DNA Phonology: Investigating the Codon Space

Giuseppe Insana

Wolfson College

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

November 2003

European Molecular Biology Laboratory,
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SD,
United Kingdom.

Email: insana@ebi.ac.uk

Summary

The main part of the thesis is concerned with large-scale studies of codon usage in completely sequenced genomes. A new compositional analysis scheme is presented, complete with a number of computation and visualisation tools. The thesis addresses the benefits of this very general scheme, named *codon profiling*, with comparisons to the very similar *synonymous codon usage*. Codon profiling is applied to the analysis of several domains of interest, with the scope of addressing several questions related to the compositional constraints of coding sequences.

The heterogeneity of codon usage in the coding sequences of each genome was examined and presented, noting the consistency of intra-genomic distributions of codon similarity and atypicality. Such distributions provide the grounds on which to elaborate practical applications that make use of these properties.

A computationally inexpensive methodology was developed to detect Horizontal Gene Transfers (and for the first time to identify donor genomes), exploiting measures of codon similarity and combining a compositional identification approach with a phylogenetic verification process.

The thesis also presents a detailed procedure for the characterisation of coding sequences with atypical codon usages, exemplified in a study conducted on a group of human RNA binding proteins whose codon usage has striking similarity to that of some human infecting retroviruses.

Finally, the concept of *codon usage space*, the space of all the possible codon usages, is discussed. After calculating the theoretical extension of this space, the part visited by known biological sequences was mapped and its dimensionality computed. The comparison with the results obtained using several algorithms for random generation of codon usages quantifies the constraints imposed on biological sequences and allows the investigation and characterisation of the unexplored regions of the space.

Acknowledgements

The work of this thesis would not have been possible without contribution, support, supervision and friendship from **Heikki Lehväslaiho**, **Park Jong Hwa**, **Cheng Dong Seon** and **Elia Stupka**.

Additional guidance and assistance was granted with extreme precision, kindness and availability by Liisa Holm, Michael Ashburner, Peer Bork, Arek Kasprzyk, Arne Stabenau, Geoffrey Richardson, Nick Goldman, Gillian Adams, Adrian Friday and Paul Sharp. *Grazie!*

The research was developed at the European Bioinformatics Institute and was funded by a predoctoral studentship from the European Molecular Biology Laboratories. I feel sincere gratitude for the opportunity that these institutes granted me and for the splendid people I met in Hinxton, Cambridge and Heidelberg.

I wish to thank Jong, Elia, Kuang, Matthieu, Arne, Pat, Laurence, Mauro, Saiko and Ya-Hsuan for their special friendship in the Cambridge years. A special thought goes to all my Friends in Italy and around the world, and to my parents and grandparents: you are lifegivers.

Finally, I would like to express my gratitude to the open source developers, in particular to those noble people who provide Linux and Perl, resources that always guaranteed possibilities and efficiency to my projects. The work presented was completed without use of commercial software.

Declaration

In accordance with university regulations, I declare that this dissertation is the result of my own work and contains nothing that is the outcome of work done in collaboration, unless stated otherwise in the text. This thesis has been typeset in 12pt font and does not exceed the specified length limit of 300 pages according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Contents

Index of figures

Index of tables

Abbreviations

A: *Adenine*

bp: *base pairs (nucleotide count)*

C: *Cytosine*

CA: *Correspondence Analysis*

CDS: *coding sequence(s)*

CPRO: *Codon profile (analysis, vector)*

CSYN: *Synonymous codon usage (analysis, vector)*

DNA: *DeoxyriboNucleic Acid*

EBI: *European Bioinformatics Institute*

EMBL: *European Molecular Biology Laboratory*

G: *Guanine*

GC3: *Guanine-Cytosine content in the third coding position*

HGT: *Horizontal Gene Transfer*

MDS: *Multi Dimensional Scaling*

MVA: *MultiVariate Analysis*

NCBI: *National Center for Biotechnology Information*

RNA: *RiboNucleic Acid*

SCOP: *Structural Classification Of Proteins*

SOM: *Self-Organising Map*

T: *Thymine*

TCAG123: *Nucleotide contents in the three coding positions*

U: *Uridine*

WWW: *World Wide Web*

Additionally, one-letter and three-letter abbreviations for amino acids are often used:

A	Ala	Alanine	M	Met	Methionine
C	Cys	Cysteine	N	Asn	Asparagine
D	Asp	Aspartate	P	Pro	Proline
E	Glu	Glutamate	Q	Gln	Glutamine
F	Phe	Phenylalanine	R	Arg	Arginine
G	Gly	Glycine	S	Ser	Serine
H	His	Histidine	T	Thr	Threonine
I	Ile	Isoleucine	V	Val	Valine
K	Lys	Lysine	W	Trp	Tryptophan
L	Leu	Leucine	Y	Tyr	Tyrosine

I *Introduction, DNA linguistics and codon usage*

A OUTLINE OF THE DISSERTATION

The dissertation is organised in almost self-contained chapters, each with its own *Introduction*, *Methods*, *Results&Discussion* and *Conclusions* sections.

The present chapter first of all gives an introduction to the biology of genetic message encoding and translation and in particular to the codon information, with special emphasis on the redundancy of genetic messages and to how this redundancy can serve the superimposition of other messages. It then introduces a number of compositional analysis methods which are used by the scientific community to investigate the form of genetic messages.

The second chapter presents a newly developed framework, called *codon profiling*, for analysis of codon usage information. It combines traditional codon usage with nucleotidic composition analysis, thus adopting a genomic base-oriented perspective which is general, elegant and extensible. Because of its similarity with the *synonymous codon usage* analysis, the two methodologies are compared, showing the respective benefits. Furthermore, since all the work presented in this dissertation was conducted *in tandem* with both methodologies, the results obtained under the two frameworks, when different, will also be discussed in the other parts of this dissertation.

Chapter three deals with intra-genomic heterogeneity for the annotated completely sequenced genomes, assessing how diverse in codon usage the genes inside a genome are. Distributions of codon similarity are plotted for archaea, bacteria, five eukaryotic genomes and for human infecting viruses. The distributions are shown to have the same shape and spread, spanning across the same range of similarity values. The observed coherence in intra-genomic heterogeneity provided the scale and thresholds for codon similarity and codon atypicality, enabling various practical applications to be developed.

The fourth and fifth chapters present two such applications of codon profiling and codon similarity measures. Chapter four details a procedure elaborated to detect Horizontal Gene Transfer events by combination of a very fast compositional approach (based on codon similarity information) and of a slower phylogenetic approach for verification. Besides the multifaceted strategy, the advantage lies in the possibility of predicting and verifying the donor species. Chapter five describes a complete methodology for the identification and characterisation of genes with highly heterogeneous codon usage, exemplified by a real case analysis of human infecting viruses in the context of the human genome and of human protein families with very atypical codon usage.

The sixth and final chapter discusses the conceptual space of all possible codon usages. After calculating the number of theoretical possibilities, the attention is focused on understanding how many of these are really employed by the biological world (although in our limited approximation of it, represented by the sequenced data) and what the portions of non-populated space are. To deal with the enormous number of possibilities, the space is mapped at a certain specified granularity level, or, in other words, with a certain binning size grouping together similar codon usages. Several algorithms to generate random codon usages have been developed and used to sample the codon space. The heterogeneity of the generated codon usages is compared to the biological one, underlining and quantifying the constraints influencing the latter.

B MOTIVATION

Genetics studies the means by which biological information is transferred and how this information can change, giving rise to different organisms and different species: the wonderful process of evolution, intrinsically bound to our concept of life.

This biological information is contained (to the best of our knowledge) in the nucleic acid molecules, long strings of *bases*. We can think of them as long sequences of letters. These letters are: A T C G (actually the three-dimensional structure that a succession of bases assumes in the whole molecule can sometimes be more important, but we concentrate on the sequence, because from the sequence it should be possible to infer the structure). DNA is the name of the molecule responsible for this genetic information: *Deoxyribo Nucleic Acid*. DNA can be thought of as a language. It is the language in which all the information “for making a new organism” is written, the blueprint of a living being (specifically, of its structural and functional parts).

Understanding this language is a complex but fascinating goal – deciphering its phonetics (this was accomplished in the 1960's, a process that has been named *cracking of the Genetic Code*), its phonology, its syntax and morphology, its semiotics and semantics.

This thesis was born from a desire to explore the mechanisms behind (to continue with the metaphor) DNA phonology. In linguistics, phonology is the study of how sounds are used in a language, how they are combined, how they are perceived. For example, phonology studies constraints against particular combinations of sounds. Words like *druping* or *grink* are perceived as possible English words, even if they do not actually exist in the language. On the other hand, *kter* or *zlatrah* can definitely not be part of (present day) English. A representation of phonological constraint for English syllables could be:

(s) + (C) + (w|y|r|l) + (V) + V + (C) + (C) + (C)

where C=consonant, V=vowel, ()=optional; this translates into the letter s, followed by any consonant, followed by a consonant or semivowel among the set w y r or l, then a vowel, and so on and so forth. For example, consider how the word *strain* fulfils those constraints.

Are there *DNA phonological constraints* in biology? What are they? That is, what rules must the sequential array of bases obey? There are constraints to the form of the messages which are encoded in DNA. Constraints coming from the need to preserve a particular three dimensional structure, a particular composition of bases (*e.g.* more biased toward a lot of Gs and Cs or a lot of As and Ts), or a particular choice of frequent or rare "sounds" (which in the DNA domain would be the codons for abundant or rare tRNAs).

How much flexibility is there in the choice of codons? How many possible ways? Are all the arrangements possible and are they adopted by the genomes we study? Are the constraints different for different organisms? Are the genomes homogeneous with respect to codon usage? Is the amount of intra-genomic variability a constant or does it fluctuate widely? Can two species be identified by their choice of codons, like two human languages can be distinguished by the phonemes they use, their arrangement, their frequency? Can genes acquired from other species be recognised and their origin identified, in the same way as a borrowed word in the lexicon can be traced to the original language it was imported from?

New methodologies were devised and several experiments conducted towards the goal of addressing the above questions.

C Concepts

C.1 The genetic code and the translation of messages

Each group of three consecutive bases in a coding sequence is called a *codon* and corresponds to either an amino acid in a protein or to a signal that terminates translation. Codons that signal termination are called *stop codons*. The mapping from codons to amino acids is called *genetic code* (Table I-1). Most genomes use the same genetic code, called the *Standard genetic code*. There are in total 64 possible codons (four bases for three positions in the codon: 4^3). The genetic code was understood and completely described in the late 1960s.

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

Table I-1: *The Standard genetic code. * indicates terminator codon (STOP).*

C.1.1 Redundancy

In all genetic codes the sixty-four codons encode usually twenty (sometimes twenty-one or a few less than twenty) different amino acids and the termination signal. There is hence more than one codon encoding for the same amino acid, a property that takes the name of *redundancy* or *degeneracy* of the code. In the Standard code, for example, there are two amino acids encoded by a single codon, nine encoded by two codons, five by four codons, one amino acid encoded by three codons and three by six codons (plus three stop codons). Codons that encode the same amino acid are called *synonymous codons*. Most synonymous codons differ by only one base at their 3' end, the base in the third position.

C.1.2 The mediator molecule tRNA

In physical terms, genetic codes are mediated by tRNA (*transfer RNA*), molecules which are responsible for the translation of the message from nucleotides (the genes) to amino acids (the proteins). The tRNA molecules consist of 75–95 nucleotides and have an RNA reading end, called an *anticodon*, while on the opposite side they are bound to an amino acid (see Figure I-1 for a schematic representation – (*a*), called the cloverleaf diagram – and (*b*), a three-dimensional model. Multiple codons may be read by the same tRNA molecule, but usually there is a preferential codon that a tRNA molecule reads most efficiently (with optimal interaction energy). This unique codon is usually the one which is Watson-Crick complementary (*i.e.* A paired to U/T, C paired to G) to the anticodon of the tRNA. The group of different tRNAs that read the same set of synonymous codons are called *isoacceptor tRNAs*.

C.1.3 Translation: the ribosomes

The real translation process happens through the ribosomes, complex structures consisting of two unequally sized subunits which are composed of RNA molecules and proteins.

During protein synthesis a messenger molecule (mRNA, transcribed from a DNA template) moves through a ribosome. As it moves, amino acids are assembled into a gradually lengthening protein chain whose sequence corresponds to the transcript sequence, translated into the amino acid code. Having reached the end of the coded message (the STOP codon), translation stops and the ribosomal subunits separate, releasing the mRNA and the completed protein. The tRNA molecules provide the “dictionary” between the two codes, recognising codons on the mRNA and allowing the corresponding amino acid, that they carry, to be attached to the growing protein chain (Figure I-1 *c*).

