Compositional bias in DNA

Christian Gautier

Experimental approaches, as well as computer analysis on genomic sequences, have revealed a large variability in base composition between regions in the same genome or between genomes of different species. In most cases, however, the biological causes of these compositional biases remain unknown. The recent large increase in the availibility of completely sequenced genomes can give new insight into evolution processes involved in these compositional biases.

Addresses
Biometry and Evolutionary Biology Laboratory (bâtiment 741), Université Claude Bernard Lyon 1 and CNRS, 43 bd 11 nov, 69622 Villeurbanne Cedex, France;
e-mail: cgautier@biomserv.univ-lyon1.fr

Current Opinion in Genetics & Development 2000, **10**:656–661

0959-437X/00/$ – see front matter

Abbreviations

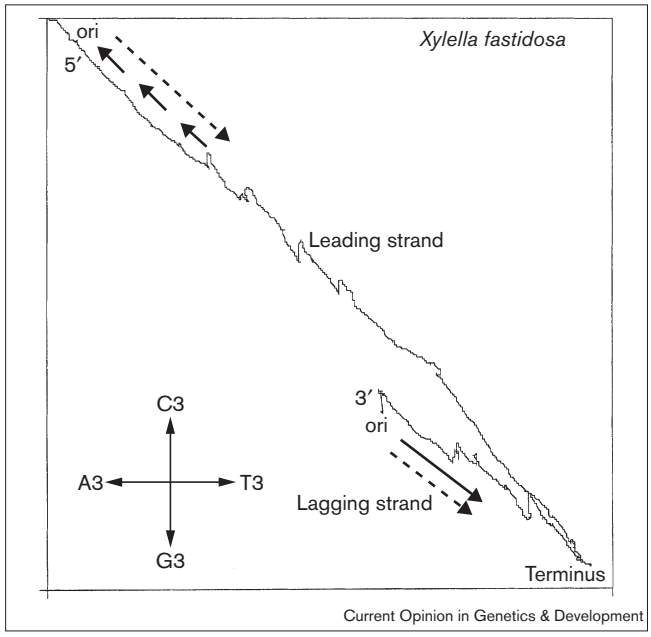
H	heavy
L	light
MHC	major histocompatibility complex

Introduction

The first compositional bias in genomic DNA that has been demonstrated is the variability of G+C content among bacterial genomes [1]. Sueoka suggested that this bias results from a mutational bias: mutation rates (C,G)→(A,T) and (A,T)→(C,G) are not equal and their ratio is species-dependent [1]. Before the emergence of neutralism, this hypothesis stated that some patterns of the genome could appear without natural selection. Since Sueoka's original hypothesis, mutational bias versus natural selection has appeared as a frequent debate when statistical structures of genomic sequences are under study. Two questions can be asked: first, can statistical properties of the genome increase organism fitness sufficiently to be selected or second, can some characteristics of the mutation/repair process generate observed structures? The proposal that an increase in G+C content would allow species to develop at higher temperature because of greater genome stability is an example of the selectionist point of view. The increase in genome data, however, has served to weaken this theory [2]. Conversely, only one mutation capable of modifying G+C content has been described [3] and no direct demonstration of mutational bias is available. An argument in favour of mutational bias is the fact that no life-history trait common to bacteria having the same G+C content, has yet been found.

Codon usage is a statistical property of the genome that has been studied since the first complete sequence became available [4]. Strong relationships between gene expressivity and preferential codon usage related to tRNA having high cellular frequencies have been shown for *Escherichia coli* [5,6], yeast [7] and more recently for *Bacillus subtilis* [8••]. This is a very strong argument for the existence of selection for codon usage in these organisms. For most organisms, however, data on cellular tRNA frequencies are not available and selection for codon usage has been inferred from correlation between a measurement of codon usage bias and expressivity [9–11]. The availability of complete genome sequences now permits a more direct approach using the number of genes for a tRNA to estimate its cellular frequency [8••,12].

Figure 1



Compositional strand asymmetry: a DNA walk. The complete sequence of *X. fastidiosa* is represented by reading the sequence in the third codon positions and walking into the plane according to the four directions indicated on the bottom left of the figure. Dashed arrows indicate the progression of the replication fork; solid arrows indicate the displacement of polymerase during replication. Replication of the leading strand results in the new lagging strand and implies synthesis of small fragments (Okazaki fragments). Replication of the lagging strand is a continuous process. Detail on the method can be found in [18], another cumulative method has been described in [19].

