA Res 1997; 4: 257–265.
48 Nitschke P, Guerdoux-Jamet P, Chiapello H, Faroux G, Henaut C, Henaut A, Danchin A: Indigo: A World-Wide-Web review of genomes and gene functions. FEMS Microbiol Rev 1998; 22: 207–227.
49 Danchin A: Homeotopic transformation and the origin of translation. Prog Biophys Mol Biol 1989; 54: 81–86.
50 Sharp PM, Li WH: An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986; 24: 28–38.
51 Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brussow H: Phage as agents of lateral gene transfer. Curr Opin Microbiol 2003; 6: 417–424.
52 Carlson CR, Kolsto AB: A small (2.4 Mb) *Bacillus cereus* chromosome corresponds to a conserved region of a larger (5.3 Mb) *Bacillus cereus* chromosome. Mol Microbiol 1994; 13: 161–169.
53 Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF: Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 2003; 37: 283–328.
54 Kurland CG, Canback B, Berg OG: Horizontal gene transfer: a critical view. Proc Natl Acad Sci USA 2003; 100: 9658–9662.
55 Lawrence JG, Hendrickson H: Lateral gene transfer: when will adolescence end? Mol Microbiol 2003; 50: 739–749.
56 Philippe H, Douady CJ: Horizontal gene transfer and phylogenetics. Curr Opin Microbiol 2003; 6: 498–505.
57 Bertin Y, Boukhors K, Livrelli V, Martin C: Localization of the insertion site and pathotype determination of the locus of enterocyte effacement of shiga toxin-producing *Escherichia coli* strains. Appl Environ Microbiol 2004; 70: 61–68.
58 Larbig KD, Christmann A, Johann A, Klockgether J, Hartsch T, Merkl R, Wiehlmann L, Fritz HJ, Tummler B: Gene islands integrated into tRNA(Gly) genes confer genome diversity on a *Pseudomonas aeruginosa* clone. J Bacteriol 2002; 184: 6665–6680.
59 Wang Y, Wang H, Xiang Q, Sun SX, Yu SY: Detection of the high-pathogenicity island of *Yersinia enterocolitica* in enterotoxigenic and enteropathogenic *E. coli* strains. Di Yi Jun Yi Da Xue Xue Bao 2002; 22: 580–583.
60 Noel L, Thieme F, Nennstiel D, Bonas U: Two novel type III-secreted proteins of *Xanthomonas campestris* pv. *vesicatoria* are encoded within the hrp pathogenicity island. J Bacteriol 2002; 184: 1340–1348.
61 van der Meer JR, Sentchilo V: Genomic islands and the evolution of catabolic pathways in bacteria. Curr Opin Biotechnol 2003; 14: 248–254.
62 Heinemann JA, Sprague GF Jr: Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. Nature 1989; 340: 205–209.
63 Chao L: Evolution of sex and the molecular clock in RNA viruses. Gene 1997; 205: 301–308.
64 Rocha EP, Sekowska A, Danchin A: Sulphur islands in the *Escherichia coli* genome: markers of the cell's architecture? FEBS Lett 2000; 476: 8–11.
65 Akashi H, Gojobori T: Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA 2002; 99: 3695–3700.
66 Rocha EP, Danchin A: Base composition bias might result from competition for metabolic resources. Trends Genet 2002; 18: 291–294.
67 Sousa C, de Lorenzo V, Cebolla A: Modulation of gene expression through chromosomal positioning in *Escherichia coli*. Microbiology 1997; 143: 2071–2078.
68 Ellis RJ: Macromolecular crowding: obvious but underappreciated. Trends Biochem Sci 2001; 26: 597–604.
69 Grantham R, Gautier C, Gouy M, Mercier R, Pave A: Codon catalog usage and the genome hypothesis. Nucleic Acids Res 1980; 8:r49–r62.
70 Ochman H, Lawrence JG, Groisman EA: Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405: 299–304.
71 Amiri H, Davids W, Andersson SG: Birth and death of orphan genes in Rickettsia. Mol Biol Evol 2003; 20: 1575–1587.
72 Shimomura S, Shigenobu S, Morioka M, Ishikawa H: An experimental validation of orphan genes of Buchnera, a symbiont of aphids. Biochem Biophys Res Commun 2002; 292: 263–267.
73 Akam M: Hox genes, homeosis and the evolution of segment identity: no need for hopeless monsters. Int J Dev Biol 1998; 42: 445–451.
74 Averof M, Akam M: Hox genes and the diversification of insect and crustacean body plans. Nature 1995; 376: 420–423.
75 Averof M: Arthropod evolution: same Hox genes, different body plans. Curr Biol 1997; 7:R634–636.