Year 2000

MÉMOIRE

Présenté devant

l'Université Claude Bernard - Lyon 1

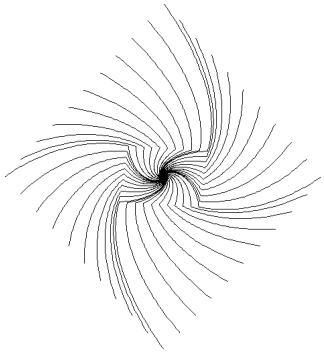
pour l'obtention de

L'HABILITATION À DIRIGER DES RECHERCHES

par

J.R. LOBRY

THE BLACK HOLE OF SYMMETRIC MOLECULAR EVOLUTION



Defended on 20-JUL-2000

Jury:

ANDERSSON, S.G.E.	(reviewer)
DANCHIN, A.	(reviewer)
GAUTIER, C.	(president)
LOUARN, J.-M.	
SUEOKA, N.	(reviewer)

CNRS UMR 5558 « Biométrie, Biologie Évolutive » Université Claude Bernard - Lyon 1

To Clara Mélina Hwa-Yung,
Thomas Eugène Kang-Min,
And their Mother Florence,
With Love.

ACKNOWLEDGMENTS

- The research corresponding to this "Habilitation à Diriger des Recherches" degree was done while I was an Assistant/Associate Professor at the University of Lyon-I from 1992 to 2000. Thanks are due in turn to your Tax Euro, as the last common ancestor of the sources of the financial supports, with a special mention to the lineages corresponding to the French "Ministère de l'Education Nationale", and to the French "Ministère de la Recherche".
- I consider as a great honour that **Noboru Sueoka** accepted to be present for this occasion. As pointed out by Christian Gautier in a recent review[58] about compositional biases in DNA, Noboru Sueoka was the first to state in 1962, before the emergence of neutralism, that some patterns of the genome could appear without natural selection in his theory of directional mutation pressure[167]. In a sense, the present study is just an extension of his seminal works to the asymmetrical case allowed by the availability of complete genome sequence data.
- **Jean-Michel Louarn** is, for a part, the voice of the functional point of view: his works[23] showed that there is a functional polarization of the *Escherichia coli* chromosome. This phenotypic trait is selectively advantageous, and controlled not by a single locus but by many asymmetric motifs (most likely 5'-RRNAGGGS-3') spread into the genome. The open question is whether there is a link between this trait and some genome structures, such as the chirochores. I would like to thank warmly Jean-Michel Louarn for sharing his ideas and expertise about this strange and exciting situation from an evolutionary point of view.
- **Siv Andersson** was the first to note[4] that "[…] the base composition of the mitochondrial tRNAs, rRNAs, and mRNAs seems to be greatly influenced by mutation pressures which are different for the two DNA strands of these genomes", and from my point of view the complete genome of *Rickettsia prowazekii*[5] in 1998 is the real start of the genomic era because for the first time a scientific question, the origin of mitochondria, was present in the title of the publication. I would like to thank her for the time to review this work.
- **Antoine Danchin** is well known in the bioinformatic and genome sequencing world. I would like to thank him first, obviously for the time to review this work, and second for writing the book *L'œuf et la Poule*[34] which was for me like oxygen because during my undergraduate studies in biology in France I have never seen or heard the word evolution or any related notion (this is a fact, not a metaphor). This way of teaching biological core notions in France looks like trying to teach Mathematics while postponing basic linear algebra to post-graduate summer schools.
- **Christian Gautier** was at the head of the team when I joined them on a tenure position in September 1992, followed by Manolo Gouy when Christian took the head of the laboratory. I am especially grateful to Christian Gautier for always letting me work really independently in the laboratory and at the computers, I appreciated deeply the freedom of working on my own research interests, in one of the best, not to say the best, molecular evolution team available in the living world. Well, at least in the known living world, our xenocompetitors being extremely poorly documented even in the Science Citation Index®. Thanks for all.
- Last but not least, this work, and his author, was supported by, and was supported by, respectively, by all the members of the team, Christian Gautier, Manolo Gouy, Guy

Perrière, Dominique Mouchiroud, and Laurent Duret during these long years of friendly fruitful interactive scientific atmosphere. Many thanks to all, and keep up the good work !

INTRODUCTION

The theory of directional mutation pressure[167] is the main background of this work, my contribution was to build a theoretical frame allowing an easy detection of asymmetrical directional mutation pressures, that is directional mutation pressures that are different between the two DNA strands. The notion of *directional mutation* looks surprising at first glance because it may suggest an underlying finalism. I will show thereafter why this is absolutely not the case by defining what a directional mutation pressure is.

Finalism and Evolution

Finalism, *i.e.* purpose driven evolution, was discarded since Darwin as a basis of evolutionary theories, in few words evolution is a Markov process. Let E be a set of *k* elements called genetic informations and designed by their ranks.

E = { 1, 2, 3, ..., k }

A genetic population is a set F of *n* genetic informations. Let $t_1, t_2, \quad , t_m$ be an increasing date sequence and $X_{t_1}, X_{t_2}, \quad , X_{t_m}$ a random variable chain. A column probability vector gives the initial state law,

$$\mathbf{P}_{t_0} = \begin{pmatrix} P(X_{t_0} = 1) \\ P(X_{t_0} = 2) \\ \dots \\ P(X_{t_0} = k) \end{pmatrix},$$

corresponding to the initial relative frequencies of the genetic informations in the population F. Let $\mathbf{S}$ be the *k*-square matrix of transitions probabilities which entry s_{ij} is the probability to obtain j at time t_{m+1} knowing there was i at time t_m:

$$s_{ij} = P\left(X_{t_{m+1}} = j \middle| X_{t_m} = i\right)$$

The state law at time t_{m+1} is defined by P_{t_m} and the transition matrix $\mathbf{S}$.

$$P_{t_{m+1}} = \mathbf{S} P_{t_m}$$

In such a Markov process the future is influenced by the past only through the present state, there is no place for finalism. The theory of directional mutation pressure is a Markov process, but we have to split the transition from t_m to t_{m+1} into two sub-steps to define it.

Chance and necessity

The results of genetics and molecular biology led to the distinction between the processes that yield diversity, random mutations $\mathbf{C}$, and the processes that select this diversity, natural selection $\mathbf{N}$. Mutations are working at the software level by modifying genetic informations while selection is working at the hardware level, a distinction coming from the irreversibility of information flux *in vivo* (DNA → RNA → Proteins), the fundamental result of molecular biology.

$$P_{t_{m+1}} = \mathbf{NC} P_{t_m}$$

Evolution is then an alternating Markov process:

$$P_{t_0} \xrightarrow{\mathbf{C}} P_{t_{0bis}} \xrightarrow{\mathbf{N}} P_{t_1} \xrightarrow{\mathbf{C}} P_{t_{1bis}} \xrightarrow{\mathbf{N}} P_{t_2} \xrightarrow{\mathbf{C}} P_{t_{2bis}} \xrightarrow{\mathbf{N}} P_{t_3} \xrightarrow{\mathbf{C}} \dots$$

In such a process chance **C** is not adjustable to the requirements of necessity **N** because population state is known only at present time, it's impossible to take advantage of the past (what were the good genetic information from a selective point of view) to anticipate future. This is the meaning of random in *random mutations*, as it was clearly explained by Graur and Li:

> ***« Are mutations random?***
>
> Mutations are commonly said to occur « randomly ». However, as we have seen mutations do not occur at random with respect to genomic location, nor do all types of mutation occur with equal frequency. So, what aspect of mutation is random? Mutations are claimed to be random in respect to their effect on the fitness of the organism carrying them. That is, any given mutation is expected to occur with the same frequency under conditions in which this mutation confers an advantage on the organism carrying it, as under conditions in which this mutation confers no advantage or is deleterious. »
>
> Graur and Li (2000) Fundamentals of molecular evolution[65].

There is a *directional mutation pressure* when mutations probabilities are not all the same: there are at least two off-diagonal entries in matrix **C** with different values. Random mutation does not mean equiprobability; directional mutations are also random mutations.

Genetic drift and selective and mutation pressures

Neutralism does not mean absence of selection but equiprobability for the selection of genetic information.

Neutralist hypothesis:	$n_{ij} = \frac{1}{k}$

Because under this hypothesis the matrix **N** does not modify the state of the population, only the mutation matrix controls evolution,

$$P_{t_{m+1}} = \mathbf{C}P_{t_m},$$

a peculiar case especially interesting as a null hypothesis: this is the theory of directional mutation pressure as stated by Sueoka in 1962[167].

Working with *relative* frequencies of genetic information to characterise the state of the population means that an implicit hypothesis is that the size of the population is large and constant over time. But in the real world population size are finite: there are sampling fluctuations from one generation to the next one yielding to genetic drift.

> As an historical sidelight, Sueoka's theory appears to be one of the first neutral theories of DNA evolution. It explicitly assumes that natural selection plays no role in the dynamics of allele frequencies. However, as it does not incorporate genetic drift, it cannot describe the fixation of nucleotides. This aspect of the theory had to wait six more years for the publication of the papers by Kimura[99] and King and Jukes[101].
>
> John H. Gillespie. The causes of molecular Evolution[59].

Things are not so simple, the problem is that the notion of *fixation* of a genetic information in a finite population has a meaning only for « irreversible » processes, for instance when *k* is large enough so that a new genetic information is not already present in the population (infinite allele model, infinite site model, irreversible mutation model), or when the total number of mutation in the whole population is very small.

Since a genetic information has only one frequency at a given time within a population, the probability density function of the steady-state distribution, $\varphi(x)$, has no direct meaning. We

have to postulate a hypothetical aggregate of an infinite number of populations evolving under the same conditions, then $\varphi(x)dx$ gives the relative frequency of populations such that the relative frequency of a genetic information is in the range $[x, x+dx]$.

For a reversible mutation pressure with $k = 2$ and denoting u and v the specific mutation rates from and to the genetic information which relative frequency is x, Wright[189] showed that the probability density function is given by:

$$\varphi(x) = \frac{\Gamma(2n(u+v))}{\Gamma(2nu)\Gamma(2nv)} x^{2nv-1}(1-x)^{2nu-1}$$

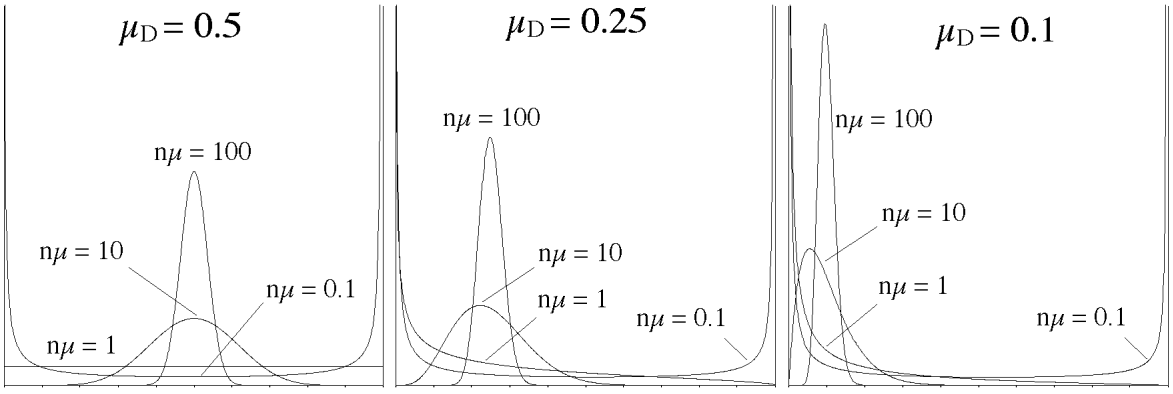
with mean:

$$\bar{x} = \int_0^1 x\varphi(x)dx = \frac{v}{u+v} = \mu_D$$

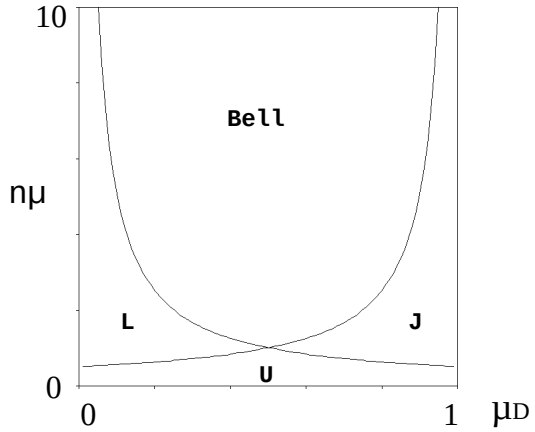
and variance:

$$\sigma_x^2 = \int_0^1 (x-\bar{x})^2 \varphi(x)dx = \frac{\mu_D(1-\mu_D)}{2n(u+v)+1}$$

The mean and the variance are both modulated by the mutation pressure but the finite population effects are visible only at the variance level. Some examples of the probability function $\varphi(x)$ are depicted thereafter.



Depending on the value of the product $n\mu$ and on the value of μ_D, $\varphi(x)$ is a bell-, U-, L-, or J-shaped distribution.



The critical value $n\mu = 1$ means that there is exactly one mutation on average within the whole population per generation, and the critical value $\mu_D = 0.5$ that the specific mutation rates *u* and *v* are equal. There are three main possibilities:

- When $n\mu >> 1$ and $\mu_D \approx 0.5$, there are many mutations within the population and the directional mutation pressure is low, there is a permanent polymorphism within the population. The mode and the mean of the distribution are very close, the most likely is to have a heterogeneous population with *x* close to μ_D. Since a homogeneous population is very unlikely, the fixation of a genetic information is not meaningful.
- When $n\mu << 1$, mutations are very rare, usual genetic drift effects are observed. The most likely is a homogeneous population and heterogeneous transients allows to alternate the two homogeneous states whose probabilities are close to μ_D and $1 - \mu_D$. The fixation of a genetic information is meaningful in this case.
- When $\mu_D \to 0.0$ or $\mu_D \to 1.0$, when there is a strong directional mutation pressure, or when $n\mu \approx 1$, when there is on average close to one mutation per generation in the population, we have something intermediate between the previous cases. The mode and the mean of the distribution are very different, the most likely is to have an homogeneous population composed only of the genetic information favoured by the directional mutation pressure, but heterogeneous populations with few unfavoured genetic information are also common. The notion of fixation is not very meaningful here because it's always the same genetic information that can be fixed; this is somewhat similar to the irreversible mutation pressure scheme[100].

Genetic drift effects are visible only at the level of the variance of the distribution, not at the level of its means. This is a justification of the interest of the previous Markov model for the evolution of the mean of relative frequencies of genetic informations. We have just to keep in mind that the variance can be very high for small populations.

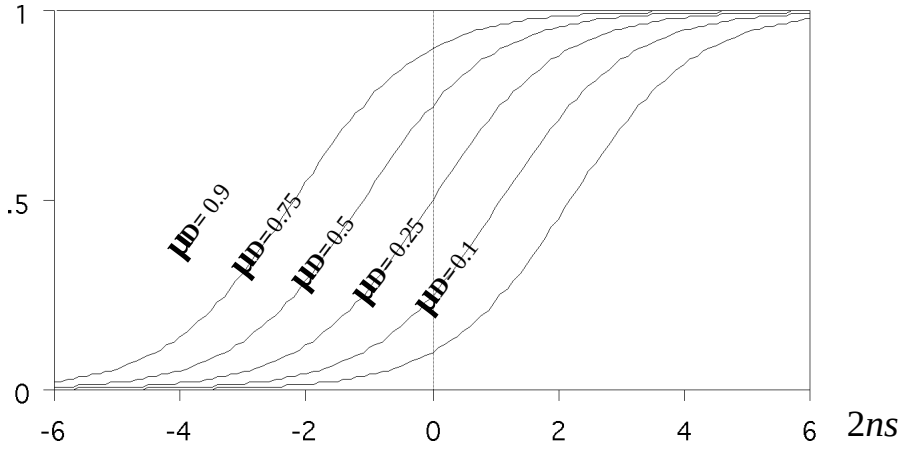
As soon as the neutralist hypothesis is relaxed, models are much more complex, because genetic drift will also influence the mean of the probability density function. Always with *k* = 2 and with a selective advantage of 0 and -*s* for the genetic information whose relative frequency are *x* and 1-*x*, respectively, Wen-Hsiung Li showed[106] in 1987 that the mean of the distribution is approximately:

$$\bar{x} \approx \frac{e^{2ns} v}{e^{2ns} v + u} = \mu_D \frac{e^{2ns}}{\mu_D e^{2ns} + 1 - \mu_D}$$

This is a sigmoidal response curve starting from μ_D at the origin, when there is no selection, and tending to 1 for high values of the product *ns*, when selection is efficient. The critical point between these two states is given by the x-coordinate, *s**, of the inflection point of the curve,

$$s^* = \frac{1}{2n} \ln\left(\frac{1 - \mu_D}{\mu_D}\right),$$

which is highly dependent on μ_D value as depicted below.



When $\mu_D = 0.5$ the inflection point is at zero and we have the usual condition $2ns >> 1$ for selection to be efficient. When $\mu_D > 0.5$ the inflection point is at a negative value, selection and the mutation are working in the same direction. When $\mu_D < 0.5$ the inflection point is at positive value, selection and mutation are working in opposite direction so that the criterion $2ns >> 1$ is not enough for selection to be efficient.