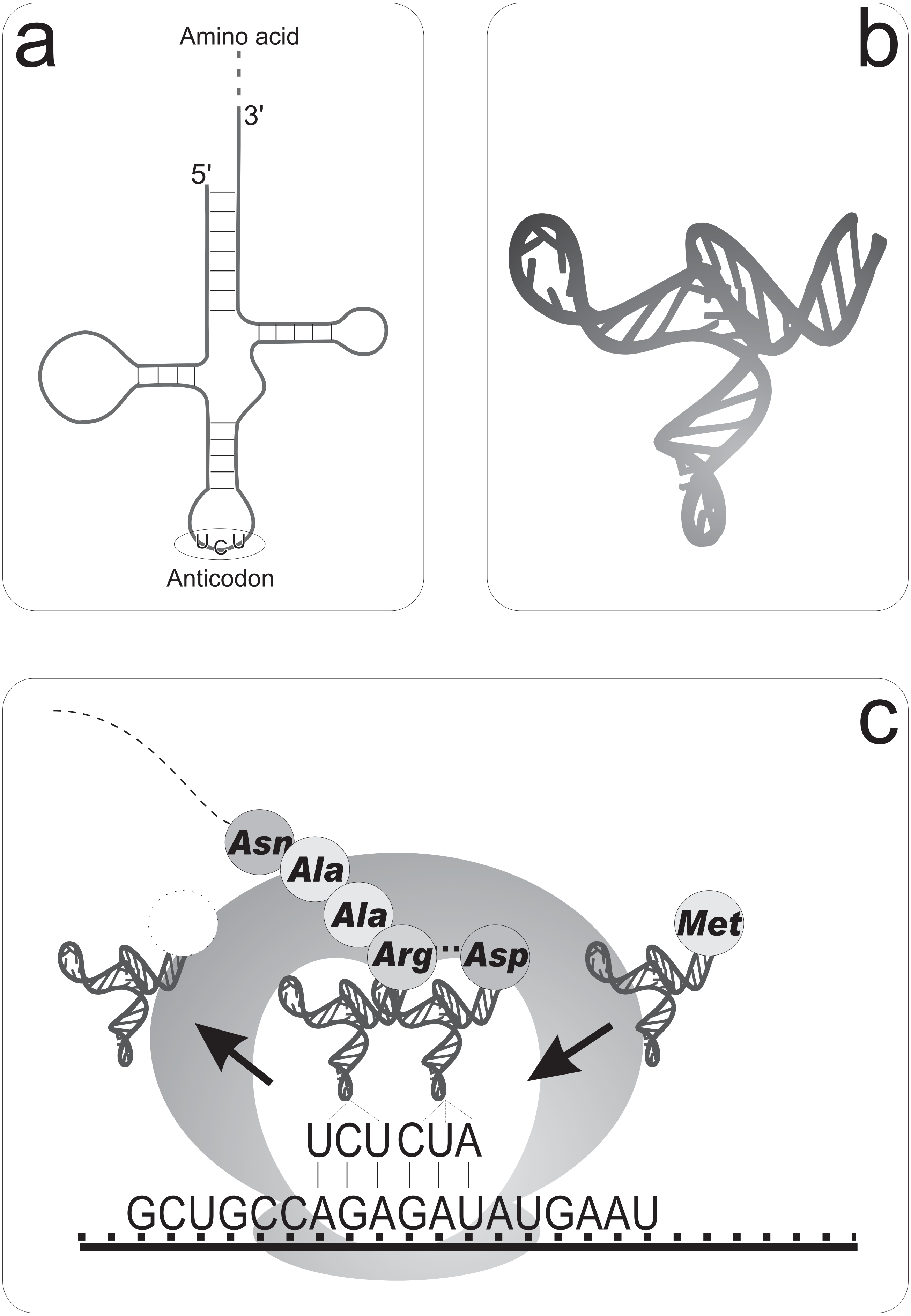


Figure I-1*: (a) Cloverleaf diagram and (b) three dimensional model of the tRNA molecule. (c) Schematic diagram of the translation mechanism and its main players: mRNA, tRNAs, amino acids and ribosomes.*

C.1.4 Codon usage

For some time after the discovery of the redundancy in the genetic code, it was often believed that synonymous codons for the same amino acid were used randomly in a genome. The simplest assumption would be that all genomes have uniform *codon usage* meaning that synonymous codons are used with equal frequency.

With more and more sequence data appearing in the late 1970s and early 1980s, it came to light that synonymous codon usage was nonrandom and that different genomes had different *preferred synonyms* for any given amino acid. These effects are known as *codon usage bias* or simply *codon usage*.

Before that time, most population geneticists imagined that the usage of codons for a given amino acid would be distributed only according to the background base composition in the genome: completely determined neutrally by the mutation processes in the replication of the genomes. For example, the bacterium *Escherichia coli* has a genomic G+C content (total sum of G and C bases over total number of bases) of about 50%. If codon usage were determined by mutation alone, then all genes would show the same frequencies in the use of synonymous codons, with a base composition having approximately equal numbers of G+C and A+T.

So in *E. coli*, the expectation would be that bases would be used in a random fashion, with G and C being used 50% of the times. In reality this is only true on average, with usually high variations of codon usage among different genes in the same genome (chapter III presents analyses of intra-genomic heterogeneity).

C.1.5 Translational efficiency

Other evolutionary forces besides mutation can influence codon usage bias. Selection for increased translational efficiency is one of them, but it is often considered to be negligible, especially in organisms like humans and other vertebrates, in intracellular bacteria and organelles (like mitochondria and chloroplasts). Although in all organisms, a set of codons may be translated more efficiently than others, a possible selective advantage, only in certain organisms (for example the above mentioned bacterium *Escherichia coli*) will this advantage actually influence the distribution of codons (in particular for the highly expressed genes). In these organisms a preferred subset of the genetic code enabling lag-free translation (in particular under conditions of high expression, and hence prevalently for highly-transcribed genes) could be positively selected, with mutations in the coding sequences favoured for those codons.

The currently accepted theory holds that the main factor for the differential translational efficiency of codons is that tRNA isoacceptors are present in different abundances in the cells. Those isoacceptors in the greatest abundance cause the codons which they recognize more efficiently (or accurately) to be translated in a more efficient way than any of their synonyms. Organisms in which selection acts on translational efficiency can have different preferred codons, according to the most abundant isoacceptors.

This was shown by Ikemura and colleagues in the 1980s. The preferred codons were the same ones as those predicted to translate the most efficiently by the tRNAs in the examined organisms. This correlation has since been made in some bacteria (like *Escherichia coli*, *Bacillus subtilis*, *Haemophilus influenzae*, *Mycobacterium tubercolosis*), in yeast and in some insects (Ikemura, 1981; Ikemura, 1982; Dong *et al.* 1996; Li and Luo, 1996; Percudani *et al.*, 1997; Kanaya *et al.*, 1999; Kanaya *et al.*, 2001a) while it was not found in mammals, where there appears to be little differential fitness for codons.

The probable reason to explain why natural selection acts with less potency on the choice of codons in mammals might be their small effective population sizes which would prevent selection from efficiently fixing preferred codons (Mooers and Holmes, 2000; Sharp *et al.*, 1993).

C.2 Code redundancy, superimposed messages

The primary function of DNA is the storage of genetic information. However, DNA also contains several signals, both compositional and structural (Schaap, 1971; Trifonov, 1989).

The redundancy of the genetic code is exploited in the genomes to superimpose the various biological messages (for instance the nucleosomal pattern or reading frame) or to satisfy constraints (like minimisation of palindromic sequences, avoidance of restriction enzyme cutting sites, DNA bendability and melting temperature), some of which will be presented in the next section.

The overlapping messages in multicode texts can only coexist due to their degeneracy: when some letters of one message can be replaced without much damage to that message, thus reaching a compromise with the other superimposed ones (Trifonov, 1989).

The constraints to the superimposition of messages are particularly relevant in organisms with small genomes, such as parasitic bacteria, in which the predominant evolutionary process is genome reduction (Koonin *et al.*, 1997), most viruses (expecially in those with fixed genome size determined by the size of the capsid in which the nucleic material is packed), and organelles.

Relying on the translation machinery of the host, viruses could try to reflect (undergo positive selection towards translational efficiency) the host codon usage, but this usually does not happen (as first reported by Grantham and collegues, 1980 and 1981) and is indirect evidence of the stronger constraints that the viral genomes need to satisfy. Examples are the avoidance of certain sequences that would be recognised by restriction enzymes (Sharp *et al.*, 1984), the maintenance of special features like palindromic regions for genomic superstructure branching, compositional deviant zones for genomic bending (Hertz *et al.*, 1987) or regions responsible for dimerization (Cain *et al.*, 2001; Andersen *et al.*, 2003).

Bacteriophages appear to be more influenced by tRNA abundances of the host (particularly in highly expressed genes). This was shown to be true for *E.coli* phages, with the exception of those that carry their own polymerase and can hence be subject to different mutational pressure (Sharp *et al.*, 1985; Kunisawa *et al.*, 1998).

But the need for superimposed message is not limited to the smaller genomes. Eukaryotes (with the exception of few protozoa and fungi) have genomes of at least two orders of magnitude larger than prokaryotes; these large genomes require specialised 'maintenance' systems and features on the DNA to regulate those systems (for example the nucleosome signal for proper chromatin packaging).

C.3 Phonological constraints: the reasons for nucleotide biases

This section presents a brief overview of several superimposed biological messages that DNA sequences can contain; in other words, of the mechanisms influencing the composition of DNA sequences and (in the case of coding sequences) the codon usage.

The two main causes thought to affect the patterns of codon usage are genome tRNA and G+C contents.

Codon usage in highly expressed genes has been positively correlated to tRNA content, in particular in bacteria, fungi and insects. In turn the tRNA content correlation can be explained as selection acting on translational efficiency (as discussed above, section C.1.5).

Besides translation efficiency, bacterial genomes meet criteria linked to G+C content, base compositional strand asymmetry and preferential gene orientation. These pressures occur independently of the coding function and they influence it: there are reports of protein constraints caused by codon usage, genomic G+C content and strand asymmetry (see Gautier, 2000, for a review). The asymmetry in gene orientation (genes are preferentially directed with their translation process occurring in the same direction as the genome replication) is usually considered as the result of selection pressure acting to avoid collisions between the replication and transcription mechanisms (Brewer, 1988). Strand compositional asymmetry, which would imply a difference in the substitution processes acting on the two strands, has been the subject of considerable research (Frank and Lobry, 1999).

In mammals, however, the pattern of synonymous codon usage appears to correlate only to the G+C content of the local genomic region (Bernardi *et al.*, 1985; Smith and Eyre-Walker, 2001). The nuclear genomes of vertebrates are mosaics of *isochores*, very long segments (more than 300kb) of DNA having different homogeneous G+C content and compositionally correlated with the coding sequences that they embed (see Bernardi, 2000, for a complete review). They are distinguished as higher-density level genomic segments (named heavy – H – isochores) and lower-density ones (light – L – isochores). G+C content of exons, introns and flanking sequences vary in accord with the isochore class in which they are located. The amino acid content of the encoded proteins is also affected, with amino acids coded by GC-rich codons (Ala, Arg, Gly, Pro) more frequent in H isochores. Furthermore, there is a higher frequency of genes in H isochores than in L ones.

Understanding of the evolutionary forces behind the evolution of isochores has generated considerable debate between neutralists and selectionists, with several models being proposed, but still remaining an unresolved question (Duret and Hurst, 2001). One of the hypotheses proposed: since GC-rich DNA is supposed to provide a double-helix more stable to heat, and a high density of protein coding genes is consistently found in homeothermic birds and mammals, a role was proposed for advantageous selection of GC-rich isochores in animals with high body temperatures. Emergence of these isochores would have accompanied the transition from cold- to warm-blooded vertebrates (Bernardi, 2000; but see Hughes *et al.*, 1999 and following works reporting isochore organisation in several reptiles).

Nevertheless, although vertebrate codon usage strongly reflects G+C content, nucleotide mutational biases are not sufficient to explain all the observed codon biases (Urrutia and Hurst, 2001).

C.3.1 DNA structure, curvature, flexibility

DNA structure, beyond the double-helix pattern, can play a fundamental role in a number of biological processes like DNA-protein interactions (Pazin and Kadonaga, 1997; Pedersen *et al.*, 1998), gene regulation and nucleosome positioning (see below). The curvature and deformability of the DNA molecule are critical for its packaging in the cell, recognition by other molecules, and transient opening during several important processes (transcription, replication, recombination and repair, to name a few).

The relation between exact sequences of DNA and their three-dimensional structure has been repeatedly shown (Brukner *et al.* 1990; Olson *et al.*, 1998). Several authors have been developing methodologies to evaluate the local sequence-directed curvature and flexibility of a DNA chain employing techniques like X-ray crystallography, electron microscopy and gel retardation (Zuccheri *et al.*, 2001). For example, sequence-dependent flexibility was found to correlate with the occurrence of AT-rich dinucleotide steps along the chain (Scipioni *et al.*, 2002). Databases of structural and flexibility properties have been compiled for dinucleotide or trinucleotide sets (see Ponomarenko *et al.*, 1999) and for octamers (Gardiner *et al.*, 2003).