Several reviews exist on G+C content in bacteria [13••] and codon usage [14–16]. In this review, I focus on two other strong statistical departures of genomic sequence from 'random': the asymmetry between genomic strands in bacteria and the isochore organisation of vertebrate genomes.

Strand asymmetry in bacterial genomes (chirochores)

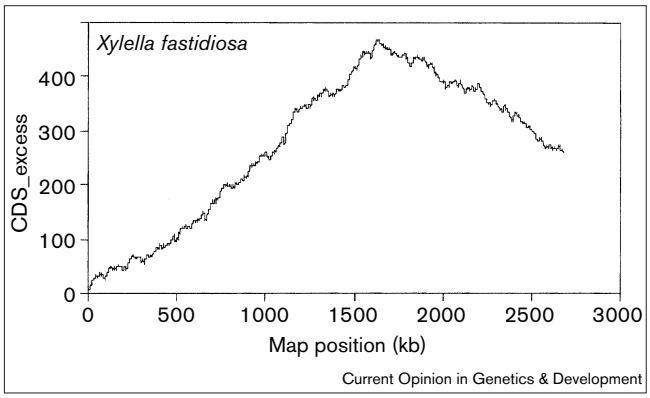
Strand asymmetries in bacteria are summarised in Figures 1 and 2 by the analyses of the *Xyllela fastidiosa* genome [17]

(the last complete genome published before this review). The DNA walk (Figure 1; [18,19]) demonstrates that the origin and terminus of replication separate the genome in two regions that differ in their base composition. Namely the leading strand is G+T-rich and the lagging strand is C+A-rich. Figure 2 indicates that genes are preferentially directed such that their transcription processes occur in the same direction as the replication of the genome.

In summary, bacterial genomes must meet criteria linked to the following: first, G+C content; second, codon usage caused by translation constraints; third, base compositional asymmetry; and fourth, preferential gene orientation (see [20,21•] for review). Moreover the amino-acid composition of proteins may also be a constraint for the genome. Mathematical developments as well as statistical analysis of complete bacterial genomes have help greatly in understanding relationships between these four constraints.

The asymmetry in gene direction is generally considered as resulting from natural selection acting to avoid head-on collisions between replication and transcription [22]. Existence of selective pressure is confirmed as asymmetry is stronger if only highly expressed genes are taken into account [20]. Pressure leading to compositional asymmetry is more difficult to determine. Compositional asymmetry implies non-identity between the substitution processes on the two strands. If the two processes where the same, at equilibrium, base composition on each strand must verify A=T and C=G [23,24]. This is not the case as, for most bacteria, the leading (respectively lagging) stand is characterised by G→C and T→A (respectively C→G and A→T) [25]. The overall G+C content varies greatly among bacterial genomes (see 'Introduction'), however, implying that the substitution matrix has changed often during evolution and the equilibrium condition may be doubtful. It can be demonstrated [26] that if the substitution matrix fulfils the requirements needed to ensure the same process on the two strands, convergence toward G=C and A=T continues even if the matrix is modified. The observed asymmetry is very clearly incompatible with an identical substitution process in the two strands. As replication is known to be different on the two strands, strand asymmetry has been proposed to result from the functioning of this molecular process. As shown in Figure 1, replication works continuously when the lagging strand is replicated (to generate the leading strand) but works discontinuously when the leading strand is processed (and the lagging strand generated). Emphasis has been placed particularly on the fact that the leading strand remains single-stranded longer than the lagging one. This may favour some particular mutations as the deamination of C and thus mutation C→T (see review in [21•]), which is coherent with the structure, depicted in Figure 1. A similar effect is observed for vertebrate mitochondria [27], even if the replication process in mitochondria is different. Confusing factors may result, however, from other constraints such as G+C content, preferential orientation of genes, codon usage or protein

Figure 2



Gene direction plotted along *X. fastidiosa* genome. Cumulative plotting of gene direction – excess of CDS in which the transcribed strand has been published – shows a majority of genes in which transcription occurs in the same direction as replication-fork progression. The maximum point of the graph corresponds to the terminus.

characteristics. In fact, it appears that these different genomic characteristics result from different pressures as shown for C+G content and asymmetry [13••,28•] or compositional asymmetry and gene orientation [29•]. Their effects can, however. interact as shown for example by protein constraints caused by codon usage [30], genomic G+C content [31] or strand asymmetry [32,33,34•].

