REVISITING THE CODON ADAPTATION INDEX FROM A WHOLE-GENOME PERSPECTIVE: GENE EXPRESSION, CODON BIAS, AND METABOLIC NETWORKS IN THE CONTEXT OF GENOMES COMPARISON

Alessandra Carbone[1]

[1] *Laboratory of Analytical Genomics, Universite Pierre et Marie Curie, INSERM U511, 91, Bd de l'Hopital, 75013 Paris, France; e-mail: carbone@ihes.fr*

Abstract

Facts and ideas presented in this short review concern some recent developments at the interface between sequence analysis, gene expression prediction and genome comparison carried on in our group. The guiding line to all results presented here is to derive biological information from genome sequences by means of a purely statistical analysis and an appropriate design of algorithms.

1 Some background and motivation

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-Ponten 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; Medigue 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.

2 An automatic detection of codon bias

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 Codon Adaptation Index (CAI) as a *universal* measure of codon bias, that is a measure for biases of possibly different origins (and not only for translational bias, as it was originally introduced for). 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 bias. This set is then used to compute the Codon Adaptation Index of genes of prokaryotic and eukaryotic organisms, including 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, to guide regulatory circuit reconstruction, and to compare species. The approach is validated on 96 slow-growing and fast-growing bacteria and archaeal genomes, *Saccharomyces cerevisiae, Plasmodium falciparum, Caenorhabditis elegans* and *Drosophila melanogaster*.

3 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 analyse 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. 2003b). 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 dierent phylogenetic relationship share similar codon bias; other genomes grouped together by various phylogenetic methods, appear to be subdivided in finer subgroups sharing dierent 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. 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 2003c). 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 2003c). This observation opens the possibility to predict rare but necessary metabolic activities from genome analysis.

As discussed above, 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.

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 S

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

A codon with maximal frequency in S is called preferred among its sibling codons. *Codon Adaptation Index* (CAI) associated by Sharp & Li (Sharp and Li 1987) to g in G, is a value in $[0, 1]$, defined as

$$CAI(g) = (\Pi_{k=1}^{L} wk)1/L$$

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 CAI value close to 1 are made by highly frequent codons.

All results cited here are obtained using very simple mathematical and algorithmic which are fully described in (Carbone et al. 2003; Carbone et al. 2003b; Carbone Madden 2003c). The statistical analysis and numerical thresholds we propose are the 64-dimensional codon space. Multivariance statistical methods have been employed visualisation tools, but none of the formal results nor the biological conclusions 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.

References

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

[2] 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.

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

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

[5] Carbone, A., Képès, F. and Zinovyev, A. (2003) Microbial codon bias and the organisation microorganisms in codon space. Submitted.

[6] Carbone, A., Madden, D. (2003) Insights on the evolution of metabolic networks from data and sequence analysis. In preparation.

[7] 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-

[8] 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.

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

[10] 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.

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

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] Sandberg, R., Bränden, 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

[32] 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.

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