> A mutation generally occurs in a single individual and give rise to an allele. If an allele achieves some frequency in a population it can be referred to as a polymorphism (not a « common [or rare] mutation ») If it has become fixed in a population it may be referred to as a substitution.
> *Molecular Biology and Evolution*, Instructions to Authors[7].

In the following the substitution specific rate r_{ij} is the probability of transition from *i* to *j* per time unit,

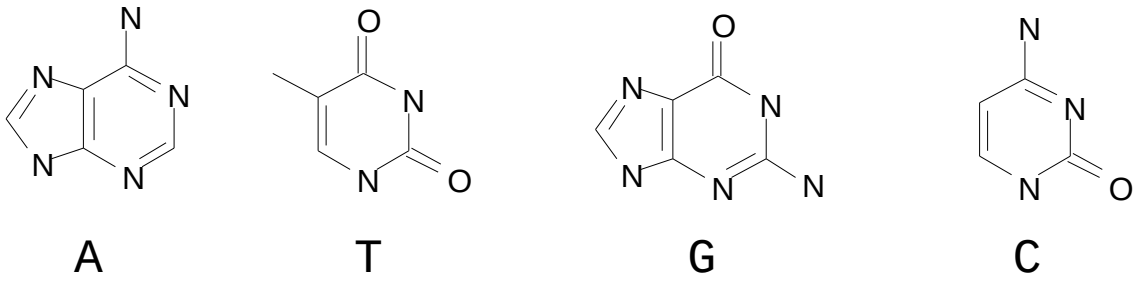
$$r_{ij} = \frac{s_{ij}}{\Delta t} = \frac{P\left(X_{t_{m+1}} = j \middle| X_{t_m} = i\right)}{t_{m+1} - t_m},$$

that is the instantaneous net result of mutation and selection. Note that the meaning is more general than for the usual allelomorphic gene substitution rate to handle the case of permanent polymorphism when there is no fixation of genetic information in the population.

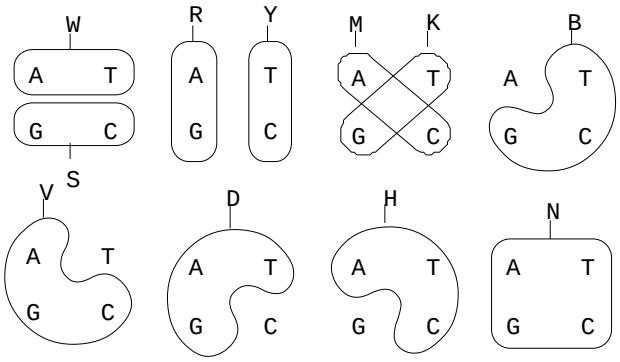
Genetic information hardware

Population genetics is completely hardware independent: its results would be valid for hypothetical populations do not working with nucleic acids as a material basis. Why should we care about the underlying hardware? It depends on the level of analysis: within the frame of the Delphic boat metaphore[37,38], a boat is better characterised by the relationships between its components than the sole list of its component properties, but when you are working at the plank level, substituting a plank in wood by a plank in sand would have dramatic effects for the global properties of the boat. In a similar way, when evolution is studied at the molecular level, it is difficult to be completely independent of the physico-chemical properties of nucleic acids.

The material basis of genetic informations is an heteropolymer of deoxyribonucleic acids (DNA) whose monomers are characterised by their nucleic basis component: adenine (A), thymine (T), guanine (G), or cytosine (C).



The following standard abbreviations[6] are used thereafter.



In double stranded DNA pairing bases are always W or S[187], they are said to be complementary bases. Let $\bar{N}$ be the complementary basis of N, we have then:

$\bar{A} = T,\ \bar{T} = A,\ \bar{G} = C,\ \bar{C} = G$

Shortly, regardless of basis N, we always have:

$\bar{\bar{N}} = N$

This fundamental property of genetic information hardware is used in the following to build the symmetric evolution model.

THE MODEL OF SYMETRIC MOLECULAR EVOLUTION

Biological hypotheses

The starting hypothesis of the model of symmetric molecular evolution is that mutation and selection are the same for the two strands of DNA. This hypothesis was called parity rule number 1 by Sueoka[172], PR1 in short. Let's consider the consequence for the structure of the substitution matrix.

Let

$r(X \rightarrow Y)$

be the substitution specific rate from basis X to Y on one strand, and

$\bar{r}(\bar{X} \rightarrow \bar{Y})$

the substitution specific rate for the complementary event on the other strand. Since these two substitution scheme yield the same result, the apparent substitution specific rate on one strand, $R(X \rightarrow Y)$, is equal to the sum of these two substitution specific rates:

$$R(X \rightarrow Y) = r(X \rightarrow Y) + \bar{r}(\bar{X} \rightarrow \bar{Y})$$

For the complementary substitution we have in the same way:

$$R(\bar{X} \rightarrow \bar{Y}) = r(\bar{X} \rightarrow \bar{Y}) + \bar{r}(\bar{\bar{X}} \rightarrow \bar{\bar{Y}})$$

Since

$\bar{\bar{N}} = N$

this can be rewritten as

$$R(\bar{X} \rightarrow \bar{Y}) = r(\bar{X} \rightarrow \bar{Y}) + \bar{r}(X \rightarrow Y)$$

PR1 hypothesis is that the substitution specific rates are the same for the two DNA strands:

PR1 hypothesis: $X,Y \quad N: \; r(X \rightarrow Y) = \bar{r}(X \rightarrow Y)$

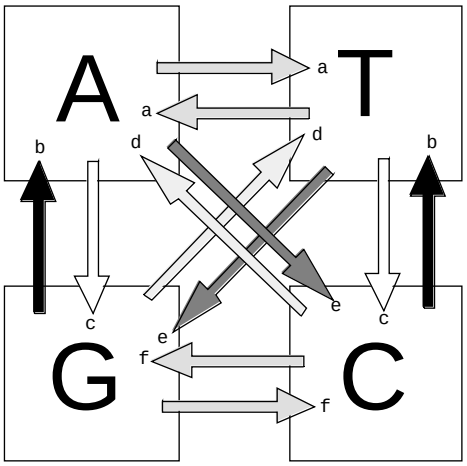
As a consequence for the apparent substitution specific rate we have under PR1:

$$R(X \rightarrow Y) = R(\bar{X} \rightarrow \bar{Y})$$

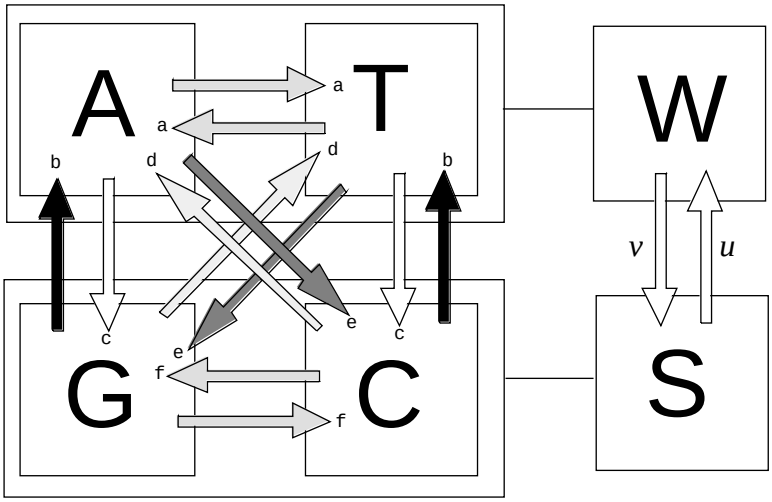
Hence, under PR1 hypothesis the apparent substitution specific rate from one basis to another one is equal to the substitution specific rate of the complementary event, for instance:

$$R(A \rightarrow G) = R(T \rightarrow C)$$

The total number of substitution specific rate, 12 in the general model, is divided by two under PR1 hypothesis as depicted below:



This model can be understand as a simplification of the general 12-parameter model or as an extension of Sueoka's 2-parameter model[167],



the connection between the two models is given by $u = b + d$ and $v = e + c$, the two parameters a and f do not appear because of the merging of W and S bases in the two-parameter model. Transitions are substitutions intra-R or intra-Y and correspond to parameters b and c, isotypic transversions are intra-W or intra-S (a and f), allotypic transversions are intra-M or intra-K (d and e).

Model notations

Let **X** be the column matrix,

$$\mathbf{X} = \begin{pmatrix} A(t) \\ T(t) \\ G(t) \\ C(t) \end{pmatrix}$$

whose elements are the nucleotide relative frequencies in one DNA strand at time *t*. Let **R** be the matrix for the continuous process of evolution of base frequencies.

$$\frac{d\mathbf{X}}{dt} = \mathbf{R}\mathbf{X}$$

The entries in the matrix **R** are the substitution rates. Many parametric forms of matrix **R** have been published[107,109,152,192], under PR1 the matrix is:

$$\mathbf{R} = \begin{pmatrix} -a-e-c & a & b & d \\ a & -a-e-c & d & b \\ c & e & -b-d-f & f \\ e & c & f & -b-d-f \end{pmatrix}$$

where the six parameters (a,b,c,d,e,f) represent the six substitution specific rates as depicted in previous figure. Parameter notations are those from Sueoka[172] and Lobry[110] in 1995. This model was also derived and studied independently by Valenzuela[183] in 1997, and also used by Wu and Maeda[191] in 1987 but without biological justification.

Equilibrium base frequencies (PR2)

The equilibrium point $\mathbf{X}^*$ is given by:

$$\mathbf{X}^* = \frac{1}{2}\begin{pmatrix} 1-\theta^* \\ 1-\theta^* \\ \theta^* \\ \theta^* \end{pmatrix}$$

where θ^* is S-base frequency at equilibrium, which is function only[110] of 4 out of the 6 substitution specific rates:

$$\theta^* = \frac{e+c}{b+c+d+e}$$

This result is consistent with Sueoka's two-parameter model whose S-base equilibrium frequency is given[54,167] by:

$$\theta^* = \frac{v}{u+v}$$

This equilibrium point is such that $A(t) = T(t)$ and $G(t) = C(t)$, a state called parity rule number 2 by Sueoka[172], PR2 in short.

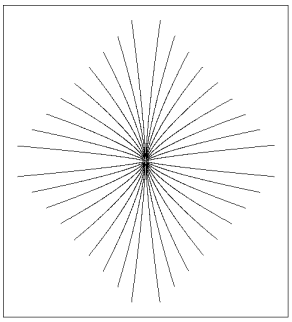
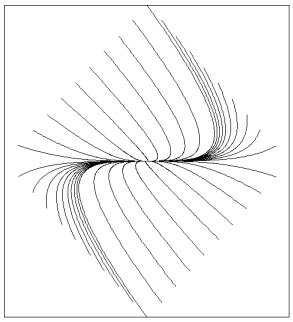
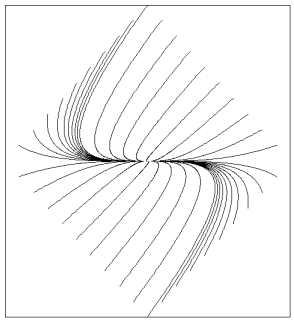
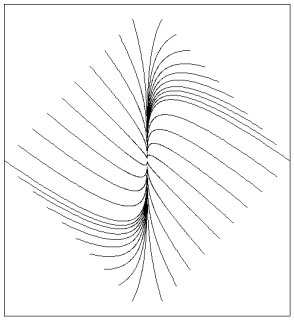
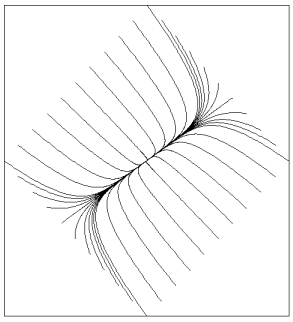
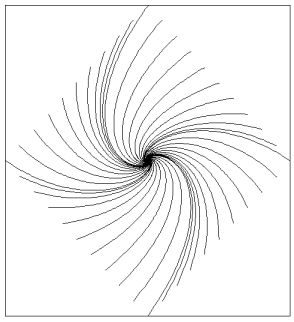
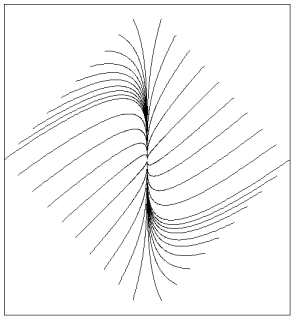
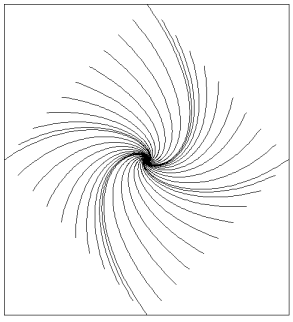
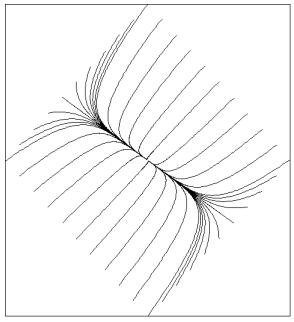
PR2 state : $A(t) = T(t)$ and $G(t) = C(t)$

This fundamental property of the model was checked independently by Sueoka[172] with numerical simulations and analytically by Valenzuela[183].

Convergence to equilibrium base frequencies

Under the hypothesis that all substitution specific rates are strictly positive, **R** belongs to the class of compartmental matrices, which are known to have no eigenvalue with positive

real part and no purely imaginary eigenvalue[78]. Moreover, as **R** corresponds to a closed system with no internal traps the multiplicity of the first eigenvalue is one by Foster-Jacquez theorem[43]. Then, There is only one equilibrium point and this equilibrium point is stable: regardless of initial conditions and substitution specific rate values trajectories will tend exponentially to frequencies at equilibrium[110]. This behaviour is depicted in the plots below where the x-axis is $A(t) - T(t)$, the y-axis $C(t) - G(t)$, PR2 state is at the origin.

	d = b	d > b	d < b
e = c			
e > c			
e < c			

Parameter values control the way to converge to the origin, there are different possible approaches, but in all the cases there is convergence. If for a given DNA sequence we knew that equilibrium is reached then there would be a simple way to reject the model just from its base frequencies. However, DNA sequences are observed only at present time so that any deviation from PR2 is also interpretable as non-equilibrium transient state under PR1 hypothesis.

Convergence to PR2

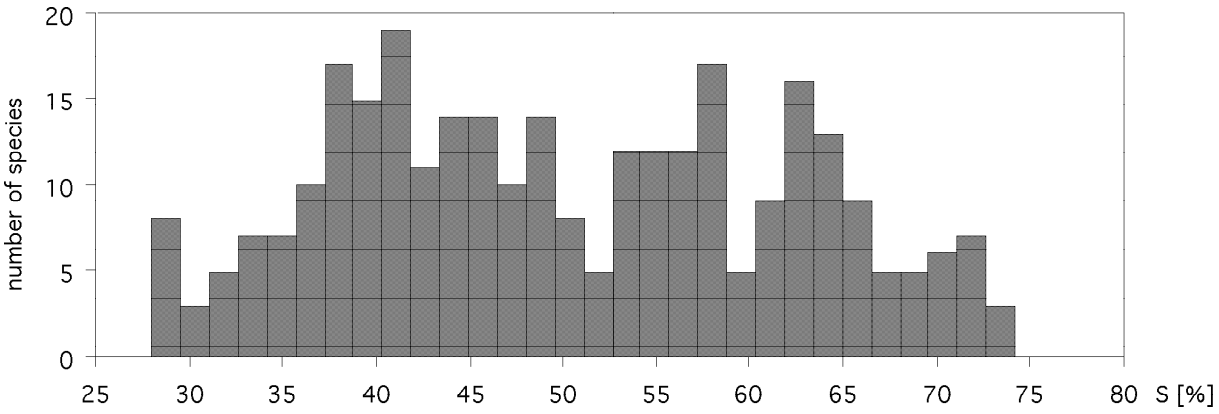
We have shown[119] that there is converge to PR2 even in non-equilibrium case under the weak requirement that all substitution specific rates are greater than a given positive threshold. This result is obtained with a more complex model whose parameter are allowed change with time,

$$\frac{d\mathbf{X}}{dt} = \mathbf{R}(t)\mathbf{X},$$

where matrix **R**(t) has the structure coming from PR1 hypothesis,

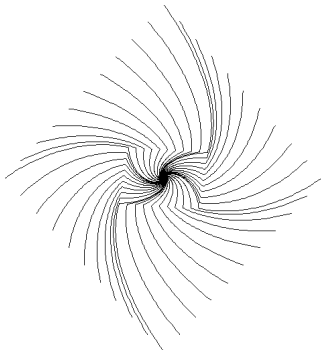
$$\mathbf{R}(t) = \begin{pmatrix} -a(t)-e(t)-c(t) & a(t) & b(t) & d(t) \\ a(t) & -a(t)-e(t)-c(t) & d(t) & b(t) \\ c(t) & e(t) & -b(t)-d(t)-f(t) & f(t) \\ e(t) & c(t) & f(t) & -b(t)-d(t)-f(t) \end{pmatrix},$$

since we are still dealing with an evolution symmetric with respect to the two DNA strands. From a biological point of view this model is more satisfactory because for long evolutionary periods it is not sensible to postulate that the substitution specific rates are constant, as is obvious from high variability of S-base frequencies in bacterial genome[166]. The figure thereafter is the distribution of S-base frequencies for 298 bacterial genomes with more than 50kb available in databases.