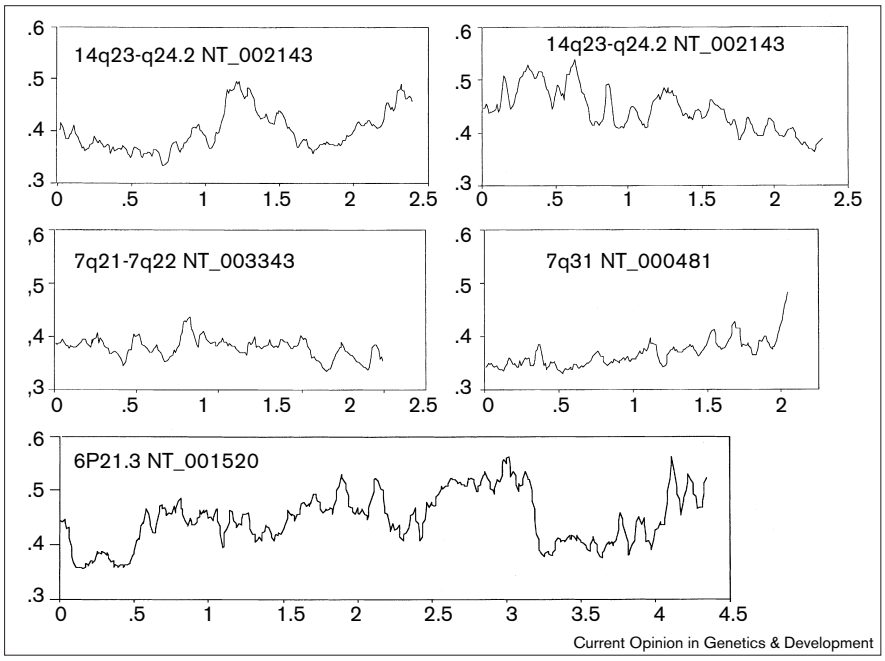
Bacterial genomes are therefore patterned strongly by either selective or neutral pressures, acting at different levels: transcription, translation, and replication. These pressures occur independently of one of the main functions of genomes: coding for proteins. Moreover, they constrain this role. Finally, it must be said that, whatever the pressure may be that maintains the asymmetry of the genome, this asymmetry may be used in a functional capacity, as seems to be the case for the expression of *dif* during *Escherichia coli* replication [35].

Isochores

Isochore description

Isochores were originally identified as a result of a gradient density analysis of fragmented genomes [36]: mammalian genomes are a mosaic of regions having different homogeneous G+C content. Higher-density level genomic segments are named heavy (H) isochores and lower-density level segments are light (L) isochores. During the past ~25 years, this definition has evolved by synergy between experimental approaches and genomic bioinformatics. If the G+C content remains the basic definition of isochores, they have been associated, very surprisingly, with a complex set of biological properties, a priori largely unrelated. Because very complete reviews have already been published [37]; a brief summary of isochores properties is as follows. First, genes are involved in isochore structure: G+C content of all three codon positions, introns and flanking sequences vary accordingly with the isochore class to which

Figure 3



G+C profile along the longest human contigs. Sequences have been retrieved from the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov:80/entrez/); they correspond to the longest contigs, with the exception of complete chromosomes (21 and 22). Cytogenetic localisation and accession number are shown on each plot. The last sequence is the MHC region that has been extended toward centromere since publication cited in the principal text. These profiles are fully compatible with isochore organisation with long L isochores (>2Mb for NT_003343) and the detection of a new structure made of long gradients with a weak but significant slope (p-value $<10^{-4}$) (NT_000481,NT001572).

the gene belongs. The stronger effect is on codon position 3 with a range of [0.28–0.95] for ~95% of human genes. The amino-acid content of protein is also constrained by isochore class, amino-acids encoded by G+C-rich codons (alanine, arginine, glycine, and proline) are more frequent in H isochores [38–40]. Second, the density of genes is higher in H isochores than in L ones. Genes in H isochores are more compact with a smaller fraction of intronic sequences and code for shorter proteins than do genes in L isochores [41,42]. Third, SINE (short-interspersed nuclear element) sequences and particularly *Alu* are preferentially found in H isochores, LINE (long-interspersed nuclear element) sequences are preferentially found in L isochores. More generally, inserted sequences are isopycnic with a C+G content similar to their insertion context [43–46]. Fourth, chromosome banding is related to isochore organisation: Giemsa bands are made of L isochores, the reverse band being more heterogeneous (H and L isochores). T bands, often found at the telomeric end of chromosomes, are made of H isochores [47–52]. Isochore organisation is therefore related to base composition, gene characteristics, repeated sequence insertion sites, cytogenetic properties and, finally, genetic characteristics.