Because structural and protein coding signals can be superimposed in coding regions, the genetic code should have a substantial degree of structural flexibility. In other words, there should be the possibility for an amino acid (or an amino acid class like *hydrophilic*) to be encoded by codons with very different structural properties, from stiff to bendable.

This was shown to be effectively true at the level of broad amino acid categories (the flexibility at the single amino acid level was reported to be mild) by Baisnée and co-workers (2001) using dinucleotide and trinucleotide models of DNA structure. They also demonstrated (in the *E.coli* genome) that there is practically no correlation between the structural properties of coding DNA and the physical properties of the encoded amino acids and proteins.

C.3.2 Nucleosomal pattern

Curved DNA is also related to nucleosomal positioning (Baldi *et al.*, 1996). The primary function of the nucleosomes (elementary repeating subunits of the chromatin structure, each formed by 146 *bp* of DNA wrapped around a protein octamer) is the packaging of DNA in a dynamic chromatin structure. However, the precise folding of regulatory sequences of genes around the histones within positioned nucleosomes is also important in controlling transcription and hence, in turn, influencing expression (Wolffe, 1994; Tsukiyama and Wu, 1997; Chen and Yang, 2001).

The nucleosome positioning pattern signal is one of the weakest (being highly degenerate) and is related to the bendability of DNA wrapped around histone octamers. A reason for this pattern being weak could be that chromatin needs to be easily unfolded to allow the processes of replication and transcription, thus the binding of histone octamers and the nucleotidic signal sequence should not be strong. Furthermore, the degeneracy of the positioning pattern guarantees the possibility of superimposition to other encoded messages (Bolshoy *et al.*, 1997).

Multiple sequence alignment shows that the main part of the signal is created by the recurrence of AA and TT dinucleotides at regular intervals (Ioshikhes *et al.*, 1996). The entire nucleosome site pattern consists of two regions, around 50 bp in length, with increased bending propensity, divided by a central 15-20 bp zone (Levitsky *et al.*, 1999).

C.3.3 RNA structure and stability

Functional and catalytic RNA molecules exhibit a characteristic secondary structure highly conserved in evolution. The most well known examples are tRNAs, rRNAs (ribosomal RNAs), and group I and II introns. RNA structure also plays an important role in the stability of mRNA molecules (transcripts), thus the conservation of RNA structure represents another message superimposed to the protein one in coding sequences. RNA binding proteins that stabilise or destabilise transcripts rarely recognise (unlike DNA binding proteins) distinct nucleotide sequences and instead bind to relatively long elements, suggesting that the RNA secondary structure of the sequences is an important factor in this process (for a very recent paper on the study of elements for mRNA stability within a yeast protein coding sequence see Vemula *et al.*, 2003).

Viral genomes in general and retroviral genomes in particular present the most striking examples of overlapping codes (including overlapping genes and messages on both forward and reverse strands). Viruses commonly use conserved RNA secondary structures located within protein-coding regions. Probably the most famous case of overlapping sequence is the *rev*-responsive element (RRE) located in the transmembrane section of the coding sequence for the *env* protein of HIV (Malim *et al.*, 1989).

Furthermore, formation of mRNA secondary structure could interfere with ribosome binding and hence negatively affect translation. Thus this additional constraint to the composition of coding sequences can be present near the initiation sites (Eyre-Walker and Bulmer, 1993).

C.3.4 Restriction avoidance

Bacteriophages are viruses which infect bacteria, injecting their DNA into the bacterial cells and taking control of their genetic machinery to replicate. The primary bacterial defense mechanism against bacteriophages are restriction enzymes, which cut DNA molecules at specific locations (*restriction sites*), usually palindromic in their sequence pattern. The presence (and the number) of a given restriction site in a phage makes it vulnerable to the respective cutting enzyme. There is considerable evidence that both phage genomes and bacterial genomes evolve to avoid the presence of restriction sites (Sharp, 1986; Karlin *et al.*, 1992; Gelfand and Koonin, 1997), thus representing an additional constraint to the form of the coding sequences.

The genetic code has been shown to be flexible enough to encode the proteins in such way as to avoid cutting sites, to the point that phages could be engineered for protection against all known restriction enzymes while still respecting a favourable codon distribution (Skiena, 2001). Interestingly, Skiena observes that since bacterial rRNA genes are usually located in the regions least depleted of palindromic sequences, this could be an indication of the tradeoff between the necessities of maintaining the functional RNA structure and that of avoiding cutting sites in the bacterial genomes.

C.4 Compositional analysis: codon usage and other techniques

Several compositional methodologies have been developed and applied to the analysis of nucleotide sequences or complete genomes. With sequence data increasing at an unprecedented pace, there have been increasing efforts to analyse, characterise and categorise this data using computational methods. The availability of several completely sequenced genomes allows comparisons which are not biased by selective sequencing.

Some of the main compositional procedures are overviewed in this section.

C.4.1 Compositional biases of dinucleotide abundances: genome signatures

There are multiple definitions of genomic signatures, however they are all based on the measurement of the frequencies of oligonucleotides (of a specific length) in genomic sequences (Karlin and Ladunga, 1994; Deschavanne *et al.*, 1999). Genomic signatures have been computed for complete genomes or just for the coding or non-coding portions, using oligonucleotides of several lengths.

Dinucleotide relative abundance values are computed from the ratio between the frequency of a given dinucleotide and the product of the frequencies of its two component nucleotides. A relative abundance sufficiently different from one shows the contrasts between the observed frequencies and those expected from random association. These ratios are reported to be constant throughout the genome, with levels of relative abundances for the dinucleotides being about the same for each 50 kilobase segment (Campbell *et al.*, 1999). Comparisons of dinucleotide abundances have been used as a measure of similarity between genomes. Their values being relatively constant for coding and non-coding DNA suggests the presence of genome-wide factors that influence and constrain the genomic compositional patterns (Karlin *et al.*, 1998).

Some general compositionally extreme trends that have been reported (Karlin and Burge, 1995; Karlin *et al.*, 1998) are: under-representation of the TA dinucleotide in both prokaryotes and eukaryotic nuclear genomes (but not in viral or organelle genomes nor in some archaea), possibly explained by the low thermodynamic stacking energy, the lowest, of this dinucleotide (Delcourt and Blake, 1991); AT being over-represented in most α-proteobacteria; CG drastically under-represented (relative abundance values of 0.23–0.37) in vertebrates, usually ascribed to methylation dependent CG$\rightarrow$TG mutations but alternatively explained by chromatin packing constraints.

Species-specific signature appears to be a common feature of the genomes, especially in prokaryotes. The species-specificity of genomic signatures was recently quantified by Sandberg *et al.* (2003), who computed classification accuracy for genomic signatures, nucleotide biases, amino acid biases and synonymous codon usage. Synonymous codon usage was shown to capture most of the species-specificity of genomic signatures of prokaryotes (better than trinucleotide signatures and at 86% of the accuracy achieved with oligonucleotides of length nine; amino acid usages capture approximately 50% or less).

C.4.2 Nucleotide biases

According to the base pairing rules, or Chargaff's rules (1951), in the double helix of DNA the nucleotide Guanine is held together with Cytosine while Adenine pairs with Thymine. In the double-strand molecule, the total amount of pyrimidine nucleotides always equals the total amount of purine nucleotides (C+T=A+G), the amount of A always equals the amount of T (A=T), the amount of C always equals the amount of G (C=G). The amount of A+T does not need to equal the amount of C+G.

A tendency was noted since the 1950s for the ratio of C+G to the total bases (A+C+G+T) being constant in a particular species, but variable between species. The total content of Guanine and Cytosine content of bacterial DNA was observed to range from approximately 25% to around 75% (Lee *et al.*, 1956), with both mutation and selection being proposed as explanations for the biases.

For single strand sequences, the genomic content of the four nucleotides or couple of nucleotides can be computed as total measures, analysed for a sliding window over the genome or calculated for specific subsets (for example coding versus non-coding elements). For coding sequences there is the additionally possibility of computing nucleotide biases at different coding positions. For example *GC3* stands for the percentage of G and C nucleotides to be found as the third base in the coding triplets.

The two strands of a DNA helix must have the same G+C content but the different bases can still vary in frequency. For example, one strand may be more rich in G than C and, by complementarity, the other one will have more C than G.

C.4.3 Codon biases

Since alternative codons for any amino acid are not used randomly, it is desirable to give quantitative measures of the degree of bias for genes or genomes in such a way as to allow comparisons both within and between species. Several methods have been devised to address this goal.

The earliest studies aimed at elucidating the nature of codon biases began in the mid 1980s (with pioneering work from Grantham, Ikemura, followed by Sharp and by other major groups) and continues with ongoing effort to the present day. One of the common aspects that these works revealed was the relationship of the nucleotide composition in third codon position and the local or global genomic nucleotide composition.

In 1987, the *Codon Adaptation Index* (CAI) was proposed by Sharp and Li as a quantitative measure of codon bias, to be used for example to predict the level of expression of a gene. An alternative measure, the *Codon bias between gene classes*, was introduced in 1998 by Karlin *et al*. and is based on gene collections like ribosomal, chaperones and translation processing factors. These two methodologies are important expression level indicators and are used in several contexts, in particular for heterologous gene expression in order to optimize codon usages and yield high expression. Codon-optimization refers to the alteration of gene sequences to make the codon usage match the available cellular tRNA pool within the species of interest.

The CAI assesses the relative merits of different codons based on translational efficiency. It is based on a reference set of highly expressed genes from which optimal codon frequencies are extracted. Ratios between the frequency of each codon and the maximal synonymous codon frequency (the frequency of the most used synonymous triplet for the same amino acid) are tabulated and used to compute the CAI of a given transcript, which is the geometric average of the ratios for all its triplets. High values (approximating 1) of the CAI correlate with high expression levels. Genes experimentally known to be highly expressed include most ribosomal protein ones, those coding for elongation factors and some membrane genes. CAI needs genome specific tables of codons for highly expressed genes. Although this scheme can appear limited because of its qualitative nature and because it was originally based on only 24 genes (half of which were ribosomal), it sufficiently captures the codon information for the most expressed genes, as was recently shown by Jansen and collegues (2003): they performed parameterization of the CAI model (and of Karlin's *codon bias between classes*) using expression data from yeast. Their results of correlation between codon usage model and expression data show that few highly expressed genes are sufficient to describe the overall bias.

If tables for highly expressed genes are not available or if the interest is not focused on expression levels and translational efficiency, then neutral measures of codon biases can be employed.

The computation of relative synonymous codon usage frequencies is the most frequently used codon bias parameter, in particular in correspondence analysis studies (Perrière and Thioulouse, 2002; for example: Grantham *et al.*, 1980; Holm, 1986; Shields and Sharp, 1987; McInerney *et al.*, 1997; Lafay *et al.*, 2000). This measure of codon usage corresponds to the observed frequency of a given codon divided by its expected value under the hypothesis of a random distribution of all its synonymous triplets.

Other methods that have been used are: simple codon frequencies independent of the genetic code (non-synonymous codon usage, where every codon is considered independently and not paired to its synonymous ones) and absolute codon occurrences (which have the drawback of reflecting the amino acid bias of the proteins encoded by the transcripts, but which are in some cases beneficial and more sensitive; Lafay *et al.*, 2000).

Another important and widely used codon usage statistic is the *effective number of codons* (N_C; Wright, 1990), which measures the amount of bias away from equal usage of synonyms with values that range between 20 (for extremely biased genes where only one codon is used per amino acid) and 61 (when all codons are used with equal probability).

C.4.4 Frequent and rare words

Some studies focused on the determination of which words (oligonucleotides) occur with unusually high or low frequency in the genomes, and with which distribution. Rare words could be binding sites for specific transcription factors, structurally deleterious sequences or restriction enzyme cutting sites. As for frequent words, they often are parts of repetitive structural, regulatory or transposable elements. In other cases, they reflect protein motifs. Comparisons between the abundance and localization of these words can identify evolutionary tendencies and genomic constraints (Burge *et al.*, 1992).

II *Codon profile and codon profiling*

A ABSTRACT

A novel methodology, called *codon profiling*, was devised to represent and compare the preferential usage of codons in genes and genomes. It computes base and position specific biases in synonymous codons as opposed to triplet relative frequencies. Automatic classification of nucleotidic sequences can be performed by analysis of codon usage information in a metric vector space. The very general scheme employed makes the methodology independent of the genetic code of the studied genome, allowing the analysis, for example, of nuclear and mitochondrial genomes together.

Codon profiling was compared to the very similar and widely used synonymous codon usage methodology, examining the relative benefits of the two schemes. All analyses presented in this dissertation were performed simultaneously with both techniques, and differences in the results have been reported.

Various programs were developed for the calculation and display of codon information and are made publicly available.