The black hole of symmetric molecular evolution

The convergence to PR2 is illustrated in the following simulation. The substitution specific rates values have been changed abruptly during the course of evolution, at the transition time the system is far from equilibrium for the S-base content, but it still converges to PR2, even if it is in a different way.

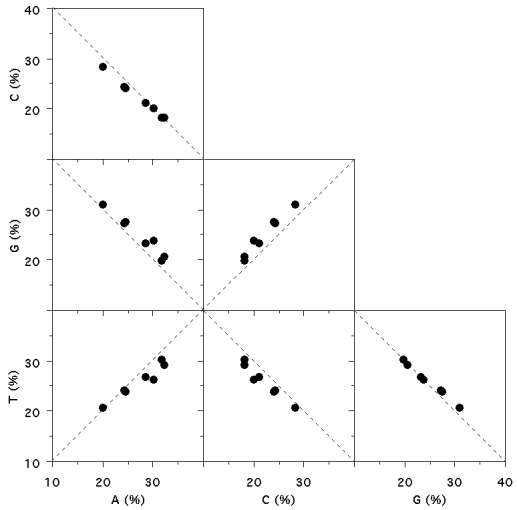


Thanks to this result we are in a much more comfortable situation to reject the model from DNA base frequencies because we don't have to work under the equilibrium hypothesis. A deviation from PR2 means that PR1 hypothesis was violated during the course of evolution of the DNA sequence under study.

PR2 as an approximation for complete genomes

> « Not unrelated to this as yet unexplained finding may be later observations from my laboratory, namely, that in microbial DNA the separated heavy and light strands, although complementary to each other with respect to base composition, both exhibit the same equivalence of 6-amino and 6-oxo bases. To my knowledge, there have been no follow-up studies of the last-mentioned observations in other laboratories. »
>
> Erwin Chargaff (1979) How genetics got a chemical education[28].

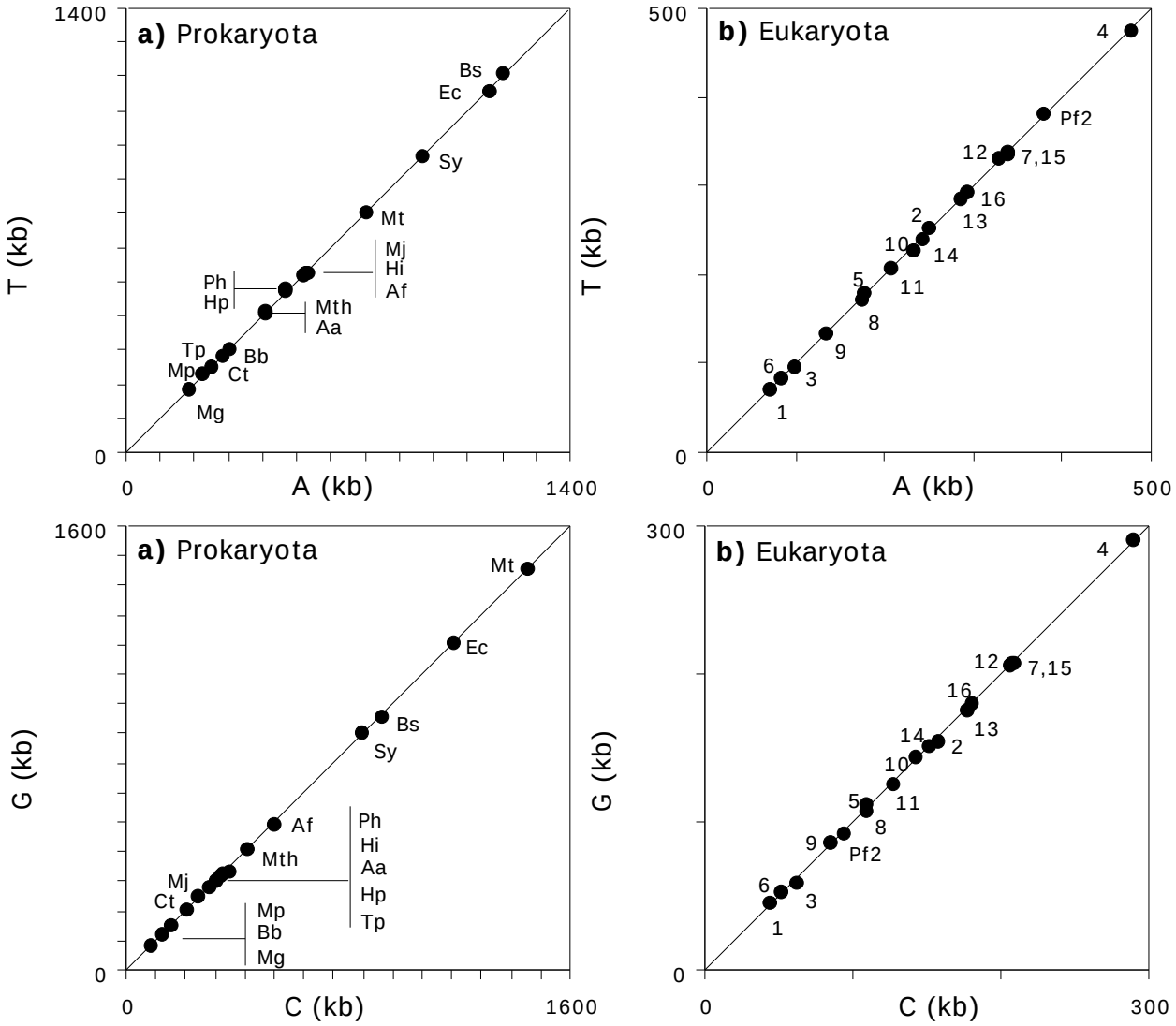
The direct experimental determination of the global base composition of a complete genome is difficult because the two strands have to be analysed separately, otherwise PR2 is obtained as a direct consequence of base pairing rules in double stranded DNA. The chemical composition of single stranded DNA was reported[155,93] in 1968 from Chargaff's lab for *Bacillus subtilis* and extended later for six more bacterial species[156]: *Proteus vulgaris*, *Bacillus megaterium*, *Bacillus stearothermophilus*, *Escherichia coli*, *Salmonella enterica* serovar Typhimurium and *Serratia marcescens*. The figure thereafter is a plot of these results for the L-strand; the dashed line is what is expected under PR2.



These results were puzzling: PR2 state is a clear consequence of the structure of double stranded DNA, but why should PR2 hold for single stranded DNA too? As an anecdotal sidelight, note that if PR2 holds for single stranded DNA then the fact that PR2 holds for double stranded DNA is no more an argument in favour of the double helix structure for DNA[187]. These results were more or less forgotten during a quarter century, with few exceptions such as studies of oligonucleotide frequencies within each strand[157], before the availability of long genomic fragments allowed for a new look at this question with a better accuracy for base frequencies values.

Nussinov pointed out in 1982 that for three complete eukaryotic viruses PR2 holds within each strand[133], but these genomes are very small, about 5 kb. A more systematic study[40] with all sequences from *Homo sapiens* and *Escherichia coli* available in 1992 showed that PR2 is usually observed for all sub-sequences ranging from 0.05 kb to 1 kb, but the problem is that this analysis merged sequences from the two strands, cancelling out a potential deviation from PR2. Prabhu's study[145] with 32 genomic fragments sizing more than 50 kb showed that PR2 holds for single stranded DNA, this result was confirmed[110] when 60 fragments with more than 50 kb become available in June 1994. These genomic fragments were from various taxonomic sources (viruses, prokaryotes, nematode, chloroplasts, insects, vertebrate, mitochondria, yeast) suggesting that the rule was general.

Mycoplasma genitalium sequence[51] is the starting point in 1995 of the complete genome area (excluding organelles and viruses). The analysis[119] of the complete genome of 4 achaebacteria, 12 eubacteria, 16 *Saccharomyces cerevisiae* chromosomes, and *Plasmodium falciparum* chromosome 2 showed that PR2 is a good approximation for single stranded DNA, as depicted in the following figure.



A selective interpretation of PR2 state in complete genome was put forward by Forsdyke[42]: this state would be the result of a selection pressure favouring mutations that generate complementary oligonucleotides in close proximity, thus creating a potential to form stem-loops. this interpretation is not very convincing because: i) based only on the *in silico* predicted[194] likelihood for DNA to adopt a cruciform structure when their *in vivo* existence is unsure[121,163]. ii) Cruciform stability decrease with temperature *in vitro*, if they were selected *in vivo* one would expect stem S-base frequencies to increase with temperature[184], as it is indeed observed for stem S-base frequencies in tRNA and rRNA, but this is not the case[57]. iii) The proportion of bases involved in an intra-strand base pairing would be[11] $\left(W-|A-T|+S-|C-G|\right)/N$. For instance in *Borrelia burgdorferi*[49] (A = 323079, T = 327196, C = 130760, G = 129646) 99.4% of bases would be involved in such structures which is hardly compatible the high proportion (93.6%) of bases involved in coding sequences.

Anyway, these results for complete genomes are not very interesting because they do not yield a rejection of the model of symmetric molecular evolution: as any null hypothesis models are informative only when rejected.

THE CHIROCHORE STRUCTURE OF BACTERIAL GENOMES

Model rejection interpretation

The entries of matrix **R**, the substitution specific rates, represent the net instantaneous result of mutation and selection. Rejection of the model does not identify the cause of the asymmetry between the two strands, and extra biological information is need before this interesting point can be discussed.

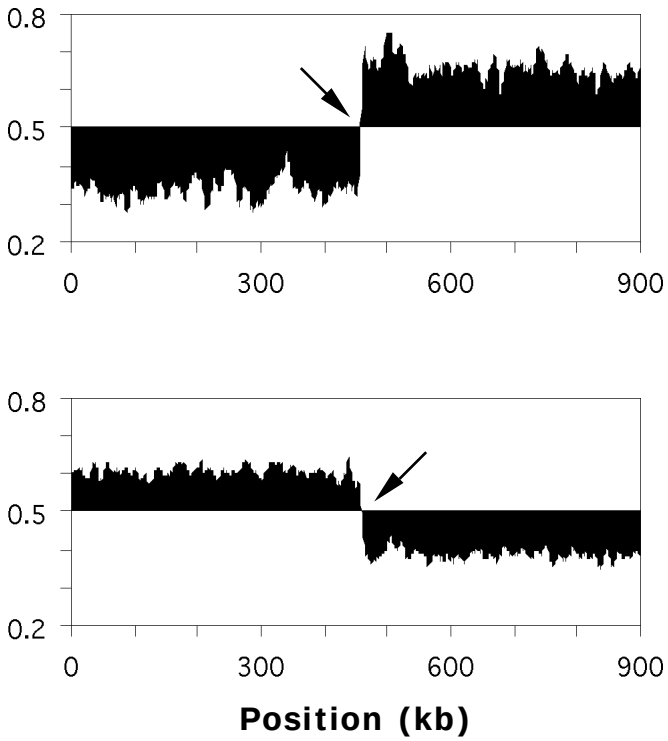
I found the chirochore structure of bacterial genomes by chance while I was challenging PR1 model prediction, *i.e.* PR2 state, along the first complete bacterial genomes[112]. In bacteria there are often segments homogeneous for the deviations from PR2 that I called chirochores by analogy with isochore that are segments homogeneous for S-base frequencies. Chirochores are a purely descriptive notion without reference to any mechanism. Replichores[18] are segments between an origin and a terminus for replication. The nice thing is that chirochores and replichores boundaries are the same[111,112,113,53,131,66,68,95,128,97,127,158,104,149,150,120,26,27,122,123,124].

A chirochore structure was also reported for the complete genomes from *Escherichia coli*[18], *Bacillus subtilis*[102], *Borrelia burgdorferi*[49], *Rickettsia prowazekii*[5], *Campylobacter jejuni*[138], *Treponema pallidum*[50], *Nesseria meningitidis*[179,137], *Chlamydia trachomatis*[147], *Chlamydia pneumoniae*[147].

The chirochore structure of bacterial chromosomes is interpreted as the result of complexes superposition of asymmetric selective and mutation pressures[48], a two way variance analysis (sense versus anti-sense strand, leading versus lagging strand) showed[182] that a significant proportion of base composition biases is due to the orientation with respect to replication: gene composition is different between the leading and the lagging strand.

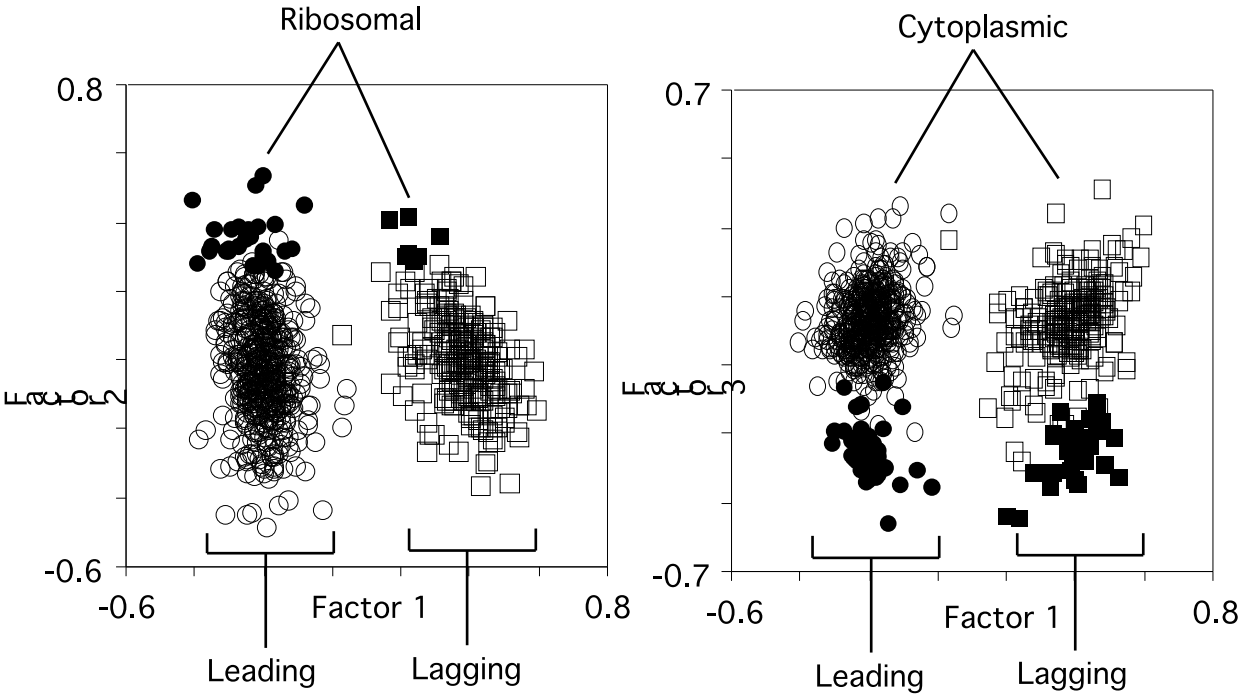
Asymmetric mutation pressure example

Up to now the most impressive chirochore structure is found in found in *Borrelia burgdorferi* major chromosome, depicted below with a 10 kb moving window and a 1 kb incremental step taking into account only third codon position bases in the published[49] strand,



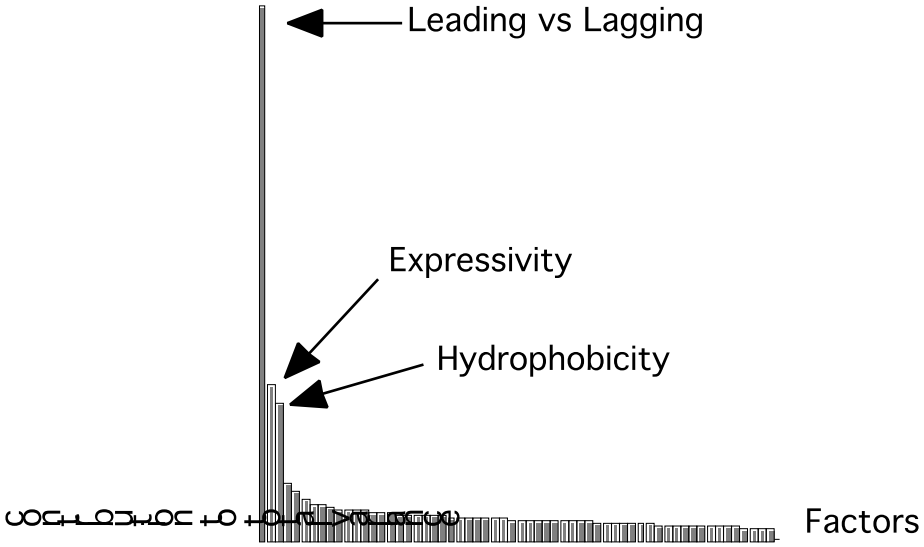
the arrows are pointing towards the experimentally mapped[144] origin of replication of the chromosome.

