CHAPTER 1

ENVIRONMENTAL AND PHYSIOLOGICAL INSIGHTS FROM MICROBIAL GENOME SEQUENCES

Alessandra Carbone and Anthony Mathelier
Génomique Analytique, Université Pierre et Marie Curie-Paris 6, UMR S511, 91, bd de l'Hôpital, 75013 Paris, France
Alessandra.Carbone@lip6.fr and Anthony.Mathelier@gmail.com

   Facts and ideas presented in this short review are addressed to those computer scientists and mathematicians that want to learn about some open questions on the bioinformatics of microbial organisms. We present some recent results of our group on the statistical analysis of microbial genomes involving the formalization of microbial spaces, metabolic network comparison, minimal gene sets, host-phage adaptation and gene chromosomal organization. The guideline to all results presented here is to derive insights on microbial physiology and habitat directly from genome sequences by means of a purely statistical analysis and an appropriate design of algorithms.

1.1   SOME BACKGROUND, MOTIVATION AND OPEN QUESTIONS

Life mechanisms have been addressed in microbiology with a model-organism approach that has dominated the history of biology of the past century. Today, the large availability of genomic data coming from genome sequencing and from new high-throughput technologies changed radically the nature of investigations that are envisaged in molecular biology. For the first time, an organism and its changing environment can be considered together and in concert with a multitude of other organisms. Evolutionary hypothesis start to be tested at large scale.

Even though the experimental power is present, our understanding of microbe-environment interactions is still in its infancy, and upcoming discoveries and advances in geomicrobiology are likely to come from all areas of this discipline. This is a particularly good time for experimentalists to interact together with bioinformaticians, biomathematicians and biophysicists for defining new experimental questions, analyze large amount of data, model phenomena that characterize the new dimension within which microbiology lives today.

Microorganisms, genome sequences and comparative genomics. The availability of a large number of genomes provides the possibility to study biodiversity between species, within species and even within strains by comparing what is missing and what is common at a genetic level. Satisfactory formal methods and models to study similarities within the diversity have been developed but methods to study differences within a diversity are much harder to envisage. On the experimental level we experience a similar situation. Interaction and back-and-forth between experimentalists and bioinformaticiens/modelers which can direct both experiments and modeling is demanded in such a setting.

Evolutionary questions concerning chromosomal integrity, genomic expansion, chromosomal rearrangements, mobility of genetic elements, lateral gene transfer, gene share and gene loss among strains and across species, gene creation, evolution of duplicated genes, modular gene organization, gene essentiality and chromosomal organization, sexual proliferation are at the heart of microbiology today and are questions that we need to address. In fact, any question concerning the way that non-coding information in genomes, genes and genome architecture influence and direct the biology of the microbe, such as its virulence or adaptiveness under specific environmental conditions, constitutes a major preoccupation in the field of microbiology. Adaptive and non-adaptive processes in a cell, including genomic adaptiveness (for bacteria, viruses and eventually eukaryotic unicellular organisms), networks adjustment under changes in environmental conditions, modular rearrangements in genome architecture are other key questions for microbiology. Together with this, the understanding of structural and functional relations in proteins, the fate of duplicated protein sequences, speciation via divergent evolution of duplicated genes, are also key concerns.

Microorganisms and their organisation. What information can we extract from genomes concerning the biology of the organism? Statistical analysis has lead to identify statistical conditions (purely based on composition and on no biological in-

formation about the life of the organism) to determine the optimal growth temperature of a bacteria, to organize bacteria with respect to their ecological niche, to determine which genes are essential for the life of a unicellular organism, to determine essential metabolic networks etc. These results are intimately connected to the evolutionary processes that the organism has undergone and any insight on the origins of evolutionary pressures are most useful for a correct understanding of the evolutionary history of the organism. It is this history that justifies the today life of the organism.

Several types of measures, that induce organismal classifications might be introduced. Some of these measures try to detect environmental organization, some other physiological similarities and others phylogenetic proximity. Differences between phylogenetic, physiological and environmental information lead to different classifications of the microbial world. Such organizational differences are bound to be important for the understanding of evolution. In this respect, it might be that the understanding of the metagenomic data (that is, DNA data coming from environmental genomic analysis, corresponding to multispecies communities of organisms which usually either they have not being attempted to culture or they have been resistant to culturing efforts) will be possible only if these three conceptually different classification paradigms for organisms will be cleared out.

Another important paradigm for classification that is rising in these last years is induced by chromosomal rearrangements. It has been observed that it does not correspond to phylogenetic organization. What sort of biological information is associated to genome reorganization? Is it dependent on environmental pressure? or perhaps on physiological constraints?