Sequences from the Human Genome Project have recently allowed the direct analysis of G+C content variation at the megabase scale along the genome. Analyses of Chromosomes 21 [53•] and 22 [54••] are coherent with the preceding description of isochores even if data is presently unavailable for testing precisely the fourth point. An extensive analysis of the MHC region [55,56•,57•] gives the same conclusion. Figure 3 shows G+C content variation along the longest contigs of the human working draft, in the exception of complete chromosomes.

Isochore taxonomic range

Variations of G+C content along bacterial genome have been described in most taxons. Isochore definition, however, is restricted to the whole pattern described above and shall be considered here only in vertebrates, even if similar structures have been described in monocots [58]. The main results are as follows. First, isochores are found unambiguously in mammals and birds but not in anoura and fishes [59]. Second, isochores in reptiles are controversial: it is the only instance for which density gradient analysis and biocomputing on genomic sequences are at odds. Although Olmo has found isochores in some reptiles by density gradient analysis [60], Bernardi in several extensive studies among vertebrate taxa (e.g. see [59]) limits the presence of isochores to birds and mammals and explains this discrepancy by the non-detection of satellite sequences. More recently, sequencing analyses of Hughes *et al.* [61•] provide a strong argument for an isochore organisation of crocodile and turtle but not snake genomes. As will be discussed later, this point is of major importance in the understanding of isochore evolution. Third, isochore organisation is highly conserved among mammals despite the numerous chromosomal rearrangements that have taken place during mammalian evolution. A specific pattern, however, is found in murids [62]. In this taxon, the variance of G+C content among isochore classes has been strongly reduced; this modification implies all position of genes and, thus, the amino-acid composition of proteins [63].

Isochore evolution

This hypothesis would clearly be falsified if the presence of isochores in some reptile genomes is confirmed either by new sequence analyses or by new experimental data. In this case, new relations between isochores and organism fitness must be found to support the selectionist point of view. The first is that isochores are formed by selection pressure. The main argument for this has been the parallel between isochores and homeothermy [37]. Until recently, the isochore was clearly limited to homeotherms (mammals and birds): an increase in C+G content in some part of the genome can be interpreted as a solution to the problems posed by high temperature. It must be said that such relationships between C+G content and temperature do not exist in bacteria (cf. 'Introduction'). Clearly, the presence of isochores in some reptile genomes being confirmed [61•], this hypothesis has now been falsified and another relationship of isochore presence to organism fitness must be found. However, indirect arguments in favour of a selection pressure have been published. Some arguments exist relating H isochores to highly recombinant regions, particularly in telomeric regions. As a higher recombination rate can allow a more efficient selection pressure, this could explain a patterning of isochore by both selection for high C+G content and recombination rate variation along the genome [64], even if other hypotheses, particularly gene conversion, can explain these relationships [65]. Variability of isochore organisation among mammals also creates certain issues for debate from the selectionist point of view. It has been estimated that the ancestor at the base of the murid lineage has a human pattern [66], therefore the murid lineage has undergone a genome homogenisation. This could indicate a decrease in selection efficiency coherent with the increase of mutation rate that is postulated for murids [67]. This last argument originated from a comparison between polymorphism and C+G content [68••]. In this study, Eyre-Walker shows that polymorphism in MHC is not compatible with a stationary process and two hypotheses are explored: selection for C+G content and biased gene conversion. Selection and neutral evolution cannot, therefore, be discriminated.