The chirochore structure has a strong influence on codon usage in *Borrelia burgdorferi* as shown by the two first factorial maps of correspondence analysis of codon frequencies in the 772 coding sequences with more than 300 b in this chromosome:



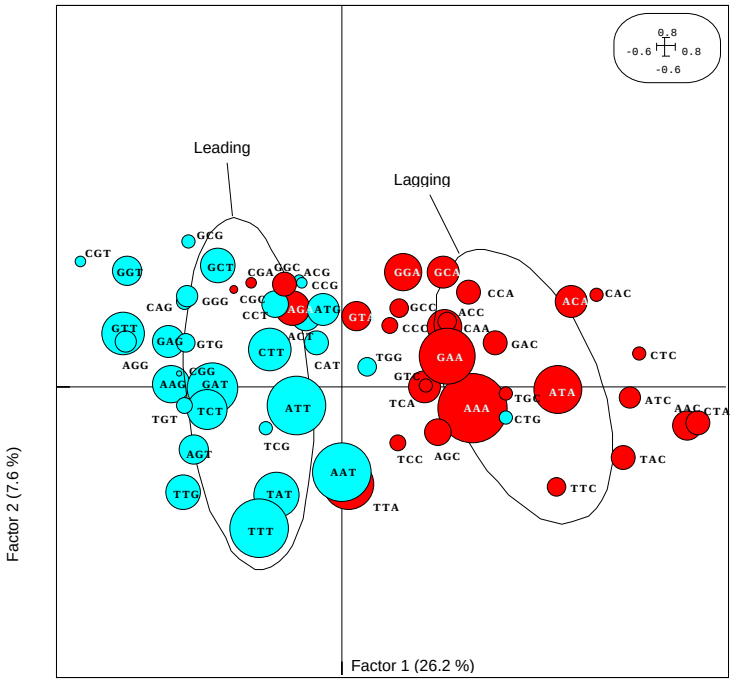
The first factor (26.2% of total inertia) is the opposition between coding sequences that are on the leading strand for replication versus those on the lagging strand. The second factor (7.6%)

is the usual[60,77] gene expressivity level effect and the third factor (6.7%) is the usual[118] opposition between sequences coding for integral membrane proteins versus those coding for cytoplasmic proteins. Note that these factorial maps are less fuzzy than those published elsewhere[127,104] because the table under analysis, as in regular correspondence analysis, contains codon absolute frequencies, and not RSCU[160] values. The eigenvalue graph thereafter is a simple visualisation of the relative contribution of interpretable and residual factors.

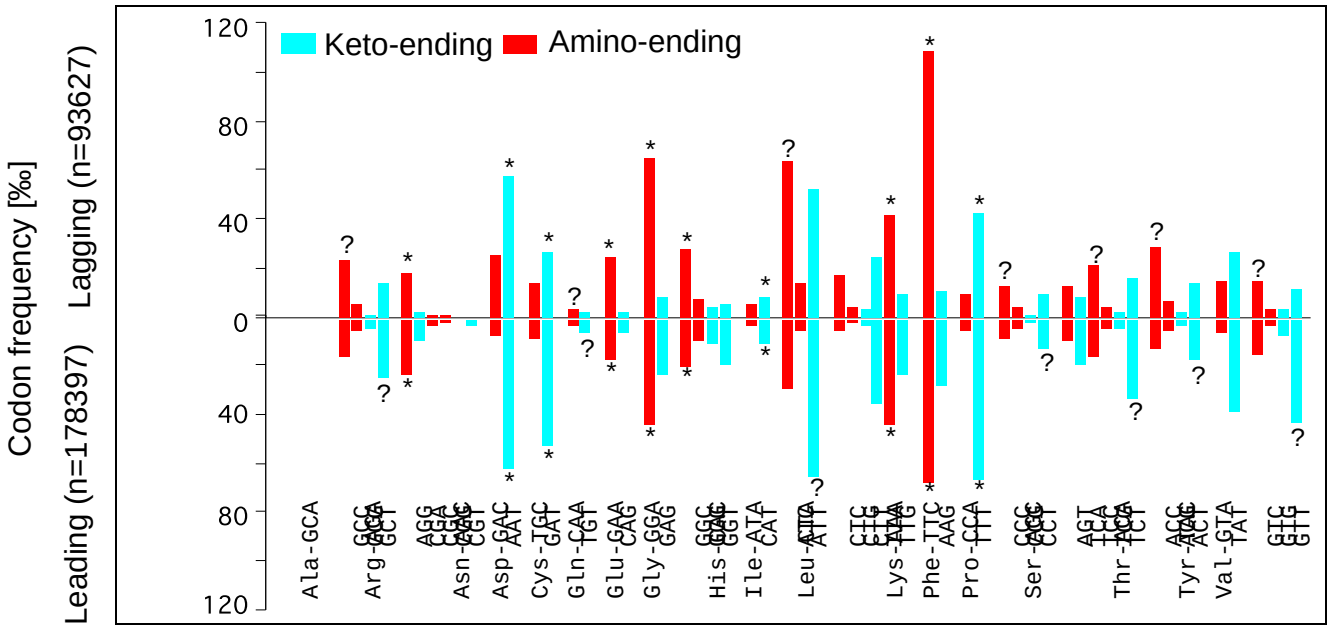


Borrelia burgdorferi genome is very special because there are two subsets of coding sequences with a completely different codon usage, a kind of molecular schizophrenia for the dialect in use within a single genome. *Borrelia burgdorferi* is in an evolutionary dead-end street for translation optimisation, the only thing to do would be to move all genes on one strand, as in some mitochondria, to stop dealing with two different codon usages.

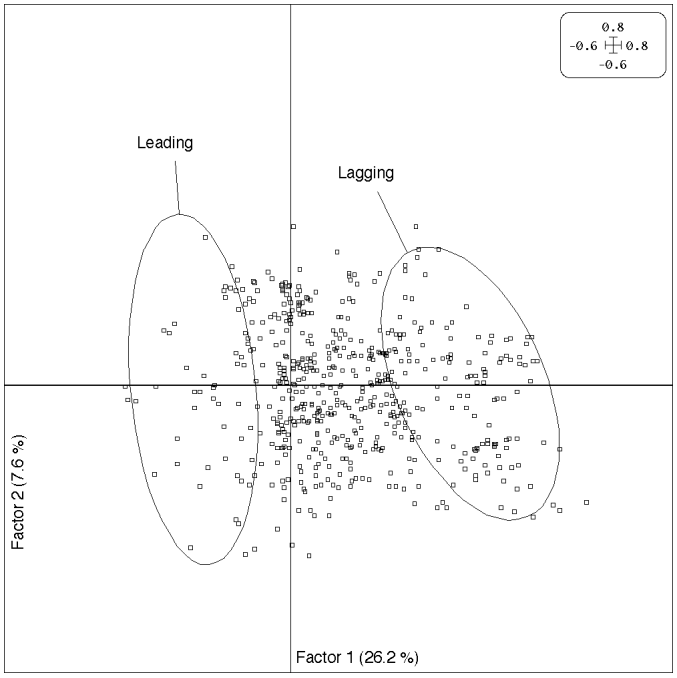
The first factorial map in codon space thereafter shows that the two subsets of coding sequences are characterised by their base composition in third codon position, with sequences from the leading group enriched in K-bases (light grey) and those from the lagging group enriched in M-bases (dark grey).



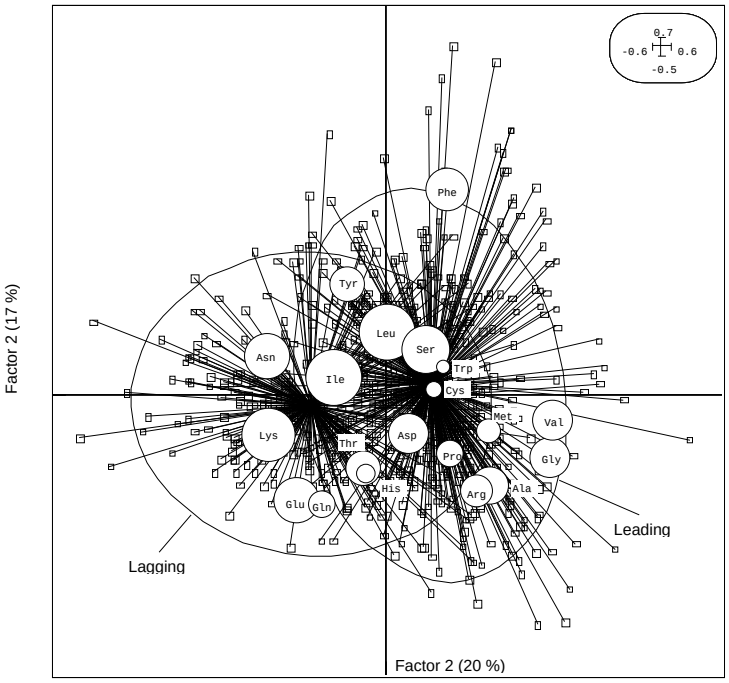
Since third codon positions are under weak selective constraints the most likely explanation is an asymmetric directional mutation pressure within this genome. In unicellular organisms, it is well known that the most important factor of codon usage variability is linked to gene expressivity[77,60,73,161,159,3,86]: frequent codons correspond to tRNA with a high intracellular concentration and this trend is exacerbated for highly expressed genes. This selective pressure is important enough to affect amino-acid composition of proteins in *Escherichia coli*[162,118]. For *Borrelia burgdorferi*, the asymmetric directional mutation pressure and the translation-linked selective pressure are not working in the same direction. For instance, among AAR codons for Lys, AAA is the most frequent for both the leading and the lagging coding sequences. Then for translation optimisation the best location is on the lagging strand to take advantage of the mutation pressure that increases A frequency on this strand. However, among AAY codons for Asn, AAT is the most frequent codon for both the leading and lagging group, so that the best location is on the leading strand to take advantage of the mutation pressure that increases T frequency on this strand. It is not possible to follow the same reasoning for all amino acids because the mutation pressure is so high that the major codon is not always the same for the two groups so that we cannot infer the optimal codon (question mark in the following graph). Optimal codons that can be determined (stars) are favoured by the mutation pressure either on the leading or the lagging group. Since a coding sequence can hardly be split between the two strands, there is no way to take advantage of the asymmetric directional mutation pressure for translation optimisation.



The 12 linear and 9 circular plasmids in *Borrelia burgdorferi* contains more than 40% of the coding potential of the cell, and it was suggested that these plasmids are in fact minichromosomes[10]. The projection of the plasmidic coding sequences onto the first factorial map below shows that base composition biases are weaker in these coding sequences, which could be a consequence of the high genomic flux, including chromosomal inversions, within these plamids[24].



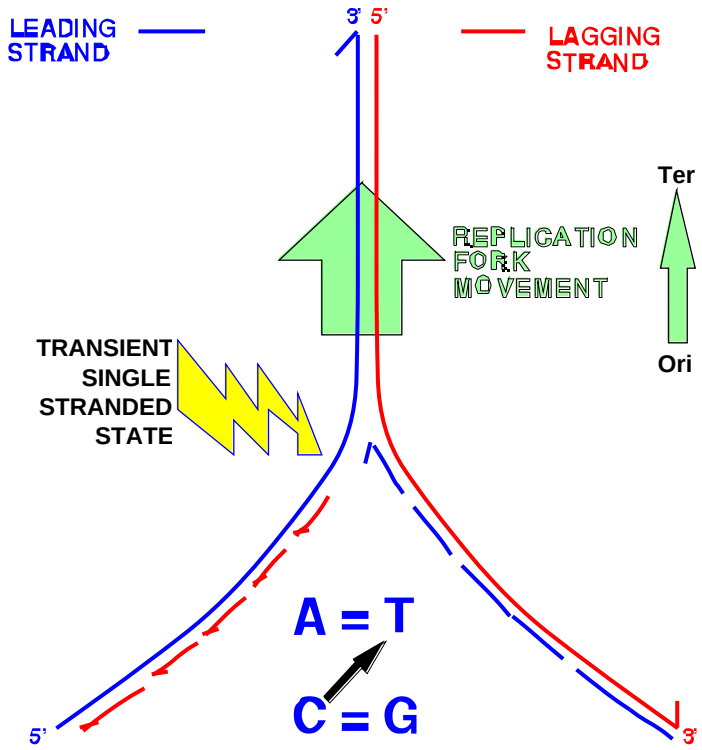
Asymmetric directional mutation pressure in strong enough in *Borrelia burgdorferi* to influence the amino acid composition of proteins[150,104,122]. Correspondence analysis of protein amino acid composition shows that first factor is the orientation with respect to replication, which is unusual since the regular first factor of variability at the amino acid level is the opposition between integral membrane proteins and cytoplasmic proteins[118].



I have focused here on *Borrelia burgdorferi* because its genome is the most spectacular to illustrate asymmetric directional mutation pressure effects. However, it should be pointed out that this phenomenon is more general: base composition biases are universal in bacteria. Universal means that when base compositions biases are visible they are always oriented in a same direction with the leading strand enriched in K-bases[149], this does not mean that biases are always present[95]. In bacteria, correspondence discriminant analysis showed[143] that the universal bias was present in *Escherichia coli*, *Haemophilus influenzae*, *Bacillus subtilis* and *Mycoplasma genitalium*, and this was extended[150] to *Borrelia burgdorferi*, *Chlamydia trachomatis*, *Helicobacter pylori*, *Methanobacterium thermoautotrophicum*, *Mycobacterium tuberculosis*, and *Treponema pallidum*. Out of 22 complete bacterial genomes, the universal bias is visible in 16 genomes[117].

It is interesting to note that the universal bias was also detected outside the bacterial world in *Euglena gracilis* chloroplast genome[130], in viruses[41,66,67,131], and mitochondria[4,8,178,79,140,148,146]. In *Homo sapiens* the controversy[191,22,190] about a possible asymmetrical directional mutation pressure in the ß-globin region is now over[47]: nothing significant is visible.

Introduced first for mitochondrial genomes[21], the cytosine deamination theory is based on the experimental evidence that the rate of this reaction is 140 times faster in single stranded DNA than in double stranded DNA[52]. During replication the template lagging strand is protected by the newly synthesized leading strand while the template leading strand has to afford a transient single strand state waiting for the newly synthesized lagging strand to be long enough to recover a double stranded state[9,125,132].



This fundamental asymmetry of replication could explain why the biases are universal. The stronger arguments are found in mitochondria[148] and viruses[66,67] genomes whose bias intensities are positively correlated with the time the single stranded state lasts during replication. Protection against cytosine deamination could be different between genomes and explain the between species variability of bias intensities. The shorter size of Okazaki's fragments in vertebrate (0.1-0.2 kb) than in bacteria (1-2 kb) could explain why no biases are visible in vertebrates[47]. This could also explain why biases are usually weaker for W-base than for S-base: Let's start from a sequence in PR2 state with Θ_0 as initial S-base frequency and suppose as *a first approximation* that the effect of cytosine deamination is to transform a fraction α of C bases into T. For the excess of G bases we have then an expression,

$$\frac{G}{G+C}(\alpha) = \frac{\frac{1}{2}\Theta_0}{\frac{1}{2}\Theta_0 + \frac{1}{2}\Theta_0 - \frac{1}{2}\alpha\Theta_0} = \frac{1}{2-\alpha}$$

which is independent of the initial S-base frequency and whose maximum value is 1 for α = 1. On the other hand for the T-base excess we have an expression,

$$\frac{T}{A+T}(\alpha) = \frac{\frac{1}{2}(1-\Theta_0) + \frac{1}{2}\alpha\Theta_0}{\frac{1}{2}(1-\Theta_0) + \frac{1}{2}(1-\Theta_0) + \frac{1}{2}\alpha\Theta_0} = \frac{1-\Theta_0+\alpha\Theta_0}{2-2\Theta_0+\alpha\Theta_0}$$

which is dependant of the initial S-base frequency and whose maximum value is $1/(2-\Theta_0)$ for α = 1. It's only in the very peculiar case of initial total absence of W base ($\Theta_0 = 1$) that we can reach 1 as maximum value as for the G-base excess.

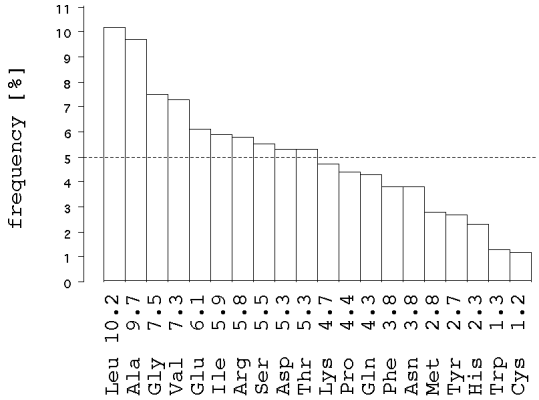
The expectation is therefore a G-base excess higher than the T one. For instance starting from a sequence in PR2 state with $\Theta_0 = 0.5$ and changing all C into T (α = 1) the G excess, G/(G+C) = 1, is higher than the T excess, T/(A+T) = 2/3. The theory of cytosine deamination is therefore compatible, at least from a qualitative point of view, with the universality of

biases, but this does not mean that this is the unique underlying source of asymmetric directional mutation pressure. Note that modelling the effect of cytosine deamination as a simple static transformation of a fraction α of C into T is very crude, but this suggests that interesting falsifiable and general predictions would result from an analytical study of asymmetrical models.