Microorganisms, metagenomics and comparative genomics. In recent years it became clear that comparative genomics will include soon metagenomic data and metagenomic reconstruction of partial metabolic information for different ecological niches. This new incoming information will no doubt bring a fresh view on more classical model organisms. The study of the relations microbe-environment demands to take into consideration the fact that the 99% of the microorganisms visualized microscopically in environmental samples are not cultivated by routine techniques. This reality underlies the difficulty to make sound ecological inferences based on metabolic properties of a few cultivable species.

Even if this point should be addressed, we need to keep in mind that cultivable species propose a setting where environmental conditions and changes in environmental conditions can be tested and where the behavior of microbial populations can be modeled and analyzed. The *hemiascomycetes* (to which bakery yeast belongs) for instance, constitute a group of species which are very useful for learning because characterized by a compact genome, by very different ecological niches, and by an easy experimental handling of several of the species.

Microorganisms and population genetics. The fundamental principle underpinning microbial population dynamics is that the survival of a given individual microorganism is ultimately dependent upon the metabolic activity of others in its ecosystem. In this respect, the inclusion of the view of population genetics within the approaches to

genome analysis for understanding diversity of microbial communities and cultures, seems necessary. The interaction microbe-environment might turn out to be entangled with population processes in microbial species, such as natural selection, demography and migrations, and the population genetics perspectives might turn out to provide the conceptual tools for this understanding.

Microorganisms, metabolism and environment. Micro-organisms are intimately involved in transforming inorganic and organic compounds to meet their nutritional and energetic needs. Because the metabolic waste from one type of species nearly always provides substrate for another, there is an interdependence between species growing in close proximity to one another, or alternatively, the communities can be spatially separated, and elemental cycling may take on more complex and convoluted pathways. The power of microbial communities is fundamental for life and today, experiments can be conceived to meaningfully study microbial communities at different scales.

Questions concerning networks response to environmental changes affecting microbial eukaryotic and/or prokaryotic communities might be addressed. The functioning of regulation networks will be understood in terms of population biology and interactions (competition or collaboration) among species.

Adaptiveness is another important phenomena concerning microbial organisms interacting with their environment. It might be addressed *in silico* within the context of genetic variability for bacterial/viral species/strains based on available data. Questions around genetic exchanges within species and across species might be also ground for interdisciplinary work.

In this paper, we shall look at microbial organisms with available complete genomic sequences and demonstrate that we can read evolution signals out of them and derive meaningful biological information about an organism. The approach is not comparative. We start from a genome, realize a statistical analysis of its genes and derive insights on the biology of the organism guided by statistical biases of codon usage. We aim to find a general method that can be applied to organisms whose genome is known but for which not much biological information is available. One of the main motivations for this work is to search for a pool of genes that are *essential* for an organism. This question is fundamental if we think of synthesizing a genome from scratch and of attaining genome minimization conditioned by specific environmental conditions and metabolic activities (Venter et al. 2003; Zimmer 2003; Smith et al. 2003).

1.2   A FIRST STATISTICAL GLIMPSE TO GENOMIC SEQUENCES

Proteins are formed out of 20 amino-acids which are coded in triplets of nucleotides, called codons. The four nucleotides (A, T, C, G) define 64 codons used in the cell. Codons are not uniformly employed in the cell, but at the contrary, certain codons are

preferred and we speak about *codon bias*. There are several kinds of codon biases and some of them are linked to specific biological functions. Statistical analysis of DNA sequences and in particular of codon bias were performed from the moment that long chunks of DNA sequences were publicly available in the early eighties (Grantham et al. 1980; Wada et al. 1990), and the roots for these studies can be traced back to the sixties (Sueoka 1962; Zuckerkandl and Pauling 1965). However with the increasing number of bacterial genome sequences from a broad diversity of species, this field of research has been revivified in the last few years (Koonin and Galperin 1997; Lin and Gerstein 2000; Radomski and Slonimski 2001; Knight et al. 2001; Sicheritz-Pont´en and Andersson 2001; Daubin et al. 2002; Lin et al. 2002; Lobry and Chessel 2003; Sandberg et al. 2003; Jansen et al. 2003).

Biased codon usage may result from a diversity of factors: GC-content, preference for codons with G or C at the third nucleotide position (Lafay et al. 1999), a leading strand richer in $G + T$ than a lagging strand (Lafay et al. 1999), horizontal gene transfer which induces chromosome segments of unusual base composition (Moszer et al. 1999), and in particular, translational bias which has been frequently noticed in fast growing prokaryotes and eukaryotes (Sharp and Li 1987; Sharp et al. 1986; M´edigue et al. 1991; Shields and Sharp 1987; Sharp et al. 1988; Stenico et al. 1994). Three main facts support the idea of "translational impact": highly expressed genes tend to use only a limited number of codons and display a high codon bias (Grantham et al. 1980; Sharp and Li 1987), preferred codons and isoacceptor tRNA content exhibit a strong positive correlation (Ikemura 1985; Bennetzen and Hall 1982; Bulmer 1987; Gouy and Gautier 1982), and tRNA isoacceptor pools affect the rate of polypeptide chain elongation (Varenne et al. 1984; Buckingham and Grosjean 1986).