B INTRODUCTION

Since the first whole genome was sequenced in 1978 (Sanger *et al.*, 1978) and the first one of a free living organism in 1995 (Fleischmann *et al.*, 1995), numerous genomes have been sequenced every year. This provides an enormous resource for comparative genomic analysis. Specifically, it allows taxonomic analysis based on the whole genome information rather than on a specific set of genes such as the 16S ribosomal gene. However, these methods require a great amount of computing power and detailed analysis by experts for all the genes involved. Simple methods such as the analysis of G+C content and synonymous codon usage can provide faster identification of regions of interest.

In this thesis, a way to integrate codon usage and genomic base composition analysis (such as G+C content) was devised and named *codon profiling*. It draws from codon information (occurrences of triplets in coding sequences) but it presents it and analyses it from a different point of view than other codon usage techniques. One of the most-employed of these techniques, and the most similar to codon profile, is *synonymous codon usage* (the *relative frequencies* method) to which codon profile will be compared in this work.

All the analysis discussed in this thesis have always been conducted in parallel with the synonymous codon usage technique, constantly noticing the near equivalence of these two techniques in the results they obtain. Codon profiling was preferred for its generality, elegance and different approach, but also for its higher (although not markedly so) sensitivity in some cases, which will be described. Nevertheless, all the results presented, obtained through codon profile vectors, were repeated and verified using synonymous codon usage vectors.

B.1 Messages beyond the triplet

Codon profiling arose from the interest towards the additional constraints in coding sequences: how the redundancy of the genetic code is exploited to superimpose various messages on top of the peptide one. The coding portions of a genome enable us to specifically search for these signals, separating the protein message and investigating the other constraints (*q.v.* I C.3 for an overview of several important ones).

Like in synonymous codon usage analysis, the amino acid bias (which amino acids are used more and which ones less frequently in protein sequences) is eliminated with codon profiles. Codon profiling also tries to reduce in part the contribution of the component related to cell tRNA content, by working with the single nucleotide as the minimal unit. This is in line with the understanding that, although a correlation was found between cellular tRNA content and codon usage in a number of organisms, the selection pressure acting on translational efficiency is considered weak (see I C.1.5 on translational efficiency).

Several forces contribute to the shaping of codon usage, and different organisms may have different constraints to the form of the coding sequences. Changing the point of view can maybe help the investigation of codon usage patterns.

C METHODS

C.1 Codon profiling

C.1.1 Shifting the point of view

A codon profile is a record of the preferential use of the four bases at the three individual triplet positions inside the codon, for all amino acids. The information of codon occurrences in coding sequences is presented and analysed from the *base-at-position* point of view. This point of view is more oriented towards biases which are due to genomic nucleotidic preferences (and hence mutational biases and compositional constraints).

The following Table II-1 illustrates a comparison between codon profile and synonymous codon usage to represent the codon usage information for the amino acid Arginine in the human genome.

Synonymous codon usage		**Codon profile**			
CGT	8%	*base*	*position in the triplet*		
CGC	19%		**1**	**2**	**3**
CGA	11%	**T**	0	0	0.08
CGG	21%	**C**	0.59	0	0.19
AGA	21%	**A**	0.41	0	0.32
AGG	20%	**G**	0	1	0.41
sum	100%	*sum*	1	1	1

***Table II-1**: Codon usage for the amino acid Arginine in the human genome; comparison between synonymous codon usage and codon profile methodologies.*

In both methods the amino acid bias (which amino acids are used more and which ones less frequently in protein sequences) is eliminated and relative frequencies (as opposed to absolute codon occurrences) are computed. Use of relative frequencies can give rise to artifactual distributions; this is circumvented by appropriate filtering of the data (see below, section C.4).

C.1.2 Combined contributions

For the majority of amino acids, codon profile analysis is equivalent to synonymous codon usage, whereas it behaves differently for 6-fold or 8-fold degenerate amino acids. For those amino acids (like Serine, Leucine and Arginine in the *Standard code*, Threonine in the *Yeast mitochondrial code*) and also for Glutamine in *Ciliata, Dasycladacean and Hexamita nuclear code* the codon profile combines the contribution of the triplets, as in the example shown above (Table II-1): the high Arg_G3 (relative frequency of Guanine in third position for codons coding for Arginine) of 0.41 reflects the abundances of both CGG and AGG codons. Section D.1 discusses the differences between the codon profile and synonymous codon usage analyses.

Codon profile is more focused on revealing genome-wide preferences such as those from the *GC3* analysis (G+C content in the third coding base), but without being restricted to that single aspect. Nucleotide composition analyses are still widely used and a source of precious information, even if their nature is coarse and prone to averaging effects. Codon profiling is the extension of G+C content (and similar base specific studies) by combining it with codon usage analysis.

C.1.3 Generality

Another advantage of the codon profile method is its generality: it can accommodate different translation tables (all the existing ones, those still to be discovered and the artificially created ones), since it does not start with a pre-defined setup for the genetic code.

Synonymous codon usage analysis would use vectors of different dimensionality for different genetic codes. Additionally, the components of these vectors would have labels that would depend on the translation table. For example, the dimension which was relative to an Isoleucine triplet in the Standard genetic code would refer to one of the two triplets which encode Methionine in the mitochondrial code (or to some other triplet, depending on how the triplets are sorted – mapped – in the vector).

In codon profiles, a single set of parameters is used for all possible genomes, allowing for example the comparison between nuclear and mitochondrial genomes, or the analysis of *Mycoplasma* bacteria together with the other bacterial species.

The generality of codon profile vectors is an important aspect of the methodology which leads to a consistent framework, a uniform labelling of vectors as well as the possibility to use the same analysis programs and the same visualization tools regardless of the problem being investigated.

C.1.4 Dimensionality

Since 20+1 coding possibilities are analysed in 12-element tables (4 bases · 3 codon positions), the resulting codon profile data vector has 252 elements. Most of the dimensions in the codon profile vectors would either always be zero or always one, according to the genetic code. These dimensions do not contribute to measures of codon composition similarity, are automatically ignored by multivariate analysis algorithms and hence do not negatively affect performance. They form the basis of the generality of codon profile vectors, making them independent of the translation table.

C.2 Codon profile vectors and measure of distance

Any two codon profiles can be thought of as two points in the multidimensional space represented by all possible codon profiles.

If two transcripts (or two groups of transcripts or two genomes) have a similar codon composition the distance between the corresponding two points in that vector space will be short. Conversely, if the relative occurrences of the codons in the two sets are very different, the distance will be long.

The Euclidean (geometric) distance in the codon profile space was adopted as a convenient and suitable measure of dissimilarity between two codon profiles. The Euclidean distance can be calculated as follows:

$$\mathrm{dist}_{252}(p,q) = \sqrt{\sum_{i=1}^{252} (p_i - q_i)^2}$$

where *p* and *q* are vectors of 252 dimensions that contain the relative frequencies of any two codon profiles.

Note that the *empty dimensions* will not affect the measure of distance: for example for all data points (in the known genetic codes) the dimension relative to Ser_G1 (content of Guanine in first position for Serine coding codons) will always be 0. Conversely, the dimension relative to Val_G1 will always be 1 for all data points. Hence the contribution of those dimensions for any two vectors (for both being the same) will cancel and will not contribute to the measure of dissimilarity.

In all analyses the STOP codons were not considered, since they are statistically under-represented, occurring only once per transcript. The terminator information contributed neither to filtering, nor to the measures of codon similarity (which are hence computed on the first 240 dimensions of the vector).

C.3 Display: single matrices and difference matrices

One of the possible ways to display codon profiles is achieved by arranging the elements of the vector in a graphical matrix form (Figure II-1 *a*).

Some amino acids have no synonymous codons, so they appear with three fixed blue (frequency=1) squares, according to the single codon coding for them. For example, Methionine (Met) can only be encoded (in the Standard code) by ATG, so the box relative to Methionine in the matrix display will have three completely blue squares in correspondence with A in the first, T in the second and G in the third position.

Other amino acids reflect the codon variability, which is usually restricted to the third position, but (in the cases of Arginine, Serine, Leucine and *STOP*) can also involve first or second positions. The scale indicates the relative frequency, as outlined above. The sum of the elements in each of the columns, indicating the triplet positions, adds up to 1.00, like in Table II-1 above.

This matrix form is especially useful for displaying differential matrices, showing at a glance which are the biggest differences between two codon profiles (which may be computed from two genomes, two protein families or two chromosomes). In this case the colouring of the scale indicates the difference in frequency between the corresponding positions of the two codon profiles. Hence the majority of the squares will be white, with no colour, indicating a difference of 0 (in which case the two codon profiles are equivalent for those vector dimensions). Slight differences will be lightly shaded squares and great differences are indicated by stronger shades. Positive differences are in blue while negative differences appear in red. The matrix in Figure II-1 *b*, for example, illustrates the difference between human and HIV-1 codon biases.

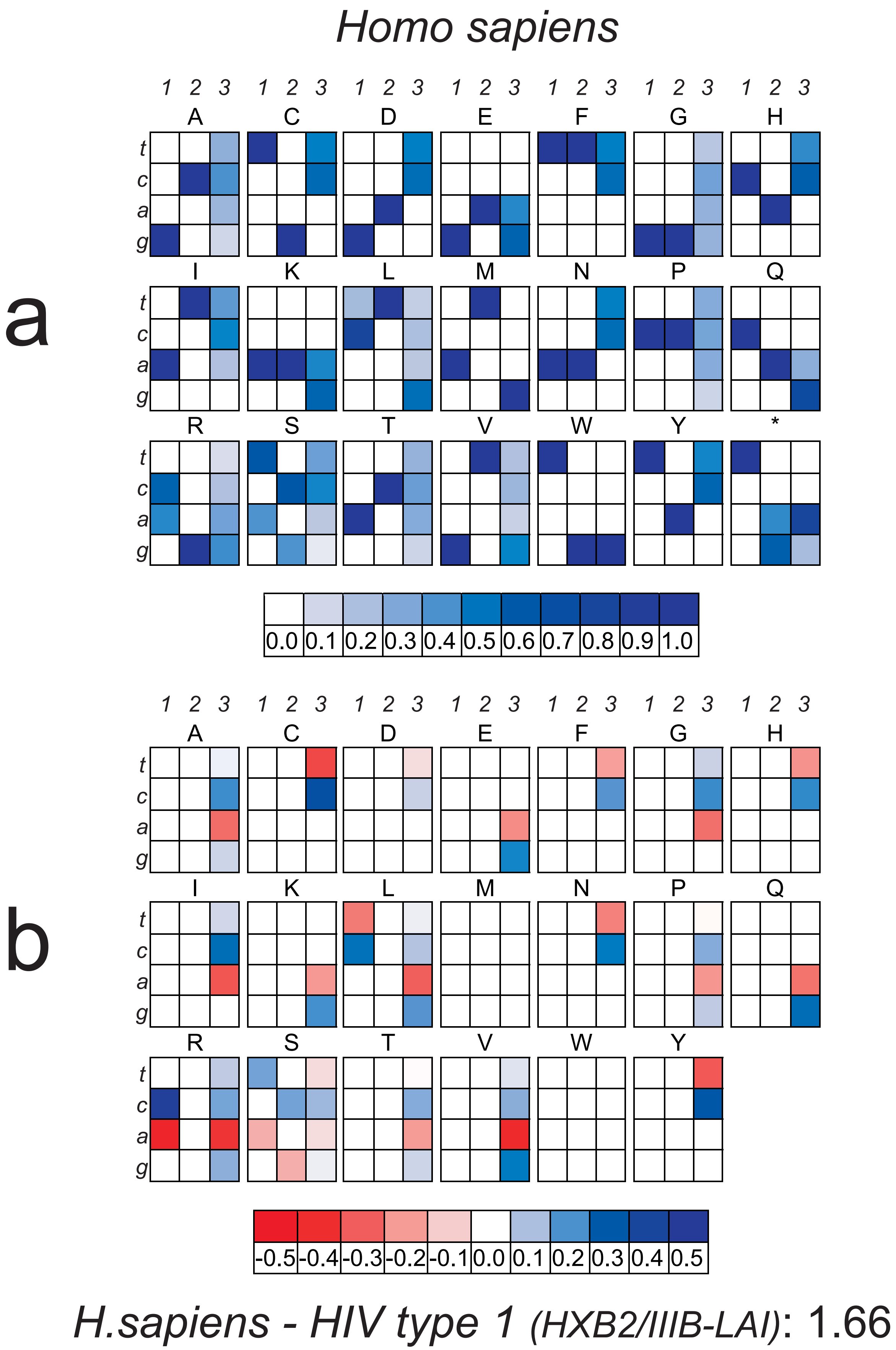


***Figure II-1**: (a) Homo sapiens codon bias (total codon usage over all sequenced transcripts, sequence data from Ensembl release 8.30a.1; Hubbard et al., 2002) displayed in the codon profile matrix form. (b) Codon profile difference matrix and Euclidean distance between Homo sapiens and Human Immunodeficiency Virus 1 (codon bias for reference strain HXB2/IIIB-LAI, sequence from GenBank entry K03455; Ratner et al., 1985).*

C.4 Filtering the datasets to prevent artifactual distributions

In the analysis of short transcripts, codon frequencies are susceptible to large stochastic variation. To minimise this, some authors select only transcripts longer than a certain amount of bases, for example longer than 300 bp (Garcia-Vallvé *et al.*, 2000; Kanaya *et al.*, 2001b). Nevertheless, the lack of particular codons in a sequence can create artifactual multivariate clusterings, especially when using – as in synonymous codon usage or codon profiling – relative frequencies (Perrière and Thioulouse, 2002). This could obscure more interesting trends in the data.