Asymmetric selective pressure example

The average amino-acid composition of proteins is affected by symmetric directional mutation pressure in bacteria[165,114,69,188,35,29], viruses[16,96,12,20], mitochondria[82,80] and eukaryotes[16,169,70,36,170,30,35,33] to a point that it's impossible to study protein thermophillic adaptation without taking this effect into account[71,126]. However, observed variations are less than would be expected if amino-acid frequencies were free of selective constraints[114]. There is most likely a stabilising selective pressure toward optimal amino-acid compositions to avoid the global physico-chemical of proteins, such as their solubility, being aberant[83].

Optimal amino-acid frequencies are unknown; we can take those from *Escherichia coli*[118] as guideline because its genomic S-base frequency (50.8%) is in the middle of the observed range in bacteria[166,174].



To compute the expected ratio T_1/A_1 we just have to compute the ratio of codon frequencies TNN/ANN compatible with the amino-acid frequencies as in the table below:

TNN	aa	aa%	min	max	uni	ANN	aa	aa%	min	max	uni
TTY	Phe	3.8	3.8	3.8	3.8	ATH	Ile	5.9	5.9	5.9	5.9
TTR	Leu2	10.2	0.0	10.2	3.4	ATG	Met	2.8	2.8	2.8	2.8
TCN	Ser4	5.5	0.0	5.5	3.7	ACN	Thr	5.3	5.3	5.3	5.3
TAY	Tyr	2.7	2.7	2.7	2.7	AAY	Asn	3.8	3.8	3.8	3.8
TGY	Cys	1.2	1.2	1.2	1.2	AAR	Lys	4.7	4.7	4.7	4.7
TGG	Trp	1.3	1.3	1.3	1.3	AGY	Ser2	5.5	0.0	5.5	1.8
						AGR	Arg2	5.8	0.0	5.8	1.9
Sum			9.0	24.7	16.1			22.5	33.8	26.2	

T1/A1 max = 24.7/22.5 = 1.10
T1/A1 uni = 16.1/26.2 = 0.61
T1/A1 min = 9.0/33.8 = 0.27

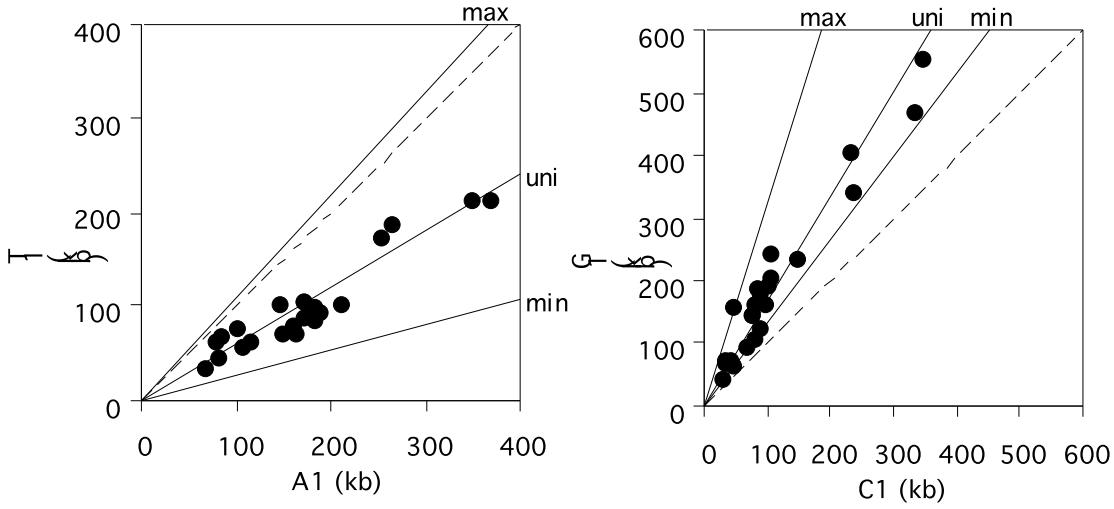
We thus expect the ratio T_1/A_1 to range from 0.27 to 1.10 depending on codon usage for Leu, Ser, and Arg with a value of 0.61 for a uniform codon usage. The low T_1/A_1 ratio is mainly

due to TRN codons: they correspond to rare amino acids (Tyr, Trp, Cys) and stop codons. In a similar way we can compute the expected G_1/C_1 ratios,

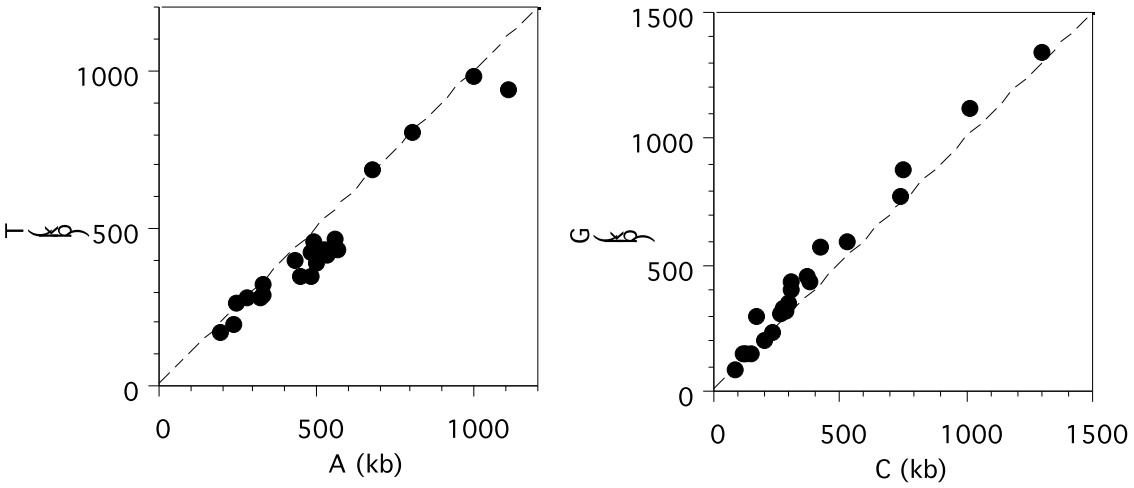
GNN	aa	aa%	min	max	uni	CNN	aa	aa%	min	max	uni
GTN	Val	7.3	7.3	7.3	7.3	CTN	Leu4	10.2	0.0	10.2	6.8
GCN	Ala	9.7	9.7	9.7	9.7	CCN	Pro	4.4	4.4	4.4	4.4
GAY	Asp	5.3	5.3	5.3	5.3	CAY	His	2.3	2.3	2.3	2.3
GAR	Glu	6.1	6.1	6.1	6.1	CAR	Gln	4.3	4.3	4.3	4.3
GGN	Gly	7.5	7.5	7.5	7.5	CGN	Arg4	5.8	0.0	5.8	3.9
Sum			35.9	35.9	35.9				11.0	27.0	21.7

G1/C1 max = 35.9/11.0 = 3.26
G1/C1 uni = 35.9/21.7 = 1.65
G1/C1 min = 35.9/27.0 = 1.33

to see that the relative G excess in first codon positions is a consequence of GNN codons corresponding to abundant amino acids in proteins. The following plot is what is observed in available complete bacterial genomes. The mean observed values ($T_1/A_1 = 0.60$, $G_1/C_1 = 1.50$) are coherent with expected values. An outlier already mentioned[186] is *Methanococcus jannaschii* with a ratio $G_1/C_1 = 3.8$ due to the low His (1.4 %) and Gln (1.4 %) frequencies in this bacteria.



In the same way we can compute expected values for remaining codon positions ($T_2/A_2 = 1.03$, $G_2/C_2 = 0.76$, $T_3/A_3 = 1.01$, $G_3/C_3 = 1.02$). Globally the A excess in first codon position is not corrected by others positions, the G excess in first codon position is partially cancelled out by second position, on average the expected ratios are T/A = 0.88 et G/C = 1.14, that is a small R-base excess in coding sequences, and this is indeed observed[177].



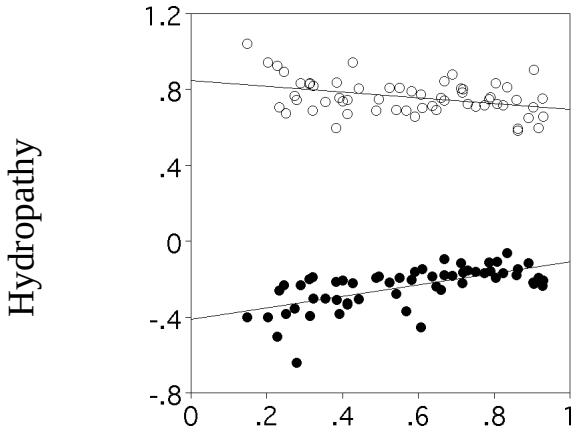
These local asymmetric selective pressures do not automatically yield a chirochore structure because if coding sequences were evenly distributed between the two strands, such biases would cancel out at a chromosomal scale. However in bacteria there is often an excess of genes on the leading strand, and this is interpreted as the result of a selective pressure to avoid head-on collisions between the RNA polymerase and the replication fork[19,193,108]. Under this hypothesis the selective pressure for a gene to be in the right orientation should increase with expressivity level, and this is effectively the case[128,142]. As a consequence, in bacteria whose gene repartition is highly biased between the two strands such as *Bacillus subtilis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, the chirochore structure could be inverted if only a given codon position is taken into account[128].

COULD A MUTATIONAL PRESSURE BE SELECTED?

Since the effect of a directional mutation pressure is a slow modification of genetic information in a population, if there is selection this should be on a much longer time scale when many populations are in competition. This is not impossible; for instance recombination in diploid and merodiploid species is an example of process which is believed to be advantageous for its long-term effect. Are there examples of adaptive utilisation of the long term effects of a directional mutation pressure?

Isochores and thermostability

Bernardi has suggested[16,13,14,15,17] that the high S-base frequency in some regions (heavy isochores) of warm-blooded vertebrate chromosomes could be advantageous for its thermostabilising properties, either directly at the DNA level or indirectly by increasing the hydrophobicity, and therefore presumably the stability, of the encoded proteins. However, heavy isochores are also present in two cold-blooded vertebrates[75] (*Crocodylus niloticus* and *Trachemys scripta elegans*). Moreover, there is complete lack of correlation between optimal growth temperature and the S-base content in bacteria[57]. Last but not least in drosophila species the highest S-base contents are observed in species living in cold environments[153]. The recent report of a correlation between S-base content and protein hydrophobicity[35] is not convincing because cytoplasmic and integral membrane proteins were analysed simultaneously despite the fact the subcellular location is known to be the first factor of protein composition variability[114]. The plot thereafter is the evolution of the average hydrophaty index[103] for proteins from 59 bacterial species[114] as a function of S-base content in third codon position. As can be seen, if there was a selective pressure one should explain why it's working in an opposite way for the two groups of proteins.



Integral Membrane Proteins

```
y = ax + b; n=59; R^2 = 0.151
a = -0.155 [-0.252,-0.057]
b = +0.849 [+0.788,+0.909]
```

Cytoplasmic Proteins

```
y = ax + b; n=59; R^2 = 0.413
a = +0.303 [+0.207,+0.399]
b = -0.413 [-0.472,-0.354]
```

<S3>

The hypothesis of a selective advantage for a high S-base content in relation with temperature is therefore not a convincing example of a selectively advantageous directional mutation pressure.

Genetic codes

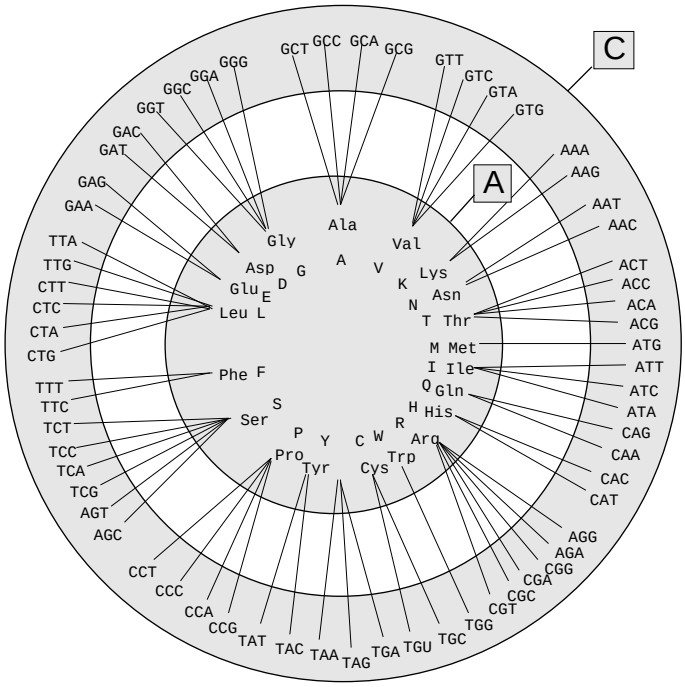
Let **C** be the set of the 64 possible codons,

C = {AAA, AAC, AAG, AAT, ..., TTT },

and **A** the set of the empty set plus the 20 possible amino acids in proteins,

A = { , Ala, Arg, ..., Val },

where the empty set represent stop codons and unassigned codons. A genetic code is a surjective function from **C** onto **A**: every element of **C** map to one element in **A** and every element of **A** is mapped to by some element of **C**, as in the example below corresponding to the so-called universal genetic code.



Genetic codes are usually represented by a 4x4 matrix crossing the first and second codon positions while the third one is listed in each entry.

« universal » genetic code							
C	**A**	**C**	**A**	**C**	**A**	**C**	**A**
TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA		TGA	
TTG	Leu	TCG	Ser	TAG		TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCC	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

A more compact notation is obtained with the one-letter code for amino acids. Neglecting the special case of the initiation codon and some translation exceptions such as selenocystein coding, a genetic code is given by a string of 64 characters with 21 possible values. Known genetic codes are represented in the alignment below where only deviations from the universal genetic code are outlined:

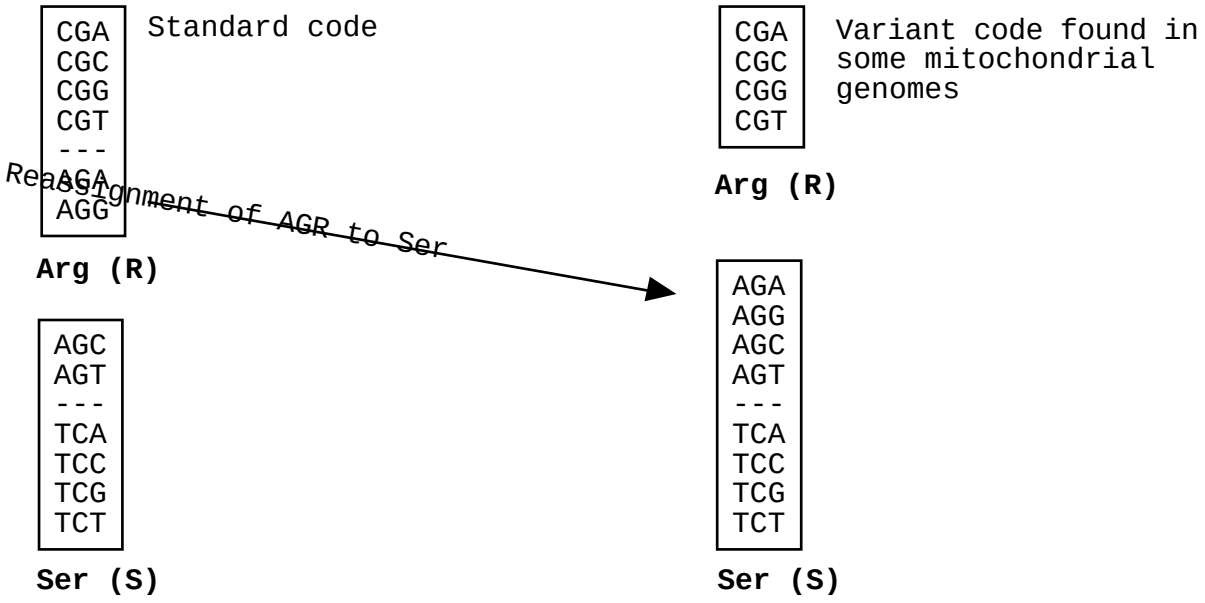
```
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG

1        FFLLSSSSYY  CC WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
4        ..............W...................................................
10       ..............C...................................................
2        ..............W...................M...........  ....................
3        ..............W.TTTT..............M...............................
5        ..............W...................M...........SS..................
21       ..............W...................M.......N...SS..................
9        ..............W...........................N...SS..................
14       ..........Y...W...........................N...SS..................
13       ..............W...................M...........GG..................
16       ..........L.......................................................
15       ..........Q.......................................................
6        ..........QQ......................................................
12       ..................S...............................................

constant XXXXXXXXXX..XX.X....XXXXXXXXXXXXXX.XXXXXX.XXX..XXXXXXXXXXXXXXXX
```

1. The Standard Code. 4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code 10. The Euplotid Nuclear Code **2**. The Vertebrate Mitochondrial Code **3**. The Yeast Mitochondrial Code **5**. The Invertebrate Mitochondrial Code **21**. Trematode Mitochondrial Code **9**. The Echinoderm Mitochondrial Code **14**. The Flatworm Mitochondrial Code **13**. The Ascidian Mitochondrial Code **16**. Chlorophycean Mitochondrial Code 15. Blepharisma Nuclear Code 6. The Ciliate, Dasycladacean and Hexamita Nuclear Code 12. The Alternative Yeast Nuclear Code

Deviations from the standard code are found at only 11 codons out of the 64. Variant codes are assumed to derive from the standard code by codon capture[81,135]: i) under a strong directional mutation pressure ($\mu_D \to 0.0$ or $\mu_D \to 1.0$) a codon is no more used in a genome ii) the codon is deassigned by a mutation in its cognate tRNA, but as it is not recognised by translation release factors a reverse mutation giving this codon is counterselected to avoid

stalled ribosomes. This intermediate situation is found in *Micrococcus luteus* ($\hat{\mu}_D = 0.95$ [174]) whose codons AGA and ATA are unassigned[87], in *Mycoplasma capricolum* ($\hat{\mu}_D = 0.07$ [174]) whose codon CGG is unassigned[134], in *Balanoglossus carnosus* mitochondria ($\hat{\mu}_D = 0.50$) whose codon AAA is unassigned[25] iii) the codon is reassigned thanks to a mutation in a tRNA or a release factor. For instance the very common reassignment of stop codon TGA to Trp was acquired independently in numerous lineages[76]; this is an example of convergence at the molecular level due to a directional mutation pressure.