To study the effect of translational bias on gene expression, Sharp & Li (Sharp and Li 1987) proposed to associate to each gene of a given genome a numerical value, called *Codon Adaptation Index or CAI* for short, which expresses its synonymous codon bias (see appendix for the definition). The idea is to compute a weight (representing relative adaptiveness) for each codon from its frequency within a chosen small pool of highly expressed genes S, and combine these weights to define the $CAI(g)$ value of each gene g in the genome. For Sharp et al., the hypothesis driving the choice of S is that, for certain organisms, highly expressed genes in the cell have highest codon bias, and these genes, made out of frequent codons, are representative for the bias. Based on this rationale, one can select a pool of ribosomal proteins, elongation factors, proteins involved in glycolysis, possibly histone proteins (in eukaryotes) and outer membrane proteins (in prokaryotes) or other selections from known highly expressed genes, to form the representative set S. Then, CAI values are computed and are checked to be compatible with genes known to be highly or lowly expressed in the cell. If this is the case, then predictions are drawn with some confidence on expression levels for genes and open reading frames, even with no known homologues. Even if conceptually clear, this framework has been misused several times in the literature and incorrect biological consequences have been derived for gene expression levels of organisms which do not display a dominant translational bias, as discussed in (Grocock and Sharp 2002). This confusion motivated us to search for a

methodology based on a precise mathematical formulation of the problem to detect the existence of translational bias.

But the main motivation for us came from the recognition that an increasing number of genome sequences will be available for organisms for which biological knowledge consists merely of a sketched morphological and ecological description. For these organisms, it might not be evident how to define the reference set S, nor how to identify a reliable testing set which can ensure that predictions meet a satisfiable confidence level. Still, one would like to detect if translational bias holds for these genomes and if so, to predict their gene expression levels. If not, one would like to know the origin of their dominating bias and use this information for genome comparison.

1.3   AN AUTOMATIC DETECTION OF CODON BIAS IN GENES

We proposed a simple algorithm to detect dominating synonymous codon usage bias in genomes (Carbone et al. 2003). The algorithm is based on a precise mathematical formulation of the problem that leads to use the Self-Consistent Codon Index ($SCCI$) (strongly correlated to the CAI measure in translationally biased organisms) as a *universal* measure of codon bias, that is a measure for biases of possibly different origins (and not only for translational bias, as CAI was originally introduced for). The formal definitions of $SCCI$ and CAI are given in the Appendix.

The idea of the algorithm is simple. It is an iterative algorithm that at iteration $i+1$ computes codon weights based on a set S of genes selected at iteration i, then ranks all genes with respect to their $SCCI$ value and selects a new set S, which has half the cardinality of the set determined at iteration i (if at the i-th iteration the selected set is already constituted by the 1% of all genes, then the new set will also be constituted by 1% of genes) and whose genes score the highest. The process is repeated until 1% of genes have been selected and convergence is reached. At the start, S is the set of all genes.

With the set of coding sequences as a sole source of biological information, the algorithm provides a reference set S of genes which is highly representative of the dominant codon bias. This set is used to compute the $SCCI$ of genes not only for organisms whose biology is well known but also for those whose functional annotation is *not* yet available. An important application concerns the detection of a reference set characterizing translational bias which is known to correlate to expression levels in many bacteria and small eukaryotes; it detects also leading-lagging strands bias, GC-content bias, GC3 bias, and horizontal gene transfer. In general, the algorithm becomes a key tool to predict gene expression levels and to compare species. The approach has been validated on 96 slow-growing and fast-growing bacteria and archaeal genomes, *Saccharomyces cerevisiae, Plasmodium falciparum, Caenorhabditis elegans* and *Drosophila melanogaster*.

1.4 GENOMIC SIGNATURES AND A SPACE OF GENOMES FOR GENOME COMPARISON

Based on this analysis, we propose a novel formal framework to interpret genomic relationships derived from entire genome sequences rather than individual loci. This space allows to analyze sets of organisms related by a common *codon bias signature* (at times, more than one kind of bias influences the same genomic sequence and the ensemble of these overlapped biases defines what we call the *signature* of a genome) (Carbone et al. 2004). We give a number of numerical criteria to infer content bias, translational bias and strand bias for genome sequences. We show in a uniform framework that genomes of quite different phylogenetic relationship share similar codon bias; other genomes grouped together by various phylogenetic methods, appear to be subdivided in finer subgroups sharing different codon bias characteristics; Archaea and Eubacteria share the same codon preferences when $AT3$ or $GC3$ bias is their dominant bias; archaeal genomes satisfying translational bias use a sharply distinguished set of preferred codons than bacterial genomes. Our analysis, based on 96 eubacterial and archaeal genomes, opens the possibility that this space might reflect the geometry of a prokaryotic "physiology space". If this turns out to be the case, the combination of the upcoming sequencing of entire genomes and the detection of codon bias signatures will become a valuable tool to infer information on the physiology, ecology and possibly on the ecological conditions under which bacterial and archaeal organisms evolved. For many organisms, this information would be impossible to be detected otherwise. More recently, our algorithm has been applied to more than 300 genomes and our hypothesis of environmental signature has been supported at larger scale (Willenbrock et al. 2006).