The second hypothesis is that isochores result from mutational biases: the main argument for this is that the small effective size of vertebrate populations implies that only high fitness increase can be selected and that is not compatible with a selection acting on one non-coding base (e.g. see [69]). The alternative to selection is a mutational bias (see 'Introduction') but at present no biological characteristics of the mutation/repair process are known to lead to an isochore pattern. A very elegant model was proposed that relates variation of mutational bias to both the difference in replication time for the different isochore classes and on different nucleotide pools that would be available in the different stage of cellular division [70] but its predictions on substitution rates have been proven incorrect (e.g. see [71]). Francino and Ochman [72] predicted that selection acting on C+G content must imply a reduced substitution rate in an H isochore submitted to purifying selection. As the opposite is found for two globin pseudogenes, they conclude that isochores result from mutation and not selection. Their analysis does not take into account the large variability of mutation rate inside mammalian genome. This variability, independent from isochore structure, could explain the observed difference between substitution rate independently of presence or absence of selection [71]. Another study on pseudogenes concludes that there was an absence of selection [73]. The authors made a careful estimation of substitutions that have occurred since the integration of several retropseudogenes originated from the same functional gene. This CG→AT substitution occurs in excess, corresponding to an insertion environment AT-rich of moderately CG-rich sequences. Insertion of the same sequence in an H isochore would have been an interesting comparison.

If the pioneering work of Bernardi and co-workers has been fully confirmed by recent advances in complete genome sequencing, the evolutionary mechanisms that have created them remain largely unknown. The most intriguing factor remains the correlation between variations of very different biological parameters. Understanding the mechanisms of these correlations is the key for reconstructing a coherent isochore evolutionary history. In this scope, building genomic maps including both compositional bias and a functional signal may be the first step toward understanding the pressures that have generated and maintained isochores [74].

Conclusions

Base composition as well as recombination and mutation rate divide genomes in a complex mosaic. In the near future, the availability of complete genome sequences will allow one to draw compositional and mutational maps of genomes. Such maps are the necessary, but not sufficient, condition for functional and evolutionary interpretations of this genome heterogeneity. They must be associated to a better knowledge of genetic (recombination etc.) and molecular (replication, transcription etc.) processes that manage the genome. Taking simultaneously into account genetic information written in the genome, evolutionary mechanisms particularly at the level of populations and ecology and genome functioning constraints are necessary for synthesising a view of genome evolution. As a result of their relatively simple structure, compositional biases provide a promising approach for integrative evolutionary studies.

Update

A very clear and possibly interesting complement to this present review has just been published: a review on the base composition of genomes and its implication for phylogenetic reconstruction [75].

Acknowledgements

I thank J Lobry for providing Figures 1 and 2, and B Spataro for Figure 3 (in each case original software they have developed was used). I thank D Mouchiroud for critically reading this article and L Duret, I Goncalves, S Hughes, and G Piganeau for discussions.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Sueoka N: **On the genetic basis of variation and heterogeneity of DNA base composition.** *Proc Natl Acad Sci USA* 1962, **34**:95-114.

2. Galtier N, Lobry JR: **Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes.** *J Mol Evol* 1997, **44**:632-636.

3. Yanofsky C, Cox EC, Horn V**: The unusual mutagenic specificity of an *E. coli* mutator gene.** *Proc Natl Acad Sci USA* 1966, **55**:274-281.

4. Grantham R: **Codon base randomness and composition drift in coliphage.** *Nat New Biol* 1972, **237**:265-266.

5. Ikemura T: **Correlation between the abundance of *E. coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translation system.** *J Mol Biol* 1981, **158**:573-597.

6. Gouy M, Gautier C**: Codon usage in bacteria: correlation with gene expressivity**. *Nucleic Acids Res* 1982, **10**:7055-7073.

7. Ikemura T: **Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes.** *J Mol Biol.* 1982, **158**:573-597.

8. •• Kanaya S, Yamada Y, Kudo Y, Ikemura T**: Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtillis* tRNAs: gene expression level and species–species diversity of codon usage based on multivariate analysis.** *Gene* 1999, **238**:143-155.

The cellular levels of individual tRNAs in *Bacillus subtilis* were determined experimentally and are found to be proportional to the number of their respective genes. Moreover, a clear constraint of tRNA contents on synonymous codon choice is shown. For 18 organisms whose genomes have been sequenced a relation is shown between the number of tRNA genes (and, thus, their cellular content) and codon usage. Moreover, the tRNA set varies among organisms, particularly in the function of genomic G+C content.

9. Shield DC, Sharp PM, Higgins DG: **'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons.** *Mol Biol Evol* 1988, **5**:704-716.

10. Chiapello H, Lisacek F, Caboche M, Henaut A: **Codon usage and gene function are related in sequences of *Arabidopsis thaliana*.** *Gene* 1998*,* **209**:GC1-GC38.