Transcripts coding for peptides without Cysteine or without Tryptophan residues are not infrequent. These transcripts would appear very atypical and cluster together, concealing other more interesting transcripts, namely those with different but not abnormal codon usage. For example, the transcripts missing codons for those amino acids would determine one of the first principal coordinates of separation in multivariate analysis.

To clean the dataset and eliminate those transcripts, three kinds of filters can be used, listed here in the order of the most restrictive to the most permissive:

* **CSYN-filter**: the transcript is kept if it has at least one member of each triplet kind; only those transcripts which contain each of the 61 species of codons would be further analysed

* **CPRO-filter**: the transcript is kept if it holds information in all codon profile dimensions (for the appropriate genetic code); due to the combination of triplet contributions, this filter is less restrictive than the previous one (*q.v.* D.1.2)

* **AA-filter**: the transcript is kept if it has at least one codon coding for each amino acid (if it encodes the full repertoire of amino acids)

In single transcripts – and in particular in archaea and in bacteria, since their genes are relatively short – it is statistically unlikely to find a full repertoire of the 61 triplets. For this reason the AA-filter is best used to clean data sets composed of single transcripts. On the other hand, a more restrictive filter can be used for transcript clusters (*e.g.* clusters of transcripts belonging to the same protein family).

The following Figure II-2 shows the application of the AA-filter on the transcripts from the genome of the bacterium *Pseudomonas aeruginosa*.

From a total of 5,565 transcripts, 1,601 are removed by the AA-filter. The risk of an artifactual distribution without application of this procedure is already obvious at this stage, considering the high number of transcripts that form a sort of *band* in correspondence with an Euclidean distance of 2 (and getting smeared for higher distances, even reaching values of 5 units) from the mean codon bias. These are all those transcripts which lack codons for one or more amino acids; they are removed from the data set when the filter is applied.

Considering all the analysed completely sequenced prokaryotic genomes (see appendix III F.1 for a list), the three amino acids which contribute the most to the removal of transcripts are Cysteine (accounting for 54.5% of the removed transcripts), Tryptophan (20.5%) and Histidine (7.5%).

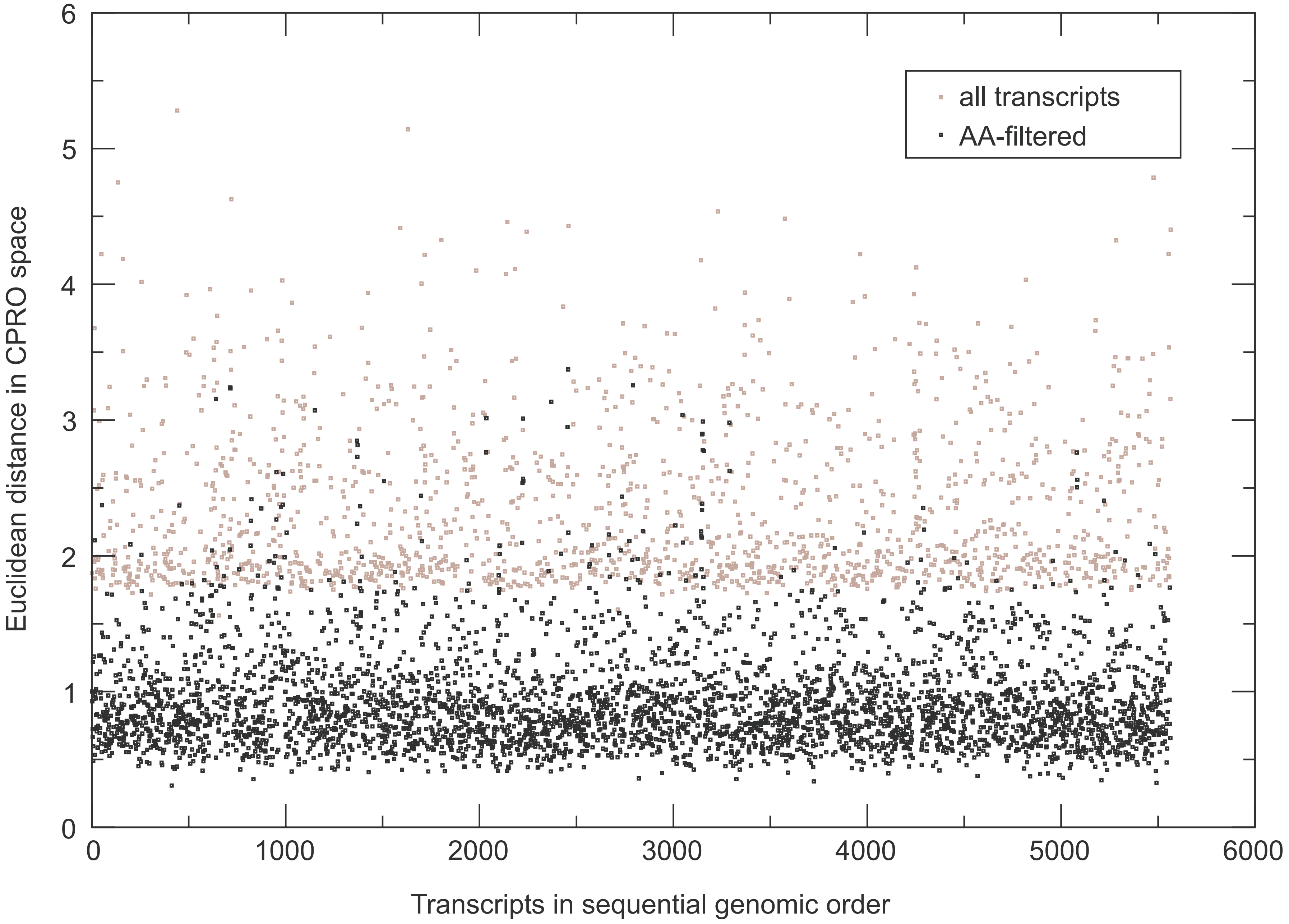


***Figure II-2**: P.aeruginosa transcripts, arranged in the same order in which they are encoded in the genome. The application of the AA-filter leaves 3,964 transcripts (in black) out of a total of 5,565. The filtered transcripts encode the full repertoire of amino acids.*

C.4.1 Masking instead of Filtering

An alternative to filtering the transcripts missing codon information (which could be felt as a loss of valuable data) is the masking of empty dimensions. For example, transcripts missing all codons for Cysteine would inherit, in the corresponding dimensions of the codon profile vector (or synonymous codon usage vector), the values relative to the genomic average distribution of Cysteine codons.

While this might seem a more desirable procedure, since it does not drop potentially valuable information, it is not exempt from problems. Namely, where to set the limit for the masking: how many dimensions are allowed to be substituted because of missing information? Without a (necessarily arbitrary) limit, the masking could include extremely short transcripts, substituting their codon information with the average bias.

For example in the case of a short transcript which lacks information for half of the codon species, the application of the masking would result in a codon profile which is half anomalous and half exactly like the genome average, and this codon profile would probably stand out as atypical in its own way.

Masking leads to chimeric and artificial codon usages with unknown and potentially misleading properties. It is therefore best avoided.

C.5 Coloured codons and musical codons

C.5.1 Coloured codons

A visual way to represent codons has been devised. A symbol (a coloured shape) is assigned to each codon. The symbols were created in a coherent way, with the three properties *shape*, *inner colour* and *border colour* mapping uniquely to the nucleotides forming the codon triplets. These coloured symbols (denoted *Coloured codons*) can be used to enhance the differences in synonymous codon usage for every transcript forming a protein family.

Figure II-3 shows the genetic code through the coloured codons symbols. It is easy to notice that the symbols were created with consistency to the nucleotides in the triplets: there are four possible border colours (which stand for the four possible bases in first codon position), four possible shapes (to represent the middle codon position, the one most linked to the chemical characteristics of the encoded amino acids) and four possible filling colours (which stand for the possibilities in the third position in the triplet, the *"wobbling"* one).

An example application of the coloured codons analysis will be given below, rendering transcript sequences with these symbols (see V D.3). Visual inspection of sequences (and in particular clusters of sequences, as in the experiment performed) is greatly eased by use of the coloured symbols (especially if compared to the normal representation, using the Latin letters TCAG – or UCAG for RNA), allowing an easier and faster discovery of synonymous triplet preferences or other strong patterns.

Additionally, this process allows recovery and observation of the sequential codon information which is normally lost in compositional analyses. The order in which the triplets appear along the gene is made apparent, leading to the possible discovery of gene positional patterns (for example certain preferences for some synonymous triplets at the beginning of the transcript) or identification of possible gene fusions when observing very different codon usages in the two halves of a sequence.

C.5.2 Musical codons

Another alternative way developed to represent biological sequences, specifically the coding ones, makes use of sound. The interest in representing genetic patterns in music is both pedagogical and aesthetic.

Most algorithms that convert DNA sequences to music (for example Hayashi and Munakata, 1984; Ohno and Ohno, 1986) adopt a one-to-one correspondence between the four nucleotides and four notes. But when representing the coding sequences it is probably best to facilitate the perception of the triplets and the following of the correct "listening frame". For this reason the *Musical codons* concept makes use of rhythm, in addition to pitches.

By employing four different rhythm structures and four notes (which can be further distinguished as appearing at the beginning or at the end of the rhythm structure), sixty-four musical combinations were defined and assigned to the triplets of the genetic code (the mapping is represented in Figure II-4). This was done with the same scheme as the coloured codons described above, following the same principles. The three properties of *shape*, *inner colour* and *border colour* described above are mirrored in the corresponding *rhythm structure*, *beginning note* and *ending note*.

The programs coded to translate sequence to music are available through a graphical interface which can be reached from the internet address: *http://www.ebi.ac.uk/~insana/codonprofile/*. The conversion of biological information into tones and rhythms enables the shifting of the pattern search and recognition processes to the musical (and hence temporal) domain. In addition to allowing a different approach to the biological sequence analysis, the musical sequences can find application in works of artistic science (or scientific art) and in popularisation of science.

While not appearing in the work of this thesis, the algorithm to generate musical codons is used by the *Sonic Genes* project, with which the author collaborates (The art of DNA, Economist April 2003). Sonic Genes is an ongoing research project – started in 2001 by Dr. Sophie Dauvois – that investigates ways of converting genetics into music. This collaboration between geneticists and musicians, which proposes to translate the human genome into music, merges scientific knowledge and artistic expression to produce soundscapes from DNA sequences. The project is being supported by the Wellcome Trust program *Science on Stage and Screen* through a Research and Development grant.