Spaces for environmental and physiological classification represent a bacterial classification alternative to phylogeny and they are closer to the living conditions of the organism. With a growing number of genomic data available, it becomes more and more important to have new alternative organizational schemes to understand bacterial populations and the biology of single organisms within their living environment. The algorithmic idea working for bacteria should be revisited for metagenomic sequences for instance and adapted for viral genomes. On such spaces, hypotheses such as adaptability of a virus to the codon bias of its host can be checked and preliminary analysis support this hypothesis (see later).

1.5 STUDY OF METABOLIC NETWORKS THROUGH SEQUENCE ANALYSIS AND TRANSCRIPTOMIC DATA

Genes with high codon bias describe in meaningful ways the biological characteristics of the organism and are representative of specific metabolic usage (Carbone and Madden 2005). In silico methods exploiting this basic principle are expected to become important in learning about the lifestyle of an organism and explain its evolution in the wild. We demonstrate that besides high expressivity during fast growth

or glycolytic activities which have been very often reported, the necessity for survival under specific biological conditions has its traces in the genetic coding (Carbone and Madden 2005). This observation opens the possibility to predict rare but necessary metabolic activities from genome analysis.

High expression of certain classes of genes, like those constituting the translational machinery or those involved in glycolysis, are correlated particularly well in the case of fast growing organisms. By shifting the paradigm towards metabolic pathways, we notice that several energy metabolism pathways are correlated with high codon bias in organisms known to be driven by very different physiologies, which are not necessarily fast growing and whose genomes might be very homogeneous. More generally, we derive a classification of metabolic pathways induced by codon analysis, show that genetic coding for different organisms is tuned on specific pathways and that this is a universal fact. The codon composition of enzymes involved in glycolysis for instance, often required to be rapidly translated, is highly biased by dominant codon composition across species (this is indicated by the high CAI value of these enzymes). In fast growers, the numerical evidence is definitely far more striking than for other organisms (that is, the absolute difference between the CAI value of these enzymes and the average CAI value for genes in the genome is "large"), but even for *Helicobacter pylori*, a genome of rather homogeneous codon composition, enzymes involved in glycolytic pathways happen to be biased above average. In the same manner, one detects the crucial role of photosynthetic pathways for *Synechocystis* or of methane metabolism for *Methanobacterium*.

mRNA transcriptional levels collected during the *Saccharomices cerevisiae* cell cycle under diauxic shift (deRisi et al. 1997) (here, glucose quantities decrease in the media during cell cycle and yeast goes from fermentation to aerobic respiration), have been used to analyze the yeast metabolic network in a similar spirit as done with codon analysis. A classification of metabolic pathways based on transcriptomic data has been proposed, and we show that the metabolic classification obtained through codon analysis essentially "coincides" with the one based on (a large and differentiated pool of) transcriptomic data. Such a result opens the way to explaining evolutionary pressure and natural selection for organisms grown in the wild, and hopefully, to explain metabolism for slow-growing bacteria, as well as to suggest best conditions of growth in the laboratory.

It is open the question of whether this kind of analysis can contribute to reconstruct metabolic information from metagenomics data.

1.6   FROM GENOME SEQUENCES TO GENOME SYNTHESIS: MINIMAL GENE SETS AND ESSENTIAL GENES

The aim on creating a synthetic genome that, when inserted into a cell, can live and replicate, possibly producing clean energy or curbing global warming (Venter 2003), recently increased the interest on the fundamental question of determining which genes are essential to a microbe.