11. Duret L, Mouchiroud D: **Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*.** *Proc Natl Acad Sci USA* 1999, **96**:4482-4487.

12. Duret L: **tRNA gene number and codon usage in *C. elegans* genome are co-adapted for the optimal translation of highly expressed genes.** *Trends Genet* 2000, **16**:287-289.

13. •• Sueoka N: **Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A=T and C=G.** *J Mol Evol* 1999, **49**:49-62.

For each of 10 species, a pattern of strand asymmetry is built by taking into account the two skews (C/G and A/T) for each amino acid. When combined, these patterns appear to have no effect on C+G content.

14. Sharp PM, Matassi G: **Codon usage and genome evolution.** *Curr Opin Genet Dev* 1994, **4**:851-860.

15. Akashi H, Eyre-Walker A: **Translational selection and molecular evolution.** *Curr Opin Genet Dev* 1998, **8**:688-693.

16. Kreitman M, Comeron J: **Coding sequence evolution.** *Curr Opin Genet Dev* 1999, **9**:637-641.

17. Simpson AJG, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LM, Araya JE, Baia GS, Baptista CS *et al.*: **The genome sequence of the plant pathogen *Xylella fastidiosa*.** *Nature* 2000, **406**:151-159.

18. Frank AC, Lobry JR: **Oriloc : prediction of replication boundaries in unannotated bacterial chromosomes.** *Bioinformatics* 2000, **16**:560-561.

19. Grigoriev A: **Analysing genomes with cumulative skew diagrams**. *Nucleic Acids Res* 1998, **26**:2286-2290.

20. Francino MP, Ochman H: **Strand asymetries in DNA evolution.** *Trends Genet* 1997, **13**:240-245.

21. • Franc AC, Lobry JR: **Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms.** *Gene* 1999, **238**:65-77.

A recent and complete review on strand asymmetry.

22. Brewer BJ: **When polymerase collide: replication and the transcriptional organization of the *E. coli* chromosome.** *Cell* 1988, **53**:679-686.

23. Sueoka N: **Intrastrand parity rules of DNA base composition and usage biases of synonymous codons.** *J Mol Evol* 1995, **40**:318-325.

24. Lobry JR: **Properties of a general model of DNA evolution under no-strand-bias conditions.** *J Mol Evol* 1995, **40**:326-330.

25. Lobry JR: **Asymetrical substitution patterns in the two DNA strands of bacteria.** *Mol Biol Evol* 1996, **13**:660-665.

26. Lobry JR, Lobry C: **Evolution of DNA base composition under no-strand-bias conditions when the subtitution rates are not constant.** *Mol Biol Evol* 1999, **16**:719-723.

27. Reyes A, Gissi C, Pesole G, Saccone C: **Asymmetrical directional mutation pressure in the mitochondrial genome of mammals.** *Mol Biol Evol* 1998, **15**:957-966.

28. • Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH : **Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases.** *Nucleic Acids Res* 1999, **27**:1642-1649.

The authors compared orthologous genes between the two spirochaetes *Borellia burgdorferi* and *Treponema pallidum*. They show that genes that are on different strands in the two species differ in base and amino acid composition, which is consistent with constraints caused by strand asymmetry.

29. • Tillier ERM, Collins RA: **The contribution of replication orientation, gene direction, and signal sequences to base composition asymmetries in bacterial genomes.** *J Mol Evol* 2000, **50**:249-257.

The authors of this paper describe a new and efficient use of ANOVA in genomics. Using as factor 1 the strand and as factor 2 the direction of the gene transcription, the authors show that both factors have a significant effect on the skew and that these effects are mostly independent.

30. Lobry JR, Gautier C: **Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes**. *Nucleic Acid Res* 1994, **22**:3174-3180.

31. Lobry JR: **Influence of genomic G+C content on average amino-acid composition of protein from 59 bacterial species.** *Gene* 1997, **205**:309-316.

32. Mclean MJ, Wolfe KH, Devine KM: **Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes.** *J Mol Evol* 1998, **47**:691-696.

33. McInerney JO: **Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi***. *Proc Natl Acad Sci USA* 1998, **95**:10698-10703.

34. • Rocha EPC, Danchin A, Viari A: **Universal replication biases in bacteria**. *Mol Microbiol* 1999, **32**:11-16.

The authors use discriminant analysis to carefully detect factors that are related to strand asymmetry in 15 complete bacterial chromosomes. They show that strand asymmetry is related to modification of base composition, codon usage and amino acid composition of proteins.