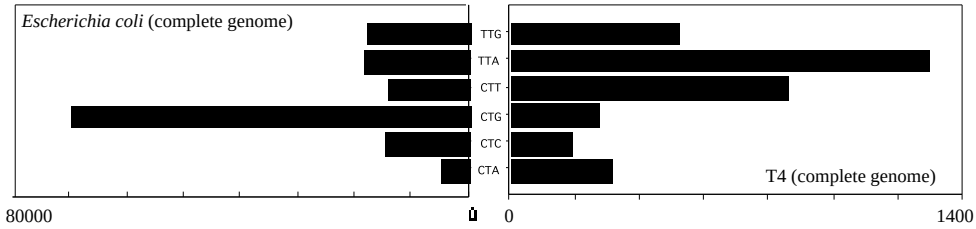
Andersson et Kurland have suggested[3,4] that codon reassignment could have an adaptive value for genome under a strong selective pressure to reduce their size such as organelles or intracellular bacteria where genome degradation is an ongoing process[1,2] as can be seen from the high proportion of non coding sequences in *Rickettsia prowazekii* (25%) or in *Mycobacterium leprae*. Because three amino acids are mapped to by six codons in the standard code, the minimum number of tRNA cannot be lower than 23[135], it is therefore significant to note that many codon reassignments allow to go below this limit such as reassignment of AGR codons to Ser,



saving one tRNA as compared to the standard code. It is then not excluded that in some cases the long terms effects of directional mutation pressure were used for translation optimisation by codon reassignment. However, in Ascidiacea mitochondria AGR codons are reassigned to Gly, this does not allow to save a tRNA and is more likely an answer to the lack of standard GGN codons for Gly because of the strong directional mutation pressure, and other codon reassignments such as AAA from Lys to Asn do not correspond to a tRNA number reduction strategy.

Bacteriophage T4

The quasi-deassignment induced by T4 infection is perhaps an example of directional mutation pressure whose long terms effect are selectively advantageous. T4 genome is under a strong directional mutation pressure ($\hat{\mu}_D = 0.22$ [92]) in contrast to its host *Escherichia coli* ($\hat{\mu}_D = 0.55$ [174]), this is *a priori* not selectively advantageous because T4 has to encode its own tRNA instead of simply following the host codon usage to optimise translation and reduce its genome size. For example, codon usage for Leucine is as follows for *Escherichia coli* and T4 phage:



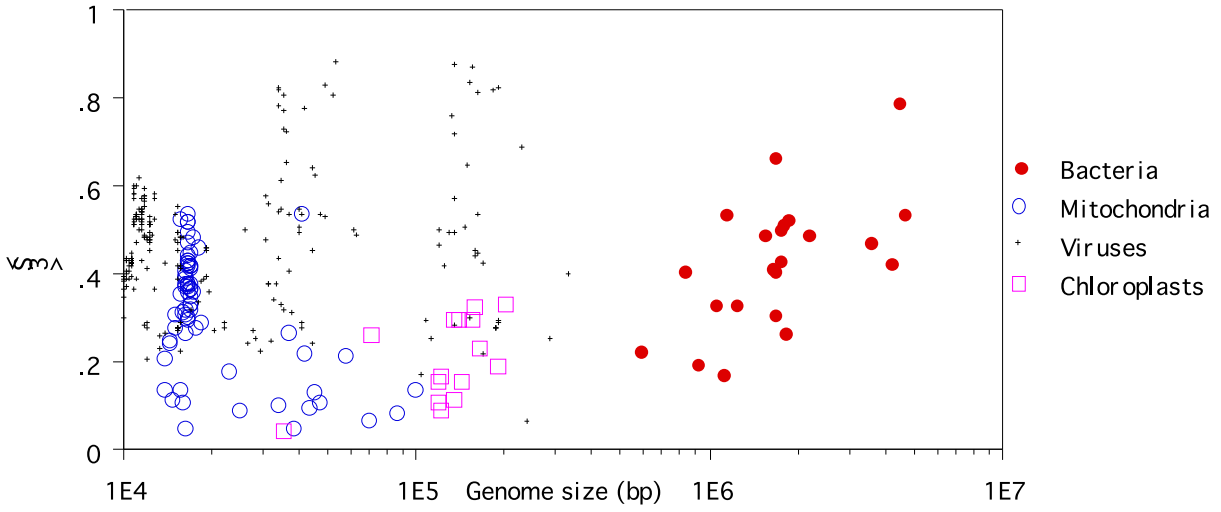
Hence, *Escherichia coli* favoured codon is CTG, its cognate Leu-tRNA$_1$ is one of the most abundant tRNA in the cell[77], which is not surprising since CTG is the most frequent codon among all codons. But in T4 genome, TTA is the favoured codon and is recognised by a T4-encoded tRNA. More generally, codons that are recognised by T4-encoded tRNA are more frequent in T4 coding sequences, and this trend is exacerbated for genes that expressed at the end of the phage infection cycle[32]. Tamiko and Noboru Sueoka have shown[88,91,84,89,90,85] that less than two minutes after infection of an *Escherichia coli* cell by T4, the Leu-tRNA$_1$ concentration decreases dramatically. This quasi-deassignment induced by T4 infection (whose molecular detail are very complex[98]) clearly advantages a codon usage in T4 genome differing from its host. As far as I know this is the only clear example of a long term selectively advantageous effect of directional mutation pressure.

Asymmetric mutation rate

Furusawa and Doi have shown[55,185,39,56] with computer simulations that a different mutation specific rate between the two DNA strands could be advantageous in the long term by allowing populations to handle high mutations rates while still preserving optimal individual thanks the asymmetric repartition of mutants in the population. Are asymmetric directional mutation pressures an example of such process? This is an open question, highly speculative because based only upon *in silico* simulations whose conditions, such as using only one strand as the "coding" strand, are questionable from a biological point of view.

Genome size and S-base frequencies

Data from complete genome do not allow inferring a clear relationship between genome size and their S-base frequencies.



There is a trend for organelles and intracellular bacteria to have small genomes and low S-base contents[129,72], there is for instance no known mitochondria with more than 55% of S-base in third codon positions. That genome size reduction induces the loss of DNA repair enzymes, and therefore a higher sensitivity to directional mutation pressures, is expectable. What is unclear, however, is why mutation pressure should always be directed toward a low S-base content.

Escherichia coli chromosome polarisation

The *dif* locus, close to *Escherichia coli* chromosome replication terminus, is essential to monomerise chromosome dimers due to homologous recombination: about 15% of cells are involved during exponential growth phase and dif^- mutants are eliminated when competing with dif^+ individuals, except if they are both $recA^-$ and deficient for recombination[139]. What is special with *dif* locus is its location and context dependence: its activity progressively decreases when it is moved away from its original position and is no more active 30 kb away, and its activity is cancelled by upstream or downstream chromosome inversions. The current model is that there are local asymmetric signals in the *dif* activity zone allowing for a correct positioning at the septum level of the two *dif* locus partners required for dimer resolution[31].

Then, under this model, there is a selective pressure to preserve asymmetric signals in the *dif* activity zone. However, deletion mutants of the *dif* site and of the whole *dif* activity zone (up to 155 kb upstream and 59 kb downstream) recover the wild-type phenotype when *dif* is reinserted at the deletion junction point! In other words, there are no asymmetric signals specific of the *dif* activity zone: how could this work? It is tempting to speculate that *dif* activity is based upon asymmetric signals present on the whole chromosome but for another reason. Salzberg's group has shown that many asymmetric oligomers are unevenly distributed between the two strands[158]. Is it and adaptive recycling of the long-term effects of the asymmetric directional mutation pressure? The base composition of the leading strand in *Escherichia coli* chromosome (A=1137535, C=1140273, G=1215935, T=1145478) correspond to a small K-base enrichment (50.9%), the expected ratio of the number of K_n oligomers on the leading strand over its number on the lagging strand, $(0.509/0.491)^n$, is only 1.33 for octamers. The asymmetric directional mutation pressure seems too weak in *Escherichia coli* to produce a highly biased repartition of oligomers between the two strands. On the other hand, it is puzzling to note that the strongest asymmetric directional mutation pressure is found in *Borrelia burgdorferi* chromosome whose linear structure with covalently closed single-strand hairpin loops at its ends makes that 100% of cells are concerned with chromosome dimer resolution after replication.

CONCLUSION AND FUTURE DIRECTIONS

> La biologie positive doit donc être envisagée comme ayant pour destination générale de rattacher constamment l'un à l'autre, dans chaque cas déterminé, le point de vue anatomique et le point de vue physiologique, ou, en d'autres termes, l'état statique et l'état dynamique. Cette relation perpétuelle constitue son vrai caractère philosophique.
>
> Auguste Compte
> Cours de philosophie positive
> 1840-1842

Assuming that a symmetric process with respect to the two DNA strands governs the evolution of DNA bases frequencies; a nice wrong model is obtained. This model is nice because it can be rejected from the sole inspection of base frequencies in DNA and its rejection, which is effective in many genomes, means that the underlying non-observable process is asymmetric.

For some species such as *Borrelia burgdorferi* the most likely biological interpretation is that there is an asymmetric directional mutation pressure, a selective alternative is extremely difficult to imagine because an *intragenomic* diversifying selective pressure should be postulate to explain the codon usage schizophrenia. An adaptive recycling of the long-term effects of the asymmetric directional mutation pressure effects is not excluded if a polarised chromosome is selectively advantageous; this is an unanswered question.

The universality of the biases induced by asymmetric directional mutation pressure suggests a common underlying mechanism. This is a puzzling question because invariants are rare in biology. The theory of accelerated cytosine deamination in single stranded DNA during replication is interesting but is hard to challenge. It is important to understand the origin of this universality because compositional biases are very high in pathogenic bacteria such as *Borrelia burgdorferi* (Lyme disease) *Chlamydia pneumoniae* (pneumonia) *Chlamydia trachomatis* (trachoma) *Rickettsia prowazekii* (typhus) et *Treponema pallidum* (syphilis) and completely absent in human. Are antimicrobial agents specifically targeted against highly biased genomes possibles? If a deficient handling of the single stranded state during replication were at their origin, transcription and then the whole metabolism would also be targeted. However, observed biases are the result of a long evolutionary story and a small difference, too small for being useful at the human time scale, could be at their origin. Whatever, a first step is to understand the reason the universality of these biases.

The modelling, that is the translation in a mathematical or computerised formal system, of the cytosine deamination theory or any alternative theory is a major bottleneck. According to my own experience modelling is extremely slow, tedious and expensive. The interpretation of base composition biases is far from being obvious, and without a clear theoretical analysis of the expected results under a given model, I don’t see how we could progress towards a better understanding of the underlying mechanism(s). We are perhaps blind to results already present in databases just because we don’t have the right approach. A research effort for a better understanding of DNA base frequency evolution under asymmetric conditions is then required.

REFERENCES

1. Andersson, J.O., Andersson, S.G.E. (1999) Insights into the evolutionary process of genome degradation. *Current Opinion in Genetics & Development*, **9**:664-671.

2. Andersson, J.O., Andersson, S.G.E. (1999) Genome degradation is an ongoing process in *Rickettsia*. *Molecular Biology and Evolution*, **16**:1178-1191.

3. Andersson, S.G.E., Kurland, C.G. (1990) Codon preferences in free-living microorganisms. *Microbiological Reviews*, **54**:198-210.

4. Andersson, S.G.E., Kurland, C.G. (1991) An extreme codon preference strategy: codon reassignment. *Molecular Biology and Evolution*, **8**:530-544.

5. Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontérn, T., Alsmark, U.C.M., Podowski, R.M., Naslund, A.K., Eriksson, A.-S., Winkler, H.H., Kurland, C.G. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**:133-140.

6. Anonymous (1986) Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). *Proceedings of the National Academy of Sciences of the United States of America*, **83**:4-8.

7. Anonymous (2000) Instructions to Authors. *Molecular Biology and Evolution*, **17**:207-212.

8. Asakawa, S., Kumazawa, Y., Araki, T., Himeno, H., Miura, K.-I., Watanabe, K. (1991) Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. *Journal of Molecular Evolution*, **32**:511-520.

9. Baker, T.A., Wickner, S.H. (1992) Genetics and enzymology of DNA replication in *Escherichia coli*. *Annual Review of Genetics*, **26**:447-477.

10. Barbour, A.G. (1993). Linear DNA of *Borrelia* species and antigenic variation. *Trends in Microbiology*, **1**:236-239.

11. Bell, S.J., Forsdyke, D.R. (1999) Accounting units in DNA. *Journal of Theoretical Biology*, **197**:51-61.

12. Berkhout, B., van Hemert, F.J. (1994) The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. *Nucleic Acids Research*, **22**:1705-1711.

13. Bernardi, G. (1989) The isochore organization of the human genome. *Annual Review of Genetics*, **23**:637-661.

14. Bernardi, G. (1993) The vertebrate genome: isochores and evolution. *Molecular Biology and Evolution*, **10**:186-204.

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

16. Bernardi, G. and Bernardi, G. (1986) Compositional constraints and genome evolution. *Journal of Molecular Evolution*, **24**:1-11.

17. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953-958.

18. Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Rilley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**:1453-1462.

19. Brewer, B.J. (1988) When polymerase collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell*, **53**:679-686.

20. Bronson, E.C., Anderson, J.N. (1994) Nucleotide composition as a driving force in the evolution of retroviruses. *Journal of Molecular Evolution*, **38**:506-532.

21. Brown, G.G., Simpson, M.V. (1982) Novel features of animal mtDNA evolution as shown by sequences oftwo rat cytochrome oxidase subunit II genes. *Proceedings of the National Academy of Sciences of the United States of America*, **79**:3246-3250.

22. Bulmer, M. (1991). Strand symmetry of mutation rates in the ß-globin region. *Journal of Molecular Evolution*, **33**:305-310.

23. Capiaux, H., Cornet, F., Corre, J., Guijo, M.-I., Pérals, K., Rebollo, J.E., Louarn, J.-M. (2001) Polarization of the *Escherichia coli* chromosome. A view from the terminus. *Biochimie*, **83**:161-170.

24. Casjens, S., Palmer, N., van Vugt, R., Huang, W.M., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R.J., Haft, D., Hickey, E., Gwinn, M., White, O., Fraser, C.M. (2000) A bacteria genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of Lyme disease spirochete *Borrelia burgdorferi*. *Molecular Microbiology*, **35**:490-516.

25. Castresana, J., Feldmaier-Fuchs, G., Pääbo, S. (1998) Codon reassignment and amino

acid composition in hemichordate mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, **95**:3703-3707.

26. Cebrat, S., Dudek, M.R. (1998) The effect of DNA phase structure on DNA walks. *The European Physical Journal B*, **3**:271-276.

27. Cebrat, S., Dudek, M.R., Gierlik, A., Kowalczuk, M., Mackiewicz, P. (1999) Effect replication on the third base of codons. *Physica A*, **265**:78-84.

28. Chargaff, E. (1979) How genetics got a chemical education. *Annals of the New York Academy of Sciences*, **325**:345-360.

29. Clark, M.A., Moran, N.A., Baumann, P. (1999) Sequence evolution in bacterial endosymbionts having extreme base compositions. *Molecular Biology and Evolution*, **16**:1586-1598.

30. Collins, D.W., Jukes, T.H. (1993) Relationship between G+C in silent sites of codons and amino acid composition of human proteins. *Journal of Molecular Evolution*, **36**:201-213.

31. Cornet, F., Louarn, J., Patte, J., Louarn, J.-M. (1996) Restriction of the activity of the recombination site *dif* to a small zone of the *Escherichia coli* chromosome. *Genes & Development*, **10**:1152-1161.