Computational and experimental attempts tried to characterize a universal core of genes representing the minimal set of functional needs for an organism. Based on the increasing number of available complete genomes, comparative genomics (Mushegian and Koonin 1996, Makarova et al. 2003, Nesbø et al. 2001, Harris et al. 2003, Brown et al. 2001, Koonin 2003, Charlebois and Doolittle 2004) has concluded that the universal core contains less than 50 genes. In contrast, experiments (Itaya 1995, Kobayashi et al. 2003, Hutchison et al. 1999, Glass et al. 2006, Akerley et al. 2002, Gerdes et al. 2003, Hashimoto et al. 2005, Salama et al. 2004, Ji et al. 2001, Forsyth et al. 2002, Thanassi et al. 2002, Winzeler et al. 1999, Giavier et al. 2002, Kamath et al. 2003) suggest a much large set of essential genes (certainly more than several hundreds, even under the most restrictive hypotheses) which is dependent on the biological complexity and the environmental specificity of the organism. Highly biased genes, which are generally also the most expressed in translationally biased organisms, tend to be over-represented in the class of genes deemed to be essential for any given bacterial species. Also, all functional classes are represented by highly biased genes and within different species, highly biased genes with the same functional role need not be homologous. This association between highly biased genes and essential genes is far from perfect, nevertheless it allows to propose a new computational method based on $SCCI$ to detect to a certain extent ubiquitous genes, non-orthologous genes, environment specific genes, genes involved in stress response and genes with no identified function but highly likely to be essential for the cell. Most of these groups of genes cannot be identified with previously attempted computational and experimental approaches. Notice, for instance, that comparative genomics infers conclusions only for homologous genes and that certain non-homologous highly biased genes could not be identified by this approach. Also, experiments are run under optimal living conditions for the organisms and stress response genes cannot be identified by experiments. The large spread of lifestyles and the unusually detectable functional signals characterizing translationally biased organisms suggest to use them as reference organisms to infer essentiality in other microbial species. In (Carbone 2006), we analyse in detail 27 organisms belonging to a large variety of phylogenetic taxa, γ and δ proteobacteria, firmicutes, actinobacteria, thermococcales and methanosarcinales; they do not display strong GC nor AT content and they are characterized by different optimal growth temperatures (Carbone et al. 2004). We also discuss the case of small parasitic genomes, and data issued by the analysis are compared to previous computational and experimental studies.

1.7   A CHROMOSOMAL ORGANIZATION OF ESSENTIAL GENES

Patterns in chromosomal locations of essential genes have been examined and large scale features of bacterial chromosomes were derived (Mathelier and Carbone, manuscript, 2008). We wanted to check whether essential genes are organized in regularly spaced groups within the genome, possibly depending on transcription regulation patterns or on common functional activities of genes in the groups. Both these possibilities explaining the distribution of genes as a product of structural periodicity are

attractive. The localization of certain essential genes along structural chromosomal "faces" would have the advantage of creating spatial subregions in which essential genes could be accessed by limited diffusion of RNA polymerase or RNA polymerase fixed in factories. The solenoid model (Képès and Vaillant 2003; Képès 2004) and the rosettes model of chromosomes have been proposed as possible functional and spatial organizations of the chromosome. The idea behind these models is to bring close in space different genes through an encoded three dimensional genomic organization. The solenoid model organizes loops of DNA along a solenoidal three-dimensional arrangement and the rosettes model organizes DNA loops radially in a flower-like three-dimensional structure.

(Képès 2004) has shown that groups of genes regulated by the same transcription factors in *Escherichia coli* reveal chromosomal periodicity, and (Wright et al. 2007) shows that evolutionarily conserved gene pairs in *E. coli* also reveal chromosomal periodicity. We considered the pool of core genes detected by the methodology described above and checked whether these genes are periodically spaced or not. Genomic core's genes have to be either highly expressed or rapidly expressed and we wanted to test the hypothesis that a structured organization of their regulatory elements could help to reach fast expression. We studied at large scale the chromosomal organization of some tens of bacterial and archaeal organisms and find that most of these genomes present periodic distribution of their core genes along the leading strand (Mathelier and Carbone, manuscript, 2008). This property is not proved to hold for all microbial genomes but, still, it generates important questions around the impact of environmental pressures, selective bias, gene rearrangement constraints in microbes. We observe that a genome might display several significant periods on the different strands and that the amplitude of the signal can vary considerably from strand to strand and from organism to organism. We computed a period of about 34kb between essential genes on the leading strand of *E. coli* with a very pronounced amplitude of the signal. The same amplitude does not appear for essential genes located over the lagging strand. This seems to indicate that a chromosomal organization is hunted to help the expression of essential genes within the leading strand. We also observed that functional grouping of core genes explains chromosomal periodicity better than shared transcription regulators.

Periods computation is based on a signal processing parameterized model and a Fourier Transform analysis. Significance of the periods is established by comparing the amplitude of this signal with a random model by generating appropriate random genomes (with the same number of genes and the same distribution of distances between pairs of adjacent genes). Further investigations on the impact of structural organization on transcription mechanisms of bacterial organisms need to be addressed.

1.8   VIRAL ADAPTATION TO MICROBIAL HOSTS AND VIRAL ESSENTIAL GENES

The notion of SCCI and the algorithmic approach used to study bacterial species have been recently used to analyze viral genomes and adaptation to their host (Carbone

2008). Size and diversity of bacteriophage population asks for methodologies to quantitatively study the landscape of phage differences. Statistical approaches are confronted with small genome sizes forbidding significant single phage analysis and comparative methods analyzing full phage genomes represent an alternative of difficult interpretation due to Lateral Gene Transfer which creates a mosaic spectrum of related phage species. Based on a large scale codon bias analysis of 116 DNA phages hosted by 11 translationally biased bacteria belonging to different phylogenetic families we observe that phage genomes are almost always under codon selective pressure imposed by translationally biased hosts and we propose a classification of phages with translationally biased hosts which is based on adaptation patterns.