35. Pérals K, Coornet F, Merlet Y, Delon I, Louarn JM: **Functional polarization of the *Escherichia coli* chromosome terminus: the *dif* site acts in chromosome dimer resolution only when located between long stretches of opposite polarity.** *Mol Microbiol* 2000, **36**:33-43.

36. Thiery JP, Macaya G, Bernardi G**: An analysis of eukaryotic genomes by density gradient centrifugation.** *J Mol Biol* 1976, **108**:219-235.

37. Bernardi G: **Isochores and the evolutionary genomics of vertebrates**. *Gene* 2000, **241**:3-17.

38. Aota S, Ikemura T: **Diversity in G+C content at the third position of codons in vertebrate genes and its cause.** *Nucleic Acids Res* 1986, **14**:6345-6355.

39. D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G: **Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins.** *J Mol Evol* 1991, **32**:504-510.

40. Clay O, Caccio S, Zoubak S, Mouchiroud D, Bernardi G: **Human coding and non coding DNA: copositional correlations**. *Mol Phyl Evol* 1996, **1**:2-12.

41. Zoubak S, Clay O, Bernardi G**: The gene distribution of the human genome.** *Gene* 1996, **174**:95-102.

42. Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, Bernardi G: **The distribution of genes in the human genome.** *Gene* 1991, **100**:181-187.

43. Soriano P, Meunier-Rotival M, Bernardi G: **The distribution of interspersed repeats is non uniform and conserved in the mouse and human genomes**. *Proc Natl Acad Sci USA* 1983, **80**:1816-1820.

44. Rynditch A, Kadi F, Geryk J, Zoubak S, Svoboda J, Bernardi G**: The isopycnic, compartmentalized integration of Rous sarcoma virus sequences.** *Gene* 1991, **106**:165-172.

45. Zoubak S, Richardson JH, Rynditch A, Höllsberg P, Hafler DA, Boeri E, Lever AML, Bernardi G: **Regional specificity of HTLV-I proviral integration in the human genome.** *Gene* 1994, **143**:155-163.

46. Jabbari K, Bernardi G: **CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families.** *Gene* 1998, **224**:123-127.

47. Ikemura T, Aota S: **Global variation in G+C content along vertebrate genome DNA. Possible correlation with chromosome band structures.** *J Mol Biol* 1988, **203**:1-13.

48. Ikemura T, Wada K: **Evident diversity of codon usage pattern of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data.** *Nucleic Acids Res* 1991, **19**:4333-4339.

49. De Sario A, Aïssani B, Bernardi G**: Compositional properties of telomeric regions from human chromosomes.** *FEBS Lett* 1991, **295**:22-26.

50. Saccone S, De Sario A, Wiegant J, Raap AK, Valle GD, Bernardi G**: Correlation between isochore and chromosomal bands in the human genome.** *Proc Natl Acad Sci USA* 1993, **90**:11929-11933.

51. Saccone S, Caccio S, Kusuda J, Andreozzi L, Bernardi G: **Identification of the gene-richest bands in human chromosomes.** *Gene* 1996, **174**:85-94.

52. Eyre-Walker A: **Recombination and mammalian genome evolution.** *Proc R Soc Lond B* 1993, **252**:237-243.

53. • Hattori H, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, Toyoda A, Ishii K, Totoki Y, Choi DK *et al.*: **The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium.** *Nature* 2000, **405**:311-319.

Chromosome 21 contains few genes, this is coherent with the low G+C content of the chromosome (40.9%). There is a 7 Mb region with very low G+C content (35%) that correlates with a paucity of both Alu sequences and genes.

54. •• Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ *et al.*: **The DNA sequence of human chromosome 22.** *Nature* 1999, **402**:489-496.

The mean G+C content of the human chromosome 22 sequence is 47.8%. This is significantly higher than the G+C content calculated for the sum of all human genomic sequence determined to date (42%). The complete pattern of isochore organisation is discussed and validated using these data. Relationships between genetic and physical distances are also studied.

55. Fukagawa T, Sugaya K, Matsumoto K, Okumura K, Ando A, Inoko H, Ikemura T: **A boundary of long range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary.** *Genomics* 1995, **25**:184-191.

56. • The MHC sequencing consortium**: Complete sequence and gene map of a human major histocompatibility complex.** *Nature* 1999, **401**:921-923.