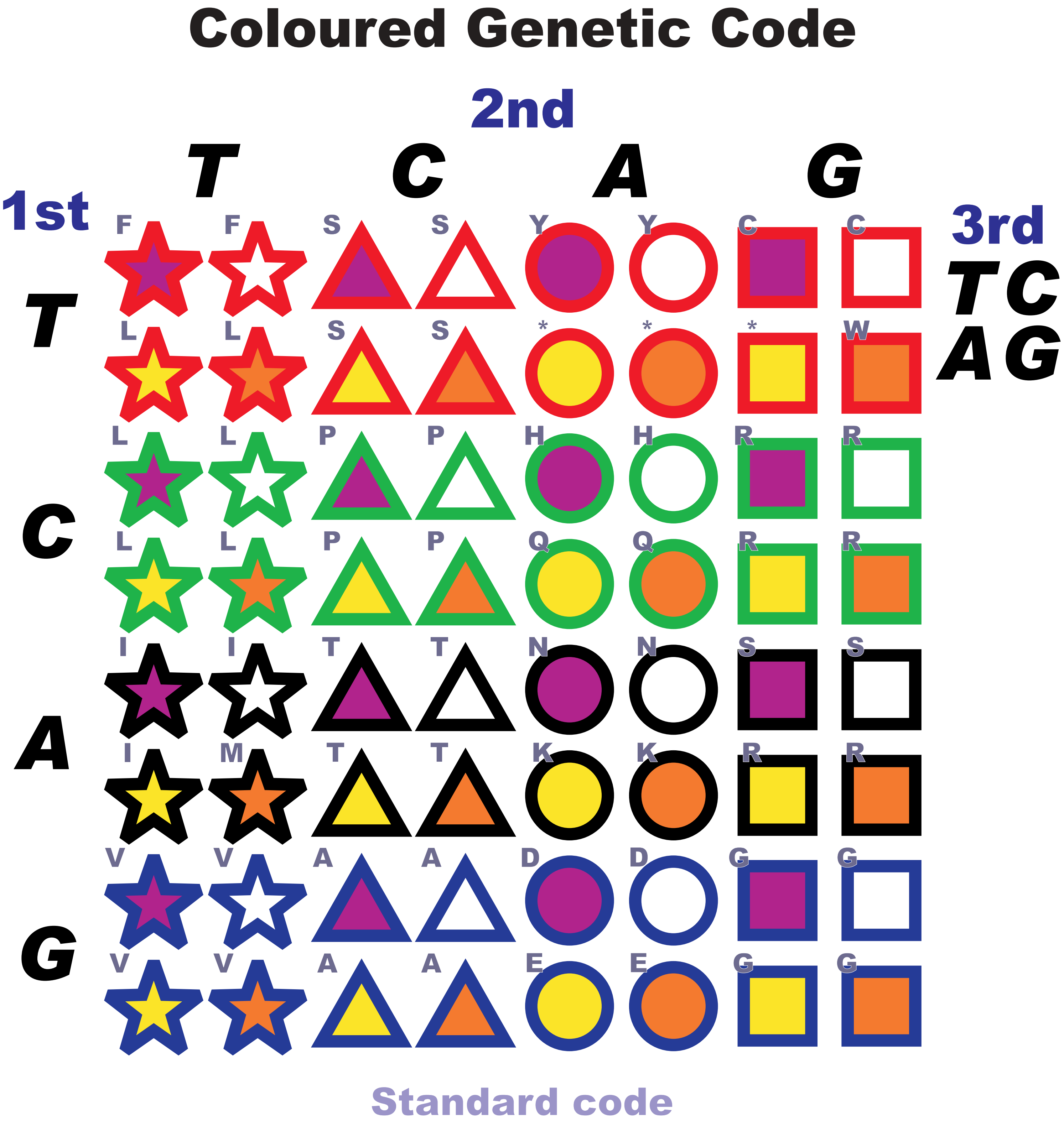


Figure II-3: *The Standard genetic code with coloured codons. Mapping of coloured shapes to the triplets.*

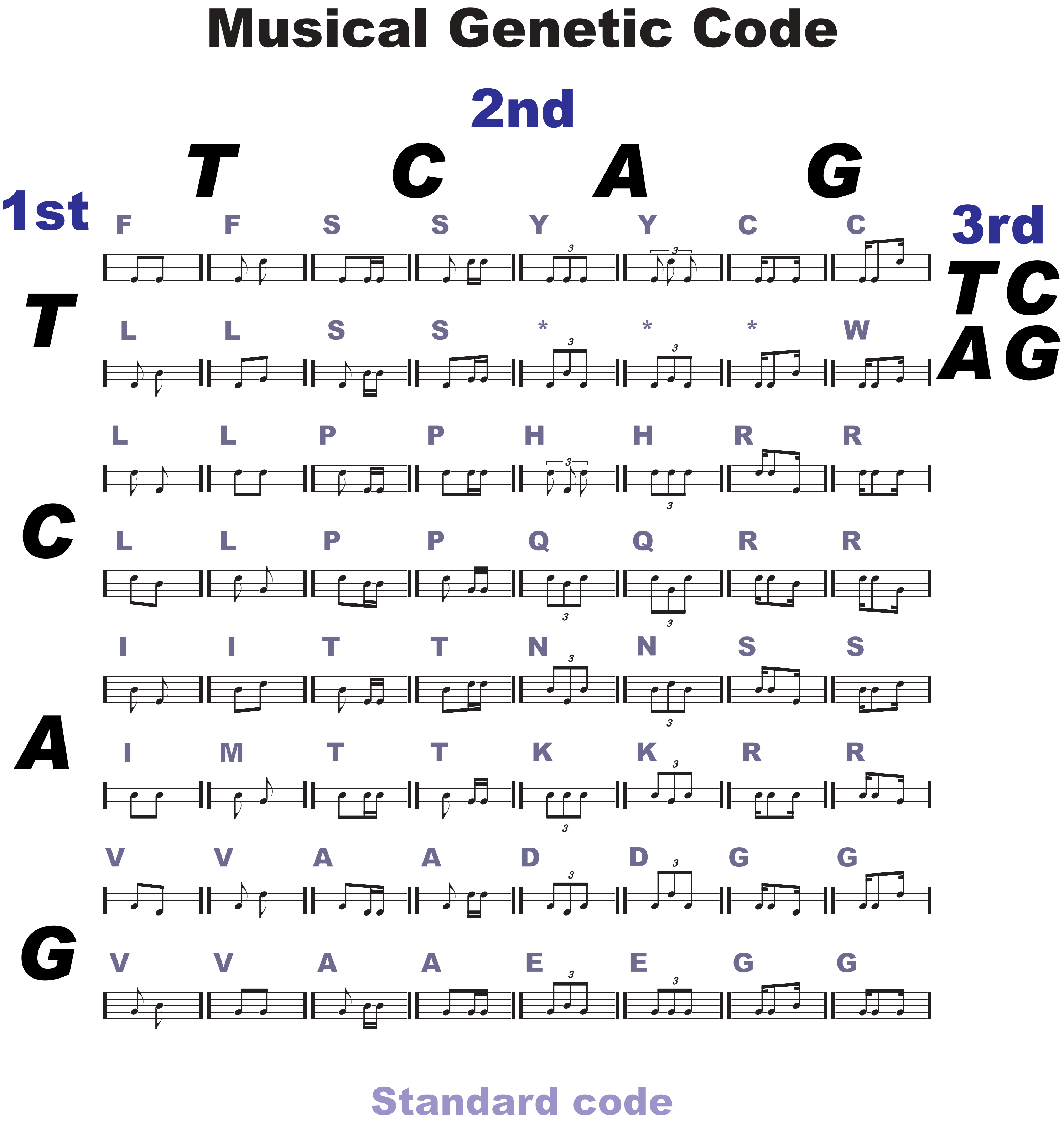


Figure II-4: *The Standard genetic code with musical codons. Mapping of melodic units to the triplets.*

D RESULTS AND DISCUSSION

D.1 Comparison with synonymous codon usage analysis

Synonymous codon usage (CSYN) and Codon profile (CPRO) vectors are computed from the same source, namely codon occurrences, which is the abundance of the 64 triplets in the coding sequences. Nevertheless they analyse this information in a different way, focusing on different aspects.

Both eliminate the amino acid bias (*i.e.* the same CSYN or CPRO would be computed from sets of data with, for example, a 30:1 or a 1:30 ratio between Alanine and Arginine residues). Furthermore, both look at relative biases rather than absolute codon occurrences. But CSYN takes the minimal unit to be the codon, while CPRO has the individual bases as the minimal unit. Apart from the different perception, labelling, display methods and generality, the shifted point of view translates into differences in the treatment of 6-fold and 8-fold degenerate amino acids. Consequently certain information will be preferentially exposed by one methodology while shielded by the other.

CPRO is targeted at individual position specific base propensities, while CSYN looks at differential codon usage. Alternative names for CPRO could hence be *synonymous base specific bias* or *synonymous TCAG123*. In fact, if two data sets are analysed, differing only in the use of codons for a single amino acid type, CPRO would be equal to TCAG123 (nucleotide propensities at individual codon positions).

For the majority of amino acids, codon profile vectors are completely equivalent to synonymous codon usage vectors. The differences are restricted to the 6-fold or 8-fold degenerate amino acids (three of them present in the Standard code: Arginine, Leucine and Serine).

For such amino acids, both CSYN and CPRO intrinsically hide and expose some information. Some examples are useful to better appreciate the differences between the two techniques; several codon usages for the amino acid Arginine will be presented and discussed in terms of CSYN and CPRO.

D.1.1 Shielded information

Some information is hidden in CPRO compared to CSYN. In other words, the codon profile vectors shield some differential codon preferences that are exposed by synonymous codon usage.

Two different CSYN sets for Arginine that give rise to the same CPRO are represented in Table II-2.

Synonymous codon usage		**Codon profile**			
CGT	0%	*base*	*position in the triplet*		
CGC	0%		**1**	**2**	**3**
CGA	50%	**T**	0	0	0.0
CGG	0%	**C**	0.5	0	0.0
AGA	0%	**A**	0.5	0	0.5
AGG	50%	**G**	0	1.0	0.5

Synonymous codon usage		**Codon profile**			
CGT	0%	*base*	*position in the triplet*		
CGC	0%		**1**	**2**	**3**
CGA	0%	**T**	0	0	0.0
CGG	50%	**C**	0.5	0	0.0
AGA	50%	**A**	0.5	0	0.5
AGG	0%	**G**	0	1.0	0.5

Table II-2*: Shielded information: two CSYN sets that CPRO considers equal because the base propensities of the two sets are the same.*

This is in the spirit of codon profile, where triplet occurrences that give rise to the same base propensities are treated equally, ignoring their difference. The decision to sum these contributions was taken because of the correlations, observed in codon usages studies, between triplets and nucleotide contents. The triplets that contribute to the same nucleotide contents are equally considered and this produces higher separation between data sets in which those triplets are correlated.

The extent of hidden information can be quantified: if considering discrete relative frequencies in multiples of 0.1 (*i.e.* 10%), there are in total 126 possible CSYN vectors for a 6-fold degenerate amino acid; these translate into 105 unique CPRO vectors. In reality, using actual biological data, it is quite rare to encounter these cases as they require a certain symmetry in the distribution of triplets (with frequency distributions such as <0.2 0.1 0.2 0.2 0.1 0.2>) for perfect complementarities of all the elements in the CPRO vector.

D.1.2 Complementary information

As the information for some dimensions in the 6-fold and 8-fold degenerate amino acids is the result of the contribution of different synonymous triplets, CPRO vectors can have data in all useful dimensions in some cases where CSYN vectors would be missing data.

Consider the following set (Table II-3), where one of the synonymous triplets is missing in the data element.

Synonymous codon usage		**Codon profile**			
CGT	20%	*base*	*position in the triplet*		
CGC	20%		**1**	**2**	**3**
CGA	20%	**T**	0	0	0.2
CGG	0%	**C**	0.6	0	0.2
AGA	20%	**A**	0.4	0	0.4
AGG	20%	**G**	0	1.0	0.2

***Table II-3**: Arginine set missing one synonymous codon. The CSYN vector lacks data for one dimension while the CPRO vector has all dimensions present because of complementary information.*

Usually the vectors which are atypical because of missing information (generally coming from short sequences or small clusters of transcripts, susceptible to large stochastic variation) need to be removed from the data set because of possible artifactual clustering during multivariate analysis (Perrière and Thioulouse, 2002).

In the cases in which the information provided by a totally missing triplet can be complemented by another synonymous triplet, the CPRO vector corresponding to this data could be kept in the data set. This is the reason why a CPRO-filter removes less elements than the CSYN-filter (see II C.4 for a description of these filtering schemes).

D.1.3 Exposed information, enhanced distance

Some information is *exposed* in CPRO compared to CSYN (in other words, CPRO enhances some differential nucleotide preferences that CSYN treats indifferently).

Assuming these very similar relative synonymous codon distributions for amino acid Arginine:

	CGT	CGC	CGA	CGG	AGA	AGG
dataset A	0.1	0.2	0.2	0.2	0.2	0.1
dataset B	0.1	0.2	0.2	0.2	0.1	0.2
dataset C	0.1	0.1	0.2	0.2	0.2	0.2

***Table II-4**: Three equidistant CSYN vectors for Arginine codons. These vectors are not equidistant from the CPRO point of view.*

The three vectors are equidistant in this 6-dimensional space, and the Euclidean distance between any two of them is equal to 0.141 units.

MultiVariate Algorithms (MVA) would not cluster any of these vectors together:

```
 / | \
A  B  C
```

If the same data is observed from the point of view of codon profile vectors, a different pattern emerges (Table II-5).

	dataset A			*dataset B*			*dataset C*		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
T	0	0	0.1	0	0	0.1	0	0	0.1
C	0.7	0	0.2	0.7	0	0.2	0.6	0	0.1
A	0.3	0	0.4	0.3	0	0.3	0.4	0	0.4
G	0	1.0	0.3	0	1.0	0.4	0	1.0	0.4

***Table II-5**: The three vectors which were equidistant in CSYN analysis are clustered in CPRO, with A and B more similar to each other than to C.*

In CPRO terms, A and B are more similar (distance of 0.141) and would be clustered together, both being at a distance of 0.2 units from C:

```
/\ \
A B C
```

The same applies to the other 6-fold degenerate amino acids.

The examples presented in the sections above (what is shown and what is shielded by the two methods) reveal the different targets of CPRO and CSYN. CPRO from this point of view is not intrinsically better or worse than CSYN. It is instead different: the different point of view enhances some information while hiding other information, in accordance with the nucleotide preferences. A real case scenario in which this difference was observed is presented below (D.4.1). A large scale comparison of Euclidean distances computed with CPRO and with CSYN vectors is reported in the next chapter (section III D.1).