The codon bias measure used in the analysis is the SCCI. Namely, SCCI values reflect codon composition of phage genes relative to host codon composition and provide a numerical index of the advantage taken by phage genes once translated in the host environment. This advantage is expected to be higher when phage gene codon composition is biased toward host codon composition. Through our computational method based on SCCI, we compare phages sharing homologous proteins, possibly accepted by different hosts, and observe that throughout phages, independently from the host, capsid genes appear to be the most affected by host translational bias. For coliphages, genes involved in virion morphogenesis, host interaction and ssDNA binding are also affected by adaptive pressure. If phage genomes were to contain a pool of essential genes, these functional classes could suggest appropriate candidate genes. Adaptation affects in a significant way long and small phages. We analyze in more details the *Microviridae* phage space to illustrate the potentiality of the approach. Surprisingly, we can reconstruct the phylogenetic tree of the large phage pool defined around phage $\phi X 174$ (Rokyta et al. 2006) using exclusively codon bias information. Also, the adaptation analysis of the set of *Microviridae* phages defined around phage $\phi M H 2 K$ shows that phage classification based on adaptation does not reflect bacterial phylogeny. This result highlights that adaptation patterns in phages might be profitably used to unravel the intricate mosaic of phage speciation.

The numerical finding provided by this and future studies of phage-host coevolution will hopefully be useful in clarifying the role of phages as therapeutic agents against bacteria (Summers 2001) and in organizing metagenomic data.

Appendix

Some comments on the mathematical methods

In this text, a coding sequence is represented by a 64-dimensional vector, whose entries correspond to the 64 relative codon frequencies in the sequence. Recall that the frequency of a codon i in a sequence g is the number of occurrences of i in g (where g is intended to be split in consecutive non-overlapping triplets corresponding to amino-acid decomposition), and that the *relative frequency* of i in g is the frequency of i in g divided by the number of codons in g. For each vector representing a coding sequence, the sum of its entries must equal 1. Hence, a coding sequence is a point

in the 64-dimensional space $[0 \cdots 1]^{64}$, where no special assumption is made on the space nor on the coordinate system.

For each genome sequence G and some set of coding sequences S in G, *codon bias* is measured with respect to its synonymous codon usage. Given an amino-acid j, its synonymous codons might have different frequencies in S; if $x_{i,j}$ is the number of times that the codon i for the amino-acid j occurs in S, then one associates to i a *weight $w_{i,j}$* relative to its sibling of maximal frequency y_i in $\mathcal{S}$

$$w_{i,j} = \frac{x_{i,j}}{y_j}.$$

A codon with maximal frequency in S is called preferred among its sibling codons. *Self-Consistent Codon Index* $(SCCI)$ associated to g in G, is a value in $[0,1]$, defined as

$$SCCI(g) = (\Pi_{k=1}^{L} w_k)^{1/L}$$

where L is the number of codons in the gene, and w_k is the weight of the k-th codon gene sequence. Genes with $SCCI$ value close to 1 are made by highly frequent codons.

When the reference set S is predefined to be a set of highly expressed genes in the organism, then the index issued by the $SCCI$ formula corresponds to the known *Codon Adaptation Index* introduced by Sharp & Li (Sharp and Li 1987). The computation of the reference set S in the definition of $SCCI$ is based on a pure statistical analysis of all genes in a genome and it does not rely on biological knowledge of the organism. This allows us to compute weights for organisms of unknown lifestyle.

The name $SCCI$ was employed for the first time by (Carbone 2006), while in (Carbone et al. 2003, Carbone et al. 2005) the notion is called CAI, even though it does not exclusively refer to codon adaptation. Notice that CAI is always employed with a manual and explicit choice of S, while the formula $SCCI$ (i.e., CAI parameterized with S) turns out to be a universal measure to study codon bias. Codon weights, reference set S, and $SCCI$ values are calculated with the program CAIJava (Carbone et al. 2003), available at www.ihes.fr/∼carbone/data.htm.

All results cited in this review are obtained using very simple mathematical and algorithmic notions which are fully described in (Carbone et al. 2003; Carbone et al. 2004; Carbone and Madden 2005). The statistical analysis and numerical thresholds we propose are realized in a 64-dimensional codon space. Multivariance statistical methods have been employed as visualisation tools, but none of the formal results nor the biological conclusions are inferred from the 3 dimensional projections. Both space of genes and space of organisms in 64 dimensions, and distances between organisms are defined as ℓ_1-distances.

Complete genomes available

In June 2008, 2623 viruses (of which 495 are phages), 56 eukaryots, 53 archaea, 729 bacteria are completely sequenced and present in the NCBI database at the address http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome.