32. Cowe, E., Sharp, P.M. (1991) Molecular evolution of bacteriophages: discrete patterns of codon usage in T4 genes are related to the time of gene expression. *Journal of Molecular Evolution*, **33**:13-22.

33. Cruveiller, S., Jabbari, K., D'Onofrio, G., Bernardi, G. (1999) Different hydrophobicities of orthologous proteins from *Xenopus* and human. *Gene*, **238**:15-21.

34. Danchin, A. (1983) L'œuf et la poule-histoires du code génétique. Fayard, Paris, France. ISBN 2-213-01302-0.

35. D'Onofrio, G., Jabbari, K., Musto, H., Bernardi, G. (1999) The correlation of protein hydropathy with the base composition of coding sequences. *Gene*, **238**:3-14.

36. D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., Bernardi, G. (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *Journal of Molecular Evolution*, **32**:504-510.

37. Danchin, A. (1998) The Delphic boat or what the genomic texts tell us. *Bioinformatics*, **14**:383-383.

38. Danchin, A. (1999) From function to sequence, an integrated view of the genome texts. *Physica A*, **273**:92-98.

39. Doi, H., Furusawa, M. (1996) Evolution is promoted by asymmetrical mutations in DNA replication - genetic algorithm with double-stranded DNA -. *FUJITSU Sci. Tech. J.*, **2**:248-255.

40. Fickett, J.W., Torney, D.C., Wolf, D.R. (1992) Base compositional structure of genomes. *Genomics*, **13**:1056-1064.

41. Filipski J (1990) Evolution of DNA sequences. Contributions of mutational bias and selection to the origin of chromosomal compartments. pp 1-54 in G. Obe (Ed.) Advances in mutagenesis research 2, Springer Verlag.

42. Forsdyke, D.R. (1995) Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. *Journal of Molecular Evolution*, **41**:573-581.

43. Foster, D.M., Jacquez, J.A. (1975) Multiple zeros for eigenvalues and the multiplicity of traps of a linear compartmental system. *Mathematical Biosciences*, **26**:89-97.

44. Francino, M.P., Chao, L., Riley, M.A. and Ochman, H. (1996) Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science*, **272**:107-109.

45. Francino, M.P., Ochman, H. (1997) Strand asymetries in DNA evolution. *Trends in Genetics*, **13**:240-245.

46. Francino, M.P., Ochman, H. (1999) A comparative genomics approach to DNA asymmetry. *Annals of the New York Academy of Sciences*, **870**:428-431.

47. Francino, M.P., Ochman, H. (2000) Strand symmetry around the ß-globin origin of replication in primates. *Molecular Biology and Evolution*, **17**:416-422.

48. Frank, A.C., Lobry, J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, **238**:65-77.

49. Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fuji, C., Cotton, M.D., Horst, K., Roberts, K., Hatch, B., Smith, H.O., Venter, J.C. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**:580-586.

50. Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G., Dodson, R., Gwinn, M., Hickey, E.K., Clayton, R., Ketchum,

K.A., Sodergren, E., Hardham, J.M., McLeod, M.P., Salzberg, S., Peterson, J., Khalak, H., Richardson, D., Howell, J.K., Chidambaram, M., Utterback, T., McDonnald, L., Artiach, P., Bowman, C., Cotton, M.D., Fujii, C., Garland, S., Hatch, B., Horst, K., Roberts, K., Sandusky, M., Weidman, J., Smith, H.O., Venter, J.C. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*, **281**:375-388.

51. Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A., Venter, J.C. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**:397-403.

52. Frederico, L.A., Kunkel, T.A., Shaw, B.R. (1990) A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, **29**:2532-2537.

53. Freeman, J.M., Plasterer, T.N., Smith, T.F., Mohr, S.C. (1998) Patterns of genome organization in bacteria. *Science*, **279**:1827.

54. Freese, E. (1962) On the evolution of the base composition of DNA. *Journal of Theoretical Biology*, **3**:82-101.

55. Furusawa, M., Doi, H. (1992) Promotion of evolution: disparity in the frequency of strand-specific misreading between the lagging and leading DNA strands enhances disproportionate accumulation of mutations. *Journal of Theoretical Biology*, **157**:127-133.

56. Furusawa, M., Doi, H. (1998) Asymmetrical DNA replication promotes evolution: disparity theory of evolution. *Genetica*, **102/103**:333-347.

57. Galtier, N., Lobry, J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, **44**:632-636.

58. Gautier, C. (2000) Compositional bias in DNA. *Current Opinion in Genetics & Develoment*, **10**:656-661.

59. Gillespie, J.H. (1991) The causes of molecular evolution. Oxford University press. ISBN 0-19-506883-1.

60. Gouy, M., Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, **10**:7055-7073.

61. Gouy, M., Milleret F, Mugnier C, Jacobzone M, Gautier C (1984) ACNUC: a nucleic acid sequence data base and analysis system. *Nucleic Acids Research*, **12**:121-127.

62. Grantham, R., Gautier, C. (1980) Genetic distances from mRNA sequences. *Naturwissenschaften*, **67**:93-94.

63. Grantham, R., Gautier, C., Gouy, M. (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, **8**:1893-1912.

64. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. (1980) Codon catalog usage and the genome hypothesis *Nucleic Acids Research*, **8**:r49-r62.

65. Graur, D., Li, W.-H. (2000) Fundamentals of molecular evolution, Second Edition. Sinauer Associates Inc., Sunderland, Massachusetts, USA. ISBN 0-87893-266-6.

66. Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research*, **26** :2286-2290.

67. Grigoriev, A. (1999) Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Research*, **60**:1-19.

68. Grigoriev, A., Freeman, J.M., Plasterer, T.N., Smith, T.F., Mohr, S.C. (1998) Genome arithmetic. *Science* **281**:1923-1924.

69. Gu, X., Hewett-Emmett, D., Li, W.-H. (1998) Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica*, **102/103**:383-391.

70. Hanai, R., Wada, A. (1988) The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. *Journal of Molecular Evolution*, **27**:321-325.

71. Haney, P.J., Badger, J.H., Buldak, G.L., Teich, C.I., Woese, C.R., Olsen, G.J. (1999) Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proceedings of the National Academy of Sciences of the United States of America*, **96**:3578-3583.

72. Heddi, A., Charles, H., Khatchadourian, C., Bonnot, G., Nardon, P. (1998) Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G+C content of an endocytobiotic DNA. *Journal of Molecular Evolution*, **47**:52-61.

73. Holm, L. (1986) Codon usage and gene expression. *Nucleic Acids Research*, **27**:244-247.

74. Holmes,W.M., Goldman,E., Miner,T.A. and Hatfield,G.W. (1977) *Proceedings of the National Academy of Sciences of the United States of America*, **74**:1393-1397.

75. Hughes, S., Zelus, D., Mouchiroud, D. (1999) Warm-blooded isochore structure in nile crocodile and turtle. *Molecular Biology and Evolution*, **16**:1521-1527.

76. Inagaki, Y., Ehara, M., Watanabe, K.I., Hayashi-Ishimaru, Y., Ohama, T. (1998) Directionally evolving genetic code: the UGA codon from stop to tryptophan in mitochondria. *Journal of Molecular Evolution*, **47**:378-384.

77. Ikemura,T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology*, **146**:1-21.

78. Jacquez, J.A., Simon, C.P. (1993) Qualitative theory of compartmental systems. *SIAM Review*, **35**:43-79

79. Jermiin, L.S., Graur, D., Crozier, R.H. (1995) Evidence from analyses of intergenic region for strand-specific directional mutation pressure in metazoan mitochondrial DNA. *Molecular Biology and Evolution*, **12**:558-563.

80. Jermiin, L.S., Graur, D., Lowe, R.M., Crozier, R.H. (1994) Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. *Journal of Molecular Evolution*, **39**:160-173.

81. Jukes, T.H. (1985) A change in the genetic code in *Mycoplasma capricolum*. *Journal of Molecular Evolution*, **22**:361-362.

82. Jukes, T.H., Bhushan, V. (1986) Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. *Journal of Molecular Evolution*, **24**:39-44.

83. Jukes, T.H., Holmquist, R. and Moise, H. (1975) Amino acid composition of proteins: selection against the genetic code. *Science* **189**:50-51.

84. Kan, J., Kano-Sueoka, T., Sueoka, N. (1968) Characterization of leucine transfer ribonucleic acid in *Escherichia coli* following infection with bacteriophage T2. *Journal of Biological Chemistry*, **243**:5584-5590.

85. Kan, J., Nirenberg, M., Sueoka, N. (1970) Coding specificity of *Escherichia coli* leucine tRNA before and after infection with bacteriophage T2. *Journal of Molecular Biololgy*, **52**:179-193.

86. Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNA: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238**:143-155.

87. Kano, A., Ohama, T., Abe, R., Osawa, S. (1993) Unassigned or nonsense codons in *Micrococcus luteus*. *Journal of Molecular Biology*, **230**:51-56.

88. Kano-Sueoka, T., Sueoka, N. (1966) Modifications of leucyl-sRNA after bacteriophage infection. *Journal of Molecular Biology*, **20**:183-209.

89. Kano-Sueoka, T., Sueoka, N. (1968) Characterization of a modified leucyl-tRNA of *Escherichia coli* after bacteriophage T2 infection. *Journal of Molecular Biology*, **37**:475-491.

90. Kano-Sueoka, T., Sueoka, N. (1969) Leucine tRNA and cessation of *Escherichia coli* protein synthesis upon phage T2 infection. *Proceedings of the National Academy of Sciences of the United States of America*, **62**:1229-1236.

91. Kano-Sueoka, T., Nirenberg, M., Sueoka, N. (1968) Effect of bacteriophage infection upon the specificity of leucine transfer RNA for RNA codewords. *Journal of Molecular Biology*, **35**:1-12.

92. Kano-Sueoka, T., Lobry, J.R., Sueoka, N. (1999) Intra-strand biases in bacteriophage T4 genome. *Gene*, **238**: 59-64.

93. Karkas, J.D., Rudner, R., Chargaff, E. (1968) Separation of *B. subtilis* DNA into complementary strands, II. Template functions and composition as determined by transcription with RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, **60**:915-920.

94. Karkas, J.D., Rudner, R., Chargaff, E. (1970) Template properties of complementary fractions of denatured microbial deoxyribonucleic acids. *Proceedings of the National Academy of Sciences of the United States of America*, **65**:1049-1056.

95. Karlin, S. (1999) Bacterial DNA strand compositional asymmetry. *Trends in Microbiology*, **7**:305-308.

96. Karlin, S., Blaisdell, B.E., Schachtel, G.A. (1990) Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. *Journal of Virology*, **64**:4264-4273.

97. Karlin, S., Campbell, A.M., Mrázek, J. (1998) Comparative DNA analysis across diverse genomes. *Annual Review of Genetics*, **23**:185-225.

98. Kaufmann, G. (2000) Anticodon nucleases. *Trends in Biochemical Sciences*, **25**:70-74.

99. Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature*, **217**:624-626.

100. Kimura, M. (1985) Diffusion models in population genetics with special reference to fixation time of molecular mutants under mutational pressure. In Population Genetics and Molecular Evolution, pp. 19-39 ed. T. Ohta and K. Aoki. Tokyo: Japan Scientific Societies Press / Berlin: Springer-Verlag.

101. King, J.L., Jukes, T.H. (1969) Non-Darwinian evolution. *Science*, **164**:788-798.

102. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.-K., Codani, J.-J., Connerton, I.F., Cummings, N.J., Daniel, R.A., Denizot, F., Devine, K.M., Düsterhöft, A., Ehrlich, S.D., Emmerson, P.T., Entian, K.D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E.J., Grandi, G., Guiseppi, G., Guy, B.J., Haga, K., Haiech, J., Harwood, C.R., Hénaut, A., Hilbet, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R.P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T.M., Portetelle, D., Porwollik, S., Prescott, A.M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Serror, S.J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanakata, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H., Danchin, A. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**:249-256.

103. Kyte, J., Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, **157**:105-132.

104. Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M., Wolfe, K.H. (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Research*, **27**:1642-1649.

105. Li, W. (1999) Statistical properties of open reading frames in complete genome sequences. *Computer & Chemistry* **23**:283-301.

106. Li, W.-H. (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *Journal of Molecular Evolution*, **24**:337-345.

107. Li, W.-H. (1997). Molecular evolution. Sinauer Associates, Sunderland Massachusetts U.S.A.

108. Liu, B., Alberts, M. (1995) Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. *Science*, **267**:1131-1137.

109. Liò, P., Goldman, N. (1998) Models of molecular evolution and phylogeny. *Genome Research*, **8**:1233-1244.

110. Lobry, J.R. (1995) Properties of a general model of DNA evolution under no-strand bias conditions. *Journal of Molecular Evolution*, **40**:326-330 ; **41**:680.

111. Lobry, J.R. (1996) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**:323-326.

112. Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, **13**:660-665.

113. Lobry, J.R. (1996) Origin of replication of *Mycoplasma genitalium. Science*, **272**:745-746.

114. Lobry, J.R. (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, **205**:309-316.

115. Lobry, J.R. (1999) A nice wrong model for the evolution of DNA base frequencies. *Physica A*, **273**:99-102.

116. Lobry, J.R. (1999) Genomic landscapes. *Microbiology Today*, **26**:164-165.

117. Lobry, J.R., Sueoka, N. (2000) Asymmetric directional mutation pressures in bacteria. *in preparation*.

118. Lobry, J. R., Gautier, C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research*, **22**:3174-3180.

119. Lobry, J.R., Lobry, C. (1999) Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Molecular Biology and Evolution*, **16**:719-723.

120. Lopez, P., Philippe, H., Myllykallio, H., Forterre, P. (1999) Identification of putative chromosomal origins of replication in Archaea. *Molecular Microbiology*, **32**:883-891.

121. Lyamichev, V., Panyutin, I., Frank-Kamenetskii, M.D. (1984) The absence of cruciform structure from pA03 plasmid DNA *in vivo*. *Journal of Biomolecular Structure and Dynamics*, **2**:291-301.

122. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M.R., Cebrat, S. (1999) How does replication-associated mutational pressure influence amino acid composition of proteins? *Genome Research*, **9**:409-416.

123. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M.R., Cebrat, S. (1999) Asymmetry of nucleotide composition of prokaryotic chromosomes. *Journal of Applied Genetics*., **40**:1-14.

124. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Szczepanik, D., Dudek, M.R., Cebrat, S. (1999) Mechanisms generating long-range correlation in nucleotide composition of the *Borrelia burgdorferi* genome. *Physica A*, **273**:103-115.

125. Marians, K.J. (1992) Prokaryotic DNA replication. *Annual Review of Biochemistry*, **61**:673-719.

126. McDonald, J.H., Grasso, A.M., Rejto, L.K. (1999) Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. *Molecular Biology and Evolution*, **16**:1785-1790.

127. McInerney, J.O. (1998) Replication and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proceedings of the National Academy of Sciences of the United States of America*, **95**:10698-10703.

128. McLean, M.J., Wolfe, K.H., Devine, K.M. (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *Journal of Molecular Evolution* **47**:691-696.

129. Moran, N. (1996) Accelerated evolution and muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **93**:2873-2878.

130. Morton, B.R. (1999) Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis*. *Proceedings of the National Academy of Sciences of the United States of America*, **96**:5123-5128

131. Mrázek, J., Karlin, S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **95**:3720-3725.

132. Nossal, N.G. (1983) Prokaryotic DNA replication systems. *Annual Review of Biochemistry*, **52**:581-615.

133. Nussinov, R. (1982) Some indications for inverse DNA duplication. *Journal of Theoretical Biology*, **95**:783-791.

134. Oba, T., Andachi, Y., Muto, A., Osawara, S. (1991) CGG: an unassigned codon in *Mycoplasma capricolum*. *Proceedings of the National Academy of Sciences of the United States of America*, **88**:921-925.

135. Osawa, S., Jukes, T.H., Watanabe, K., Muto, A. (1992) Recent evidence for evolution of the genetic code. *Microbiological Review*s, **56**:229-264.

136. Panyutin, I., Jlishko, V., Lyamichev, V. (1984) Kinetics of cruciform formation and stability of cruciform structure in superhelical DNA. *Journal of Biomolecular Structure and Dynamics*, **1**:1311-1324.

137. Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., Davies, R.M., Davis, P., Devlin, K., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Leather, S., Moule, S., Mungall, K., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., Simmonds, M., Skelton, J., Whitehead, S., Spratt, B.G., Barrell, B.G. (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*, **404**:502-506.

138. Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A.V., Moule, S., Pallen, M.J., Penn, C.W., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., van Vhet, A.H.M., Withehead, S., Barrell, B.G. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**:665-668.

139. Pérals, K., Cornet, F., Merlet, Y., Delon, I., Louarn, J.-M. (2000) Functional polarization of the *Escherichia coli* chromosome terminus: the *dif* site acts in chromosome dimer resolution only when located between long stetches of opposite polarity. *Molecular Microbiology*, **36**:33-43.

140. Perna, N.T., Kocher, T.D. (1995) Patterns of nucleotide composition at fourfold degenerate site of animal mitochondria genomes. *Journal of Molecular Evolution*, **41**:353-358.