D.2 Complete, extensible: generality

CPRO can accommodate any genetic code, all the present ones (16 different tables, 2 of which equal apart from translation initiation codons) and all the ones that will be discovered or created in the future, without any *ad hoc* adjustment.

CSYN vectors would have labels for the 64 dimensions that depend on the genetic table. Hence 15 different CSYN vector types would have to be managed. Codon usage information coming from organisms with different translation tables (*e.g. Mycoplasma pneumoniae* – translation table 4 – and *Bacillus subtitlis* – translation table 11) would not be straightforwardly comparable because the dimensions in the two vectors would refer to different codons. This problem is not present in CPRO, allowing:

1) the use of CPRO vectors without any adjustment for any genome or sequence coming from any organism. In particular: no need to redefine the labelling of dimensions in multivariate analysis programs

2) the possibility of comparing codon usages across different genetic codes (between organisms with different translation tables). In particular: the possibility of comparing nuclear and mitochondrial genomes of the same organism

3) the possibility of using CPRO also for novel (artificially created or computationally modelled) genetic codes. For example, the following CPRO would come from a modified genetic code with four codons for Methionine (Table II-6)

Codon profile			
Base	*position in the triplet*		
	1	**2**	**3**
T	0	1	0.2
C	0	0	0.2
A	1	0	0.1
G	0	0	0.5

***Table II-6**: Example from a hypothetical modified genetic code having 4 synonymous codons for Methionine.*

D.3 Dimensionality

Codon profile uses a high number of dimensions (vectors of 252 components), but the majority is composed of invariable dimensions. Invariable dimensions are those containing either always 1 or always 0 (according to the genetic code).

These components do not contribute to the actual analyses: they are ignored by MVA algorithms. They also do not contribute to the computation of Euclidean distances because all vectors would have the same value in these dimensions, so the contribution of them is actually zero.

The real dimensionality of CPRO vectors is the same one of CSYN vectors, in the Standard code. Out of the 64 triplets, the three terminator (STOP) codons are usually ignored in compositional analysis (due to their statistical under-representation: only one STOP codon per transcript) and so are the relative frequencies for Methionine and Tryptophan (which would always be 100% as they are both coded by a single triplet in the Standard code). Thus the number of variable dimensions for CSYN vectors is 59. The three 6-fold degenerate amino acids contribute six variable dimensions each to the CPRO vectors (relative usage of the two nucleotides that can appear in first position and relative usage of the four possible nucleotides in third position; actually, there are eight variables for Serine but those for the second position are completely correlated to those in the first). Since the information for the other synonymous sets is equivalent under both schemes, CPRO vectors have also 59 variable dimensions.

Codon profiling does not hence require higher computational power or a larger amount of data than other synonymous codon usage techniques.

It is important here to note that although there are 59 variable dimensions in total, the effective space of synonymous codon usages has a lower effective dimensionality since the triplets are not independent when measured by their relative frequency (a higher usage of one triplet implies a lower usage of the synonymous alternatives). The number of theoretically uncorrelated dimensions in the Standard code is 41 for CSYN and 38 for CPRO (see section VI B.6 from chapter six), with the lower number for CPRO due to the summarisation of codon information for Arginine, Leucine and Serine, by combination of triplet contributions. In practice there are always correlations among the triplets of observed codon usages (both within and between synonymous sets), like the correlation of the triplets contributing to the GC3 content.

The CPRO scheme was developed with a focus on the correlation of triplets according to nucleotide contents and hence combines together the triplets that contribute to the same base-at-position contents.

For example, the difference in the usage of AGR and CGN codons for Arginine was found to be one of the major contributions to the separation of bacteria from archaea (VI D.3.2) and of vertebrates from invertebrates (III D.3.4). AGR codons (AGA and AGG) are the rarest codons in *Escherichia coli* (6% relative usage) and amongst the rarest in many bacterial species (it has been suggested that in these organisms they play a role as modulators, regulating gene expression; Chen and Inoue, 1990; Ohno *et al.*, 2001).

The CPRO approach lowers the effective number of uncorrelated dimensions (from 41 to 38) and hence the maximum number of orthogonal axes that can be found by a multivariate ordination procedure to separate the data (*q.v.* III C.4). Nevertheless, for practical purposes a high number of axes such as 38 or 41 is never used in dimensional-reduction multivariate analyses (such as correspondence analysis or multidimensional scaling) for the following reasons: 1) the first few axes account for the largest fraction of variation (usually the number of axes is chosen so that at least 60% of the percentage of total variation can be accounted for); 2) choosing too high a number of axes increases the variability within the groups more than the variability between groups, thus lowering the usefulness of the procedure in discriminating among labelled sets (Anderson and Willis, 2003); 3) it is difficult to graphically represent a high number of orthogonal axes. In codon usage studies two, three or four axes are commonly used, usually sufficient to account for most of the variation.

If the CPRO vectors can retain useful discrimination power (the ability to separate transcripts or genomes based on the codon usage information) in practical analyses, making it easier to see codon usage patterns, then their summarisation of codon information for the 6-fold degenerate amino acids can be justified.

D.4 Resolving power and sensitivity

In all the analyses presented in this work the CPRO scheme obtained almost identical results to the CSYN one, since the two methods differ only in the treatment of the codon information for three amino acids in the Standard code. In some cases a higher resolving power was noted, namely the ability to discriminate based on codon similarity.

Although CPRO can help the simple visualisation of codon usage patterns, its discrimination power will be in absolute terms lower than the one of CSYN vectors, which contain the information on the relative usages of those triplets whose contributions are summed in CPRO vectors. There is hence a maximum of 38 orthogonal separation axes that can be found in CPRO compared to the 41 in CSYN, as discussed in the previous section.

The Euclidean distance computed between two codon profile vectors is generally greater than the one computed on synonymous codon usage vectors relative to the same data, as exemplified in the study of heterogeneity on prokaryotic genomes (where intra-genomic codon profile distances are greater than those computed with synonymous codon usage vectors; *q.v.* III D.1).

The different scale between the measures obtained under the two vector schemes was taken into consideration during the detection and identification of Horizontal Gene Transfers (described in the IV chapter), lowering the thresholds of codon similarity and atypicality when repeating the codon profile analysis with synonymous codon usage vectors. The codon profile appeared more sensitive in that comparison since, even taking in account the different scale, similarity thresholds had to be relaxed more when using synonymous codon usage vectors to detect the probably transferred regions (*q.v.* IV D.6). Nevertheless, this sensitivity is only slightly higher, and this is not surprising, since it would be due only to three amino acids which are treated differently under the two schemes.

D.4.1 Hierarchical clustering of bacterial genomes

Another case where the codon profile approach produced different results was a clustering performed on the completely sequenced bacterial genomes. The sequenced bacterial genomes were hierarchically clustered according to their total synonymous codon bias and according to their codon profile bias. The whole-genome bias was computed from all the transcripts coming from the CDS (coding sequences) in the genomic entries of EMBL database (Stoesser *et al.*, 2003; the list of genomes and accession codes is reported in III F.1).

Having no pretence of being a taxonomically accurate analysis (as the codon information is not considered sufficient for reconstructing taxonomic relationships and comparative analysis of homologous sequences is best used for this), it originated as a parallel representation of the clustered maps of the prokaryotic codon space (see VI D.3.2). Although the exact branching pattern between the taxa cannot be reliably resolved, nevertheless the clustering shows some consistency at family level: bacteria belonging to the same family (like for example *Rhizobiales*, *Chlamydiaceae* or *Bacilli*) appear to have very similar genomic codon biases and are clustered together (Figure II-5; compare also with the multidimensional scaling map of VI D.3.2, Figure VI-9).

The two diagrams are more or less equivalent, with many bacterial families kept together. One of the biggest differences is the placement of some *Enterobacteriaceae*, which appears to be more in agreement with taxonomical views in the CPRO scheme. In the clustering according to CSYN vectors, the *Y.pestis* strains and *Vibrio cholerae* (two Gram negative enterobacteria; green arrow in the figure) are clustered together with Gram positive Bacilli (near *B.subtilis* and *B.halodurans*). The clustering according to CPRO correctly places these species next to the other Enterobacteriaceae (*E.coli*, *S.typhi*, *S.typhimurium*).

The reason for this result is to be found in the greater distances (for CPRO in relation to CSYN) in various dimensions of the vectors relative to, for example, *Y.pestis* and *B.halodurans*. The greater distances are due to the combined contributions of codons, as discussed above (D.1.3), in the 6-fold degenerate amino acids.

For example, inspecting the contribution of the amino acid Serine (Table II-7) to the clustering under the two vector systems, CSYN considers equally (dis)similar *Y.pestis* and *B.halodurans* (which belong to different bacterial families and Gram stain groups) or *Y.pestis* and *E.coli* (both Enterobacteriaceae). The contribution of Serine to the total Euclidean distance between the genome vectors is 0.101 for *Y.pestis* to *B.halodurans* and 0.102 between *Y.pestis* and *E.coli*. Using CPRO vectors, combining the contributions of the triplets in the base composition point of view, *B.halodurans* becomes clearly more distant from *Y.pestis* (0.177 distance units for the dimensions relative to Serine) while *E.coli* and *Y.pestis* are more similar (0.122 distance units) and hence get clustered together.

	TCT	**TCC**	**TCA**	**TCG**	**AGT**	**AGC**
B.halodurans	0.175	0.142	0.201	0.159	0.153	0.170
E.coli	0.146	0.149	0.124	0.154	0.151	0.277
Y.pestis	0.158	0.118	0.173	0.114	0.207	0.230

***Table II-7**: CSYN contents for Serine codons in three bacterial genomes.*

The different treatment of triplets contributions produces different separation. The synonymous triplets are not considered as equal possibilities but are summed according to their contribution towards base composition.

Another interesting aspect that can be noticed in the hierarchical clustering is the placement of *T.tengcongensis*, which is a Gram negative anaerobic but is reported to have 60% sequence similarity with *B.halodurans* (a Bacillus, hence a Gram positive). In the hierarchical cluster this is confirmed, with *T.tengcongensis* close to *B.halodurans*, in both CPRO and CSYN clusterings.

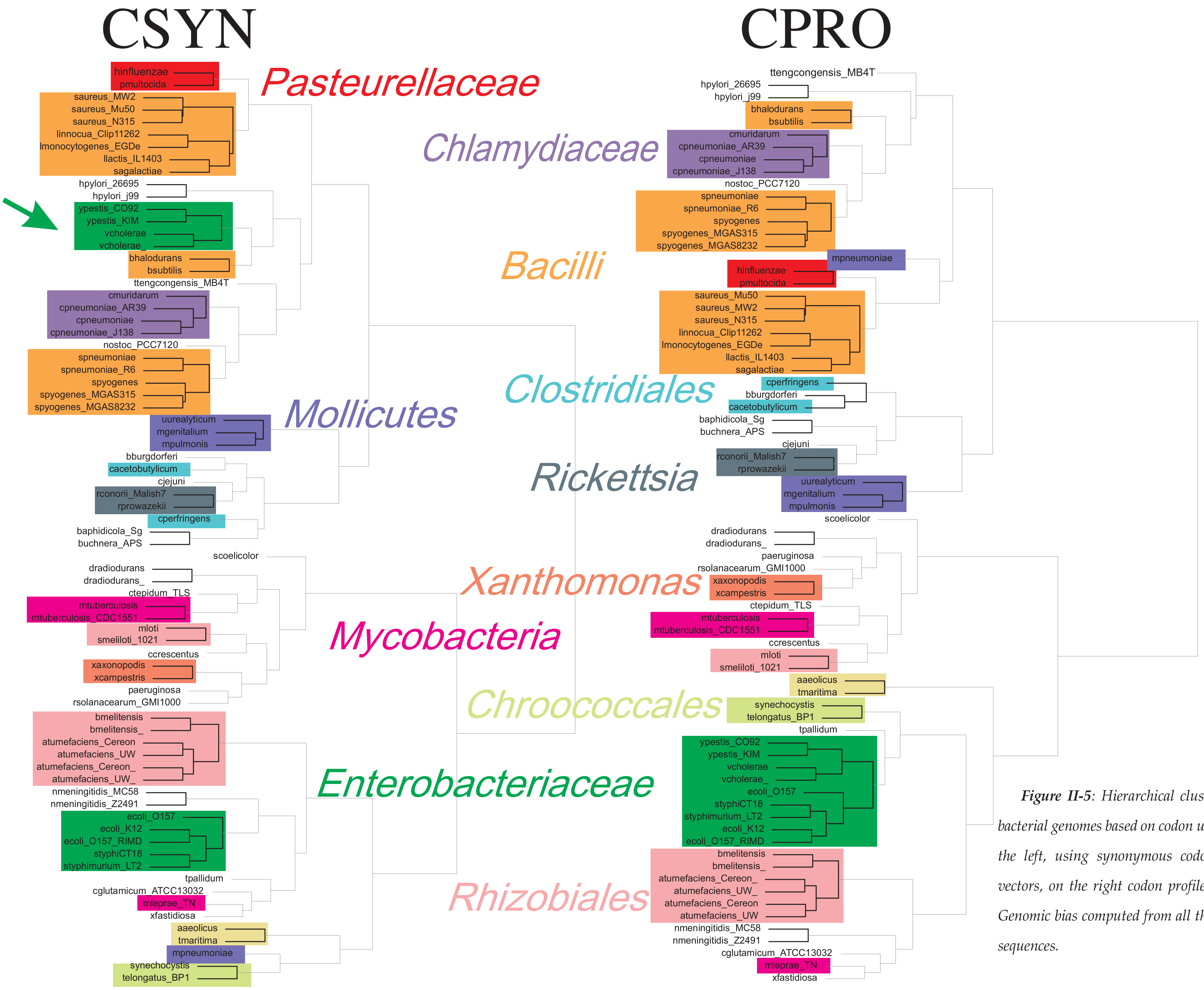


***Figure II-5**: Hierarchical clustering of bacterial genomes based on codon usage. On the left, using synonymous codon usage vectors, on the right codon profile vectors. Genomic bias computed from all the coding sequences.*

E CONCLUSIONS

Apart from the 6-fold and 8-fold degenerate amino acids, codon profile analysis (CPRO) behaves exactly like synonymous codon usage analysis (CSYN). For those amino acids, it instead shows a slightly different picture. Some distinct CSYN vectors are equivalent from the point of view of CPRO; some vectors which are treated in the same way under CSYN are differentiated under CPRO. Thus great care should be exercised when using codon profile vectors; whenever possible analyses should be carried out using both approaches, comparing the results. Sometimes the codon profile combination of triplet contributions could hide important differences, while in other cases it would enhance the bias, and hence be preferable.