References

1. Akerley, B.J., Rubin, E.J., Novick, V.L., Amaya, K., Judson, N., Mekalanos, J.J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proceedings of the National Academy of Sciences USA*, **99**, 966–971.

2. Bennetzen, J.L., Hall, B.D. (1982) Codon selection in Yeast. *Journal of Biological Chemistry*, **257**, 3026-3031.

3. Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J.(2001) Universal trees based on large combined protein sequence data sets. *Nature Genetics*, **28**, 281–285.

4. Buckingham, R.H., Grosjean, H. (1986) The accuracy of mRNA-tRNA recognition. In : *Accuracy in molecular processes: its control and relevance to living systems*, ed. T.B.L. Kirkwood, R. Rosenberger and D.J. Galas, Chapman & Hall Publishers, London, 83-126.

5. Bulmer, M. (1987) Coevolution of codon usage and transfer RNA abundance. *Nature*, **325**, 728-730.

6. Carbone, A., Zinovyev, A. and Képès, F. (2003) Codon Adaptation Index as a measure dominating codon bias. *Bioinformatics*, **19**, 2005-2015.

7. Carbone, A., Képès, F. and Zinovyev, A. (2004) Microbial codon bias and the organisation microorganisms in codon space. *Molecular Biology and Evolution*, 22(3):547–561.

8. Carbone, A., Madden, D. (2005) Insights on the evolution of metabolic networks from data and sequence analysis. *Journal of Molecular Evolution*, 61:456–469.

9. Carbone,A. (2005) Revisiting the codon adaptation index from a whole-genome perspective: gene expression, codon bias, and metabolic networks in the context of genomes

comparison", *Handelingen van de Koninklijke Vlaamse Academie van Belgie voor Wetenschappen en Kunsten*, Proceedings of the Belgian Royal Academy of Sciences and Arts, Study Day on Genomics, 18 October 2003.

10. Carbone, A. (2006) Computational prediction of genomic functional cores specific to different microbes, *Journal of Molecular Evolution*, 63(6):733-746, 2006.

11. Carbone, A. (2008) Codon bias is a major factor explaining phage evolution in translationally biased hosts, *Journal of Molecular Evolution*, 2008. Online Feb 20.

12. Charlebois, R.L., Doolittle, W.F. (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Research*, **14**, 2469–2477.

13. Daubin, V., Gouy, M., Perrière, G. (2002) A phylogenetic approach to bacterial evidence of a core of genes sharing a comon history. *Genome Research*, **12**, 1080-

14. DeRisi, J.L., Iyer, V.R., Brown, P.O. (1997) Exploring the metabolic and genetic control expression on a genomic scale. *Science*, **278**, 680-686.

15. Forsyth, R.A. *et al.* (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol Microbiol.* **43**, 1387–1400.

16. Gerdes, S.Y., Scholle, M.D., Campbell, J.M., Balazsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., Bhattacharya, A. *et al.*(2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of Bacteriology*, **185**, 5673–5684.

17. Giaever, G. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.

18. Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison III, C.A., Smith, H.O., Venter, J.C. (2006) Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences USA*, **103**, 425—430.

19. Gouy, M. and Gautier, Ch. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10, 7055-7070.

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

21. Grocock, R.J., Sharp, P.M. (2002) Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene*, **289**, 131-139.

22. Harris, J.K., Kelley, S.T., Spiegelman, G.B., Pace, N.R. (2003) The genetic core of the universal ancestor. *Genome Research*, **13**, 407–412.

23. Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., Fujimitsu, K., Keyamura, K., Ote, T., Yamakawa, T., Yamazaki, Y., Mori, H., Katayama, T., Kato, J. (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol Microbiol*, **55**:137-49.

24. Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., Venter, J.C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, **286**, 2165–2169.

25. Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13-34.

26. Itaya, M. (1995) An estimation of the minimal genome size required for life. *FEBS Lett.*, **362**, 257–260.

27. Jansen, R., Bussemaker, H.J., Gerstein, M. (2003) Revisiting the codon adaptation index from a whole-genome prespective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Research*, **31**, 2242-2251.

28. Ji, Y. *et al.* (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science*, **293**, 2266–2269.

29. Kamath, R.S. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.

30. Képès, F., Vaillant, C. (2003) Transcription-Based Solenoidal Model of Chromosomes. *Complexus* **1**, 171–180.

31. Képès,F. (2004) Periodic Transcriptional Organization of the *E. coli* Genome. *Journal of Molecular Biology*. **340**, 957–964.

32. Knight, R.D., Freeland, S.J., Landweber, L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, **2**, at http://genomebiology.com/2001/2/4/research/0010.

33. Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P. *et al.* (2003), Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A*, **100**, 4678—4683.

34. Koonin, E.V. (2003) Comparative genomics, minimal gene sets and the last common ancestor. *Nature Reviews Microbiology*, **1**, 127–136.