The sequence of 3.6 Mb of MHC region is here reported and isochore organisation is studied (see also our Figure 3) and correlated to replication timing.

57. • Stephens R, Horton R, Humphay S, Rowen L, Trowsdale J, Beck S: **Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC**. *J Mol Biol* 1999, **291**:789-799.

G+C profile is shown for the telomeric end of MMHC region showing sharp transition between an L isochore and its two neighbouring H isochores. Discussion of relationships with banding is presented.

58. Matassi G, Montero LM, Salinas J, Bernardi G: **The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants.** *Nucleic Acids Res* 1989, **17**:5273-5291.

59. Bernardi G, Bernardi G**: Compositional patterns in the nuclear genome of cold-blooded vertebrates.** *J Mol Evol* 1990, **31**:265-281.

60. Olmo E: **Evolution of genome size and DNA base composition in reptiles.** *Genetica* 1981, **57**:39-50.

61. • Hughes S, Zelus D, Mouchiroud D: **Warm-blooded isochore structure in Nile crocodile and turtle.** *Mol Biol Evol* 1999, **16**:1521-1527.

Ten genes were sequenced in the crocodile (*Crocodylus niloticus*) and six in the red-eared slider turtle (*Trachemys scripta elegans*). Some of this genes appears to have a very high G+C content in codon position 3 (four above 70% and one above 90% for the crocodile) and a strong correlation exists between the G+C content in codon position 3 in crocodile, turtle, human and chicken. Hence, although few genes are available, these results strongly suggest the existence of an isochore organisation in crocodile and turtle.

62. Mouchiroud D, Gautier C, Bernardi G: **The compositional distribution of coding sequences and DNA molecules in human and murids.** *J Mol Evol* 1988, **27**:311-320.

63. Mouchiroud D, Gautier C: **Codon usage changes and sequence dissimilarity between human and rat.** *J Mol Evol* 1990, **31**:81-91.

64. Charlesworth B: **Pattern in the genome.** *Curr Biol* 1994, **4**:182-184.

65. Wu CI, Li WH: **Evidence for higher rates of nucleotide substitution in rodents than in man.** *Proc Natl Acad Sci USA* 1985, **82**:1741-1745.

66. Eyre-Walker A: **Recombination and mammalian genome evolution.** *Proc R Soc Lond B* 1993, **252**:237-243.

67. Galtier N, Mouchiroud D**: Isochore evolution in mammals: a human-like ancestral structure.** *Genetics* 1998, **150**:1577-1584.

68. •• Eyre-Walker A: **Evidence of selection on silent site base composition in mammals: potential implications for the evolution of junk DNA.** *Genetics* 1999, **152**:675-683.

If G+C equilibrium is reached, GC→AT and AT→GC substitution must be equal. Simplifying the argument, it can be concluded that, in the absence of bias in the fixation process, this equality must also be true for mutations. The author uses this argument to show the existence of a fixation bias for the MHC genes, which could take the form of either selection pressure or conversion bias. This appears to be a strong argument in favour of the selectionist point of view. This approach is very interesting as it associates population genetics and genome evolution.

69. Shields DC, Sharp PM, Higgins DG, Wright F: **'Silent' sites in *Drosophila* are not neutral: evidence of selection among synonymous codons.** *Mol Biol Evol* 1988, **5**:704-716.

70. Wolfe KH, Sharp PM, Li WH: **Mutation rates differ among regions of the mammalian genome.** *Nature* 1989, **337**:283-285.

71. Matassi G, Sharp P, Gautier C: **Chromosomal location effects on gene sequences evolution in mammals**. *Curr Biol* 1999, **9**:786-791.

72. Francino MP, Ochman H: **Isochores results from mutation not selection.** *Nature* 1999, **400**:30-31.

73. Casane D, Boissinot S, Chang BHJ, Shimmin LC, Li WH: **Mutation pattern variation among regions of the primate genome.** *J Mol Evol* 1997, **45**:216-226.

74. Tenzen T, Yamagata T, Fukagawa T, Sugaya K, Ando A, Inoko H, Gojobori T, Fujiyama A, Okumura K, Ikemura T: **Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex.** *Mol Cell Biol* 1997, **17**:4043-4050.

75. Mooers AO, Holmes EC: **The evolution of base composition and phylogenetic inference.** *Trends Ecol Evol* 2000, **15**:365-369.