141. Perrière, G., Bessières, P., Labedan, B. (2000) EMGLib: the enhanced microbial genomes library (update 2000). *Nucleic Acids Research*, **28**:68-71.

142. Perrière, G., Lobry, J.R. (1998) Asymmetrical coding sequence repartition and codon adaptation index values between leading and lagging strands in seven bacterial species. *in* Proceedings of the first international conference on bioinformatics of genome regulation and structure (Novosibirsk, Russia, August 24-31, 1998) **2**:254-255.

143. Perrière, G., Lobry, J.R., Thioulouse, J. (1996) Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acids sequences. *Computer Applications in the Biosciences*, **12**:519-524.

144. Picardeau, M., Lobry, J.R., Hinnebusch, B.J. (1999) Physical mapping of an origin of bidirectional replication at the center of the *Borrelia burgdorferi* linear chromosome. *Molecular Microbiol*ogy, **32**:437-445.

145. Prabhu, V.V. (1993) Symetry observation in long nucleotide sequences. *Nucleic Acids Research*, **21**:2797-2800.

146. Rand, D.M., Kann, L.M. (1998) Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. *Genetica* **102/103**:393-407.

147. Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S.L., Eisen, J., Fraser, C.M. (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Research*, **28**:1397-1406.

148. Reyes, A., Gissi, C., Pesole, G., Saccone, C. (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Molecular Biology and Evolution*, **15**:957-966.

149. Rocha, E.P.C., Danchin, A., Viari, A. (1999) Bacterial DNA strand compositional asymmetry: Response. *Trends in Microbiology*, **7**:308-308.

150. Rocha, E.P.C., Danchin, A., Viari, A. (1999) Universal replication biases in bacteria. *Molecular Microbiology*, **32**:11-16.

151. Rocha, E.P.C., Viari, A., Danchin, A. (1998) Oligonucleotide bias in *Bacillus subtilis* : general trends and taxonomic comparisons. *Nucleic Acids Research*, **26**:2971-2980.

152. Rodríguez F., Olivier, J.L., Marín, A., Medina, J.R. (1990) The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, **142**:485-501.

153. Rodríguez -Trelles, F., Tarrío, R., Ayala, F.J. (2000) Fluctuating mutation bias and the evolution of base composition in *Drosophila*. *Journal of Molecular Evolution*, **50**:1-10.

154. Rosso, L., Lobry, J.R., Flandrois, J.P. (1993) An unexpected correlation between cardinal temperatures of microbial growth highlighted by a new model. *Journal of Theoretical Biology*, **162**:447-463.

155. Rudner, R., Karkas, J.D., Chargaff, E. (1968) Separation of *B. subtilis* DNA into complementary strands, III. Direct analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **60**:921-922.

156. Rudner, R., Karkas, J.D., Chargaff, E. (1969) Separation of microbial deoxyribonucleic acids into complementary strands. *Proceedings of the National Academy of Sciences of the United States of America*, **63**:152-159.

157. Rudner, R., LeDoux, M. (1974) Distribution of pyrimidine oligonucleotides in complementary strand fractions of *Escherichia coli* desoxyribonucleic acid. *Biochemistry*, **13**:118-125.

158. Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., Tomb, J.-F. (1998) Skewed oligomers and origins of replication. *Gene*, **217**:57-67.

159. Sharp, P.M., Matassi, G. (1994) Codon usage and genome evolution. *Current Opinion in Genetics and Development*, **4**:851-860.

160. Sharp, P.M., Li, W.-H. (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**:1281-1295.

161. Sharp, P.M., Tuohy, T.M.F., Mosurski, K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, **14**:5125-5143.

162. Shpaer, E.G. (1989) Amino acid composition is correlated with protein abundance in Escherichia coli: can this be due to optimization of translational efficiency? *Protein Sequences and Data Analysis*, **2**:107-110.

163. Sinden, R.R., Broyles, S.S., Pettijohn, E. (1983) Perfect palindromic *lac* operator DNA sequence exists as a stable cruciform strucutre in supercoiled DNA *in vitro* but not *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America*, **80**:1797-1801.

164. Sueoka, N. (1959) A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. *Proceedings of the National Academy of Sciences of the United States of America*, **45**:1480-1490.

165. Sueoka, N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proceedings of the National Academy of Sciences of the United States of America*, **47**:1141-1149.

166. Sueoka, N. (1961) Variation and heterogeneity of base composition of deoxyribonucleic acids: a compilation of old and new data. *Journal of Molecular Biology*, **3**:31-40.

167. Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America*, **48**:582-592.

168. Sueoka, N. (1964) On the evolution of informational macromolecules. pp 479-496 in Bryson, V. and Vogel, H.J. (eds), *Evolving genes and proteins*. Academic Press, New York, USA.

169. Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **85**:2653-2657.

170. Sueoka, N. (1992) Directional mutation pressure, selective constraints, and genetic equilibria.*Journal of Molecular Evolution*, **34**:95-114.

171. Sueoka, N. (1993) Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. *Journal of Molecular Evolution*, **37**:137-153.

172. Sueoka, N. (1995) Intrastrand parity rules of DNA base composition and usages biases of synonymous codons. *Journal of Molecular Evolution*, **40**:318-325; **42**:323.

173. Sueoka, N. (1999) Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene*, **238**:53-58.

174. Sueoka, N. (1999) Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. *Journal of Molecular Evolution*, **49**:49-62.

175. Sueoka, N., Marmur, J. and Doty, P. (1959) Heterogeneity in deoxyribonucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine. *Nature* , **183**:1427-1431.

176. Sueoka, N., Kano-Sueoka, T. (1964) A specific modification of leucyl-sRNA of *Escherichia coli* after phage T2 infection. *Proceedings of the National Academy of Sciences of the United States of America*, **52**:535-1540.

177. Szybalski, W., Kubinski, H., Sheldrick, P. (1966) Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. *Cold Sping Harbor Symposium on Quantitative Biology*, **31**:123-127.

178. Tanaka, M., Ozawa, T. (1994) Strand asymmetry in human mitochondrial DNA mutations. *Genomics*, **22**:327-335.

179. Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., Nelson, W.C., Gwinn, M.L., DeBoy, R., Peterson, J.D., Hickey, E.K., Haft, D.H., Salzberg, S.L., White, O., Fleischmann, R.D., Dougherty, B.A., Mason, T., Ciecko, A., Parksey, D.S., Blair, E., Cittone, H., Clark, E.B., Cotton, M.D., Utterback, T.R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, V., Masignani, V., Pizza, M., Grandi, G., Smith, H.O., Fraser, C.M., Moxon, E.R., Rappuoli, R., Venter, J.C. (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, **287**:1809-1815.

180. Thioulouse, J., Dolédec, S., Chessel, D. and Olivier, J.M. (1995) ADE software multivariate analysis and graphical display of environmental data. In Guariso,G. and Rizzoli,A. (eds), *Software per l'Ambiente*. Pàtron, Bolonia, pp. 57-62.

181. Thioulouse, J., Lobry, J.R. (1995) Co-inertia alaysis of amino-acid physico-chemical properties and protein composition with the ADE package. *Computer Applications in the Biosciences*, **11**:321-329.

182. Tillier, E.R.M., Collins, R.A. (2000) The contribution of replication orientation, gene direction, and signal sequences to base composition asymmetries in bacterial genomes. *Journal of Molecular Evolution*, **50**:249-257.

183. Valenzuela, C.Y. (1997) Non random DNA evolution. *Biology Research*, **30**:117-123.

184. Vologodskii, A.V., Frank-Kamenetskii, M.D. (1982) Theoretical study of cruciform states in superhelical DNAs. *FEBS Letters*, **143**:257-260.

185. Wada, K.-N., Doi, H., Tanaka, S.-I., Wada, Y., Furusawa, M. (1993) A neo-darwinian algorithm: asymmetrical mutations due to semiconservative DNA-type replication promote evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **90**:11934-11938.

186. Watanabea, H., Gojobori, T., Miura, K.-I. (1997) Bacterial features in the genome of

Methanococcus jannaschii in terms of gene composition and biased base composition in ORFs and their surrounding regions. *Gene*, **205**:7-18.

187. Watson, J.D., Crick, F.H.C. (1953) A structure for deoxyribose nucleic acid. *Nature*, **171**:737-738.

188. Wilquet, V., Van de Casteele, M. (1999) The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. *Research in Microbiology*, **150**:21-32.

189. Wright, S. (1969) Evolution and the genetics of populations. Volume 2, The theory of gene frequencies. The University of Chicago Press. ISBN 226-91050-4.

190. Wu, C.-I. (1991). DNA strand asymmetry. *Nature*, **352**:114-114.

191. Wu, C.-I., Maeda, N. (1987) Inequality in mutation rates of the two strands of DNA. *Nature*, **327**:169-170

192. Zharkikh, A; (1994) Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, **39**:315-329.

193. Zeigler, D.R., Dean, D.H. (1990) Orientation of genes in the Bacillus subtilis chromosome. *Genetics*, **125**:703-708.

194. Zuker, M. (1989) Computer prediction of RNA secondary structure. *Methods in Enzymology*, **180**:262-289.

CURRICULUM VITÆ

Personal informations

Born:	July 1st, 1966
Place of birth:	Grenoble, Europe
Nationality:	French Citizen
Married:	Florence Le Coz Two childs (Clara & Thomas Lobry)
Social security ID:	1 66 07 38 185 002 47
Home adress	Le Bourg, F-01510 TALISSIEU Fax: +33 479 87 31 60
Lab adress	Laboratoire BBE-CNRS UMR 5558 University Lyon I 43 Bd 11-NOV-1918 F-69622 VILLEURBANNE CEDEX Phone +33 472 43 12 87 Fax: +33 478 89 27 19

Education

1983: High Scool Diploma

1986: B.S.

1988: M.S. (major)

1991: Ph.D, Lyon Univerity

Professional experience

1988-1991: Ph.D Fellow at bioMérieux Inc. and teatching assistant

1991-1992: Research Fellow at bioMérieux Inc.

1992-1997: Assist. Professor at the University of Lyon

1998-present: Assoc. Professor at the University of Lyon (2300 _ $month^{-1}$)

Computer experience

Languages: Java, C, Fortran, Pascal

Operating systems: UNIX, MacOS

Society and professional memberships

Editorial Board Member of *Applied & Environmental Microbiology* (1998-2001)

Research interests

Mathematical modeling of microbial growth and growth media optimization (1988-1992)

Molecular biometry, evolution and structure of genomes (1992-present)

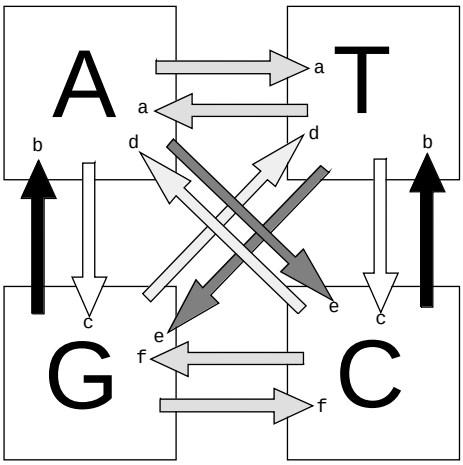
Honors

Philips Inc. prize for young scientists (1986)

Present research summary

Assuming that the evolution of DNA bases frequencies is governed by a symmetric process with respect to the two DNA strands, a nice wrong model is obtained. This model is nice because it can be rejected from the sole inspection of base frequencies in DNA and its rejection, which is effective in many genomes, means that the underlying non-observable process is asymmetric.

My current research interest is about the analysis and modeling of biological sequences to study evolution and genome features. The starting point was the study of a model of DNA base frequencies evolution under the simplifying assumption that there is no mutational or selective bias between the two strands of DNA. As compared with the general model of DNA evolution with 12 parameters, there are only 6 parameters left (as depicted below) so that the mathematical study of the model is simplified.



An interesting property of this model is that at equilibrium the intra-strand equalities [A]=[T] and [C]=[G] should be observed, regardless of substitution rates values. My present work is based on theses equalities, to test an alternative selective hypothesis that could explain them, to use them as a simplifying assumption, and to interpret deviations from them.

Thanks to the availability of complete bacterial genomes, I was able to describe a new genomic structure called *chirochore*. The term *chirochore* was coined to describe fragments

of the genome more or less homogeneous for the base composition biases. This is a purely descriptive term without reference to any mechanism, reminiscent of *isochore* for the description of DNA fragments with a homogeneous G+C content in some vertebrate chromosomes. On the other hand, the term *replichore* was introduced to designate in bacteria the two oppositely replicated halves of the chromosome between the origin and the terminus. The nice thing is that chirochore and replichore boundaries are the same in bacteria. This allowed for a simple method to predict the origin and terminus of replication in bacteria. For instance, in *Borrelia burgdorferi* the chirochore structure predicted that the origin of replication was at the center of the linear chromosome, and we found that experimental data were not in contradiction with this hypothesis. The chirochore structure, evidenced mainly with GC skew analyses, is now routinely used to predict replication boundaries in complete bacterial genome sequences. The chirochore structure is important to provide strong candidate replication origins to test in shuttle vector development, the lack of genetic tools being a research bottleneck.

Bibliography

Frank, A.C., Lobry, J.R. (2000) Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes, *Bioinformatics*, **16**:560-561.

Lobry, J.R. (1999) A nice wrong model for the evolution of DNA base frequencies. *Physica A*, **273**:100-103.

Lobry, J.R. (1999) Genomic landscapes. *Microbiology Today*, **26**:164-165.

Kano-Sueoka, T., Lobry, J.R., Sueoka, N. (1999) Intra-strand biases in bacteriophage T4 genome. *Gene*, **238**:59-64.

Frank, A.C., Lobry, J.R. (1999) Asymmetric substitution patterns : a review of possible underlying mutational or selective mechanisms. *Gene*, **238** :65-77.

Picardeau, M., Lobry, J.R., Hinnebusch, B.J. (1999) Physical mapping of an origin of bidirectional replication at the center of the *Borrelia burgdorferi* linear chromosome. *Molecular Microbiol*ogy, **32**:437-445.

Lobry, J.R., Lobry, C. (1999) Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Molecular Biology and Evolution*, **16**:719-723.

Charles, H., Mouchiroud, D., Lobry, J.R., Goncalves, I., Rahbe, Y. (1999) Gene size reduction in the bacterial aphid endosymbiont, Buchnera. *Molecular Biology and Evolution*, **16**:1820-1822.

Perrière, G., Lobry, J.R. (1998) Asymmetrical coding sequence repartition and codon adaptation index values between leading and lagging strands in seven bacterial species. *in* Proceedings of the first international conference on bioinformatics of genome regulation and structure (Novosibirsk, Russia, August 24-31, 1998) **2**:254-255.

Lobry, J.R. (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, **205**:309-316.

Galtier, N., Lobry, J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, **44**:632-636.

Perrière, G., Lobry, J.R., Thioulouse, J. (1996) Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acids sequences. *Computer Applications in the Biosciences*, **12**:519-524.

Lobry, J.R. (1996) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**:323-326.

Lobry, J.R. (1996) Origin of replication of *Mycoplasma genitalium*. *Science*, **262**:745-746.

Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, **13**:660-665.

Lobry, J.R. (1995) Properties of a general model of DNA evolution under no-strand bias conditions *Journal of Molecular Evolution*, **40**:326-330; **41**:680.

Thioulouse, J., Lobry, J.R. (1995) Co-inertia alaysis of amino-acid physico-chemical properties and protein composition with the ADE package. *Computer Applications in the Biosciences*, **11**:321-329.

Rosso, L., Lobry, J.R., Bajard, S., Flandrois, J.P. (1995) Convenient Model to Describe the Combined Effects of Temperature and pH on Microbial Growth. *Applied and Environmental Microbiology*, **61**:610-616.

Lobry, J.R. (1995) Unexpected behaviour of Monod's bacterial growth model. pp 149-154 *in* Mathematical population dynamic: analysis of heterogeneity (Arino, O., Axelrod, D., Kimmel, M. *eds*). Wuerz, Winnipeg.

Lobry, J.R., Gautier, C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research*, **22**:3174-3180.

Rosso, L., Lobry, J.R., Flandrois, J.-P. (1993) An Unexpected Correlation between Cardinal Temperatures of Microbial Growth Highlighted by a New Model. *Journal of Theoretical Biology*, 162:447-463.

Lobry, J.R., Carret, G., Flandrois, J.-P. (1992) Maintenance requirements of Escherichia coli ATCC 25922 in the presence of sub-inhibitory concentrations of various antibiotics, *Journal of Antimicrobial Chemotherapy*, **29**:121-127.

Lobry, J.R., Flandrois, J.-P., Carret, G., Pavé, A. (1992) Monod's bacterial growth model revisited. *Bulletin of Mathematical Biology*, **54**:117-122.

Carret, G., Flandrois, J.-P., Lobry, J.R. (1991) Biphasic kinetics of bacterial killing by quinolones. *Journal of Antimicrobial Chemotherapy*, **27**:319-327.

Lobry, J.R., Flandrois, J.-P. (1991) Comparison of Estimates of Monod's growth model from the same data set. *Binary*, **3**:20-23.

Lobry, J.R., Rosso, L., Flandrois, J.P. (1991) A FORTRAN subroutine for the Determination of Parameter Confidence Limits in Non-linear Models. *Binary*, **3**:86-93.