35. Koonin, E.V., Galperin, M.Y. (1997) Prokaryotic genomes: The emerging paradigm of genomebased microbiology. *Curr. Opin. Genet. Dev.*, **7**, 757763.

36. Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M. and 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.

37. Lin, J., Gerstein, M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research*, **10**, 808-818.

38. Lin, J., Qian, D.,Bertone, P., Das, R., Echols, N., Senes, A., Stenger, B., Gerstein, M. (2002)GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucleic Acids Research*, **30**, 4574-4582.

39. Lobry, J.R., Chessel, D. (2003) Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J.Appl Genet*, **44**, 235-261.

40. Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov,R.L., Wolf, Y.I. and Koonin, E.V. (2003) Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Research*, **9**, 608–628.

41. Mathelier, A., Carbone, A. (2008) Chromosomal arrangement of essential genes in bacteria and archaea. In preparation.

42. Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. (1991) Evidence for horizontal gene transfer in Escherichia coli speciation. *Journal of Molecular Biology*, **222**, 851-856.

43. Moszer, I., Rocha, E.P.C., Danchin, A. (1999) Codon usage and lateral gene transfer in *Bacillus Subtilis. Current Opinion in Microbiology*, **2**, 524-528.

44. Mushegian, A.R. and Koonin, E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Science USA*, **93**, 10268–10273.

45. Nesbø, C.L., Boucher, Y., Doolittle W.F. (2001) Defining the core of non-transferable prokaryotic genes: the euryarchaeal core. *Journal of Molecular Evolution*, **53**, 340–350.

46. Radomski, J.P., Slonimski, P.P. (2001) Genomic style of proteins: concepts, methods and analysis of ribosomal proteins from 16 microbial species. *FEMS Microbiology Reviews*, **25**, 425-435.

47. Rokyta, D.R., Burch, C.L., Caudle, S.B., Wichman, H.A. (2006) Horizontal gene transfer and the evolution of microvirid coliphage genomes. *J Bacteriol*, **188**, 1134–1142.

48. Salama, N.R. *et al.* (2004) Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori. J. Bacteriol.*, **186**, 7926–7935.

49. Sandberg, R., Brändén, C.I., Ernberg, I., Cöster, J. (2003) Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino-acids usage and G+C content. *Gene*, **311**, 35-42.

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

51. Sharp, P.M., Tuohy, T.M.F., Mosurski, K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiate highly and lowly expressed genes. *Nucleic Acids Research*, **14**, 8207-8211.

52. Shields, D.C. and Sharp, P.M. (1987) Synonymous codon usage in Bacillus subtilis reflects both traditional selection and mutational biases. *Nucleic Acids Research*, **15**, 8023-8040.

53. Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F. (1988) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomices pombe, Drosophila melanogaster and Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Research*, **16**, 8207-8211.

54. Stenico, M., Loyd, A.T., Sharp, P.M. (1994) Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acid Research*, **22**, 2437-2446.

55. Sicheritz-Pontén, T. and Andersson, Siv G.E. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Research*, **29**, 545-552.

56. Smith, H.O., Hutchison, C.A. III, Pfannkoch, C., Venter, C. (2003) Generating a synthetic genome by whole genome assembly: AX174 bacteriophage from synthetic oligonucleotides. *Proc Natl Acad Sci USA*, **100**, 15440–15445.

57. Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad of Sci USA*, **48**, 582-592.

58. Summers, W.C. (2001) Bacteriophage therapy. *Annu Rev Microbiol*, **55**, 437–451.

59. Thanassi, J.A. *et al.* (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. *Nucleic Acids Res.*, **30**, 3152–3162.

60. Varenne, S., Buc, J., Lloubès, R. and Lazdunski, C. (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of Molecular Biology*, **180**, 549-576.

61. Venter, J.C., Levy, S., Stockwell, T., Remington, K., Halpern, A. (2003) A massive parallelism, randomness and genomic advances. *Nature Genetics*, **33**, 219–227.

62. Wada, K.S., Aota, R., Tsuchiya, F., Ishibashi, T., Gojobori, T. and Ikemura, T. (1990) Codon usage tabulated from GenBank genetic sequence data. *Nucleic Acids Research*, **18**(Suppl), 2367-2411.

63. Willenbrock, H., Friis, C., Friis, A.S., Ussery, D.W. (2006) An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biol* **7**:R114.

64. Winzeler, E.A. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901-–906.

65. Wright, M.A., Krarchenko, P., Church, G.M., Segré, D. (2007) Chromosomal periodicity of evolutionary conserved gene pairs, *Proc Natl Acad of Sci USA*, **104**, 10559–10564.

66. Zimmer, C. (2003) Genomics. Tinker, tailor: Can Venter stitch together a genome from scratch?, *Science* **299**, 1006–1007.

67. Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J Theor Biol*, **8**, 357-366.