The CPRO point of view is more biased towards genomic nucleotidic preferences and less towards individual triplet preferences, and this is one reason for adopting it in large scale genomic studies, like those presented in this dissertation. Additionally, its generality and coherent scheme makes it a very suitable and extensible tool for large scale genomic analyses. For example it allows comparison of nuclear and mitochondrial genomes, or the analysis of all bacteria together, including those with non standard genetic code.

A number of tools to compute, compare and display codon biases and codon similarity have been developed and their use will be presented throughout the following chapters.

F APPENDIX

F.1 Web services and programs

Various public services to perform codon profile and synonymous codon usage analysis have been set up as web-interfaced tools and are accessible at the internet address: *http://www.ebi.ac.uk/~insana/codonprofile/*. Submitting as input either a transcript sequence (or a concatenation of sequences) or an entry from the *CUTG* database (Nakamura *et al.*, 2000), the user can retrieve the result of a series of calculations based on codon usage. The calculations included are: codon profile vector, synonymous codon usages, nucleotide contents (total or for individual triplet positions) and amino acid relative frequencies. Various forms of displaying the results can be used, including codon difference matrices.

The programs used in this work to calculate, manipulate, visualise and characterise the codon information are, with the obvious exception of those explicitly mentioned and cited in the Methods sections, scripts written in the *Perl* (http://www.perl.org/) programming language by the author; they are available upon request.

F.2 Databases

Codon profile vectors, synonymous codon usage frequencies, amino acid relative frequencies, position specific base propensities, total nucleotide contents and other similar calculations, which were performed on all completed genomes and on their transcripts, are available in flat-file format at the above mentioned internet address.

III *Genomic heterogeneity*

A Abstract

The sequence composition of genomes displays species-specific frequencies with genome-specific preferential codon usages. This can be used for hierarchical classification, screening of Horizontal Gene Transfer events and studies on biodiversity. On the other hand, codon usage can vary substantially among the genes within one genome, and average codon biases often conceal the intra-genomic differences.

A study of codon heterogeneity was hence performed, computing the variability of codon usages inside a genome and providing the statistical background for the subsequent analyses and a scale for the measures of codon usage similarity.

The distributions of the codon usage of all the completed genomes have been plotted, showing a consistent range of intra-genomic variability and the amount of atypical transcripts, which are the transcripts with codon usage significantly different from the genome bias.

B Introduction

In 1980 the precursor work of Grantham and co-workers revealed a high degree of consistency between the preferential usage of codons among genes of the same or similar organisms. One of the first observations was that viruses and mammals have widely separate coding strategies. The descriptive hypothesis they stated was *"all genes in a genome, or more loosely genome type, tend to have the same coding strategy"*. This was called the *genome hypothesis* and suggested that each type of genome preferentially employs a certain subset of the genetic code, using it differently from other kinds of species, with choices among synonymous triplets being consistently similar among its genes. Following research confirmed the initial finding and investigated the possible causes of such species-specific patterns (from translational efficiency to mutation biases, as overviewed in the first chapter). It even appears possible to capture whole-genome characteristics with compositional tools and predict the genome of origin of a genetic sequence from this species-specific pattern (Kanaya *et al.*, 1999; Sandberg *et al.*, 2003).

After some years, with more sequence data available, it became clear that most species displayed also considerable intra-genomic difference, with codon usage found to vary substantially among the genes within one genome. To prevent concealment of the underlying heterogeneity, averaging of codon usage over all the genes was discouraged and the trends in the variability of genes and classes of genes inside a genome began to be investigated (Sharp *et al.,* 1988). The differences between classes of genes with highly biased usage of synonymous codons and others with more even usage attracted considerable research.

To verify the relative merits of the two opposite (but complementary) views, the intra-genomic heterogeneity was studied in several domains for which complete genome sequences are available: archaea and bacteria, human infecting viruses and some animal genomes. In addition to a better picture of the codon heterogeneity, the results of the analyses provided the statistical background for the subsequent investigations (presented in the following chapters) and the scale on which to compare the measures of codon similarity.

C METHODS

C.1 Completely sequenced prokaryotic genomes

Codon information from all available archaeal and bacterial genomes (see appendix III F.1 for complete list and accession numbers) was computed in the form of codon profile vectors for all the individual transcripts – annotated coding sequences – and for the whole genomes. The coding sequences were obtained from EMBL database entries (Stoesser *et al.,* 2003) using the *coderet* program from the EMBOSS package (Rice *et al.*, 2000).

Only transcripts that encode at least one of each amino acid species were analysed, removing the others from the data set (application of the AA-filter, see II C.4). 140,207 transcripts out of a total of 227,434 (61.6%) were kept in the data set and this resulted in codon-dissimilarity distributions of the same skewedness but with less deviation, lower average and shorter tail. Relatively more transcripts are dropped for archaea than for bacteria.

The three amino acids which contributed the most to the removal of transcripts are Cysteine (accounting for 54.5% of the removed transcripts), Tryptophan (20.5%) and Histidine (7.5%).

For each genome, the Euclidean distance (see II C.2) between all the transcripts and the genome average was computed and the resulting distributions of distances were plotted in the form of histograms and in boxplot representations.

C.1.1 Boxplots

A boxplot is a way of visualising one-dimensional data presenting the distribution of values in a more compact way than histograms. This is particularly useful when comparing two or more sets of sample data. Differences in the medians and spreads of the datasets are clearly visible with a boxplot. It gives a picture of the symmetry of a dataset, and shows statistical outliers very clearly.

A boxplot comprises the following elements:

1) A central box within which half of the data lies. The central box is bounded below by the first quartile (also called the $x_{0.25}$ *quantile*: the middle number in the first half of the data set) and above by the third quartile ($x_{0.75}$). A central line marks the median.

2) Two protruding lines (*whiskers*) extending from the central box. The commonly accepted method for drawing the whiskers prescribes a maximum length for each whisker of 1.5 times the interquartile range (*IQR*). The whisker above the third quartile can reach the largest data value that is less than (or equal) to the value being 1.5 IQRs above the third quartile.

3) *Outliers* marked individually: those data points lying beyond the whiskers.

C.2 Human infecting viruses

Complete genome sequences of human infecting viruses were obtained from GenBank (Benson *et al.*, 2000). Removing those that do not contain any *CDS* (coding sequence) information and keeping (for brevity) only the sequence of four out of the 76 strains of papillomavirus, left 39 complete viral sequences (appendix III F.2 reports the list and accession numbers). These contained 1,318 transcripts, of which 304 were removed by *AA-filtering* (II C.4).

The remaining transcripts were used to compute codon profile distances from *self* and from *human*: the distances between the transcripts to their own genome bias and the distances to the human genome bias, respectively.

The distributions of these two groups of distances were plotted side by side using the boxplot representation.

C.3 Completely sequenced eukaryotic genomes

C.3.1 Genomes from the Ensembl project

Transcript sequences for eukaryotic genomes were obtained from Ensembl (Hubbard *et al.*, 2002; http://www.ensembl.org), a project that provides automatic annotation for a number of eukaryotic genomes, including the human one, which started the project. The analysed genomes, and their Ensembl release version, are: human (*Homo sapiens*, 8.30a.1), pufferfish (*Takifugu rubripes*, 8.1.1), mouse (*Mus musculus*, 8.3c.1) and mosquito (*Anopheles gambiae*, 8.1b.1).

The transcripts were grouped according to the annotation from the *Tribes* protein family clustering algorithm (Enright *et al.*, 2002). Families containing less than 5 transcripts were discarded. Additionally, the most restrictive filtering, *CSYN filtering* (see II C.4) was applied, keeping only those clusters with a complete set of all codons. The total number of protein families kept in the analysis is higher than 80% of the total number for the vertebrates and around 71% for the mosquito genome (Table III-1).

Genomes **Families**	*Homo sapiens*	*Takifugu rubripes*	*Mus musculus*	*Anopheles gambiae*
Total	1012	1414	1050	762
Analysed	824	1261	881	540

Table III-1*: number of total and analysed transcript clusters for each Eukaryotic genome. Those clusters not comprising a complete set of all codons were discarded.*

C.3.2 The fly genome

The heterogeneity of the mosquito genome has been compared to the genome of the main model species for insects: the fruitfly *Drosophila melanogaster*. At the time of the analysis, this genome was not yet in the Ensembl project, so the transcripts sequences were obtained from the website of FlyBase (The FlyBase Consortium, 2002;

http://www.flybase.org/). 5,016 coding sequences from the fruitfly genome satisfy the CSYN-filter and were compared to 1,386 CSYN-filtered transcripts from the mosquito genome.

C.4 Multivariate analysis

The study of synonymous codon usage is a high-dimensional data analysis problem, as it involves the simultaneous investigation of the contributions from all the triplets. Multivariate analysis has hence been frequently used to study codon usage (Grantham *et al.*, 1980; Holm, 1986; Shields and Sharp, 1987; Médigue *et al.* 1991; McInerney, 1997; Kanaya *et al.*, 1999; Lafay *et al.*, 2000; Kanaya *et al.*, 2001a).

The major trends accounting for the variation among codon usages were studied using correspondence analysis (CA; Benzécri, 1973; Greenacre, 1984) and multidimensional scaling (MDS; Cox and Cox, 1994), two methods that project high-dimensional information onto low-dimensional spaces.

In fact, these methodologies yield a series of ordered orthogonal axes (also referred to as factorial axes) that account for smaller and smaller proportions of the original variance present in the data set.

Both are unconstrained ordination procedures, in that they do not use *a priori* hypotheses but they reduce dimensions according to general criteria, such as maximizing dispersion or keeping distances in the new dimensional space equal to the original distances. These unconstrained procedures are extremely useful for the visualisation and discovery of broad patterns across the set of points in a multidimensional space; in particular, where the data is classified into labelled groups, they enable the visualisation of potential patterns of relative dispersion or location differences among the groups.

For both methods it is possible to estimate the accuracy of the representation, namely the amount of variation information retained when lowering the number of dimensions, which can be approximated by the cumulative percentage contributions from the eigenvectors associated with the projection to the low-dimensional space:

$$\frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

where l is the number of axes chosen for the low-dimensional representation, n is the total number of positive eigenvalues and λ_i are the eigenvalues sorted in descending order (i.e. $\lambda_1 > \lambda_2 \ldots > \lambda_n$).

Since smaller eigenvalues contribute much less weight to the total distance between the points, these can be usually truncated with less error for low-dimensional display.

The total number of positive eigenvalues is related to the number of uncorrelated dimensions: the number of orthogonal axes needed to account for the total variation among the points of the multidimensional space. Both procedures assess the departure from a null hypothesis of no dependence between the original dimensions of the data. If there is no correspondence, the number of orthogonal axes needed to account for all the variation among the points is equal to the number of original dimensions.

It is also important to note that, when looking for patterns among labelled groups of data in a low-dimensional overview, choosing too many axes has (apart from the difficulty in the visualisation) the drawback of increasing, after a certain number of axes, the intra-group variability compared to the inter-group one, and hence diminishing the ability of discriminating among the groups (Anderson and Willis, 2003).

The obvious shortcoming of these low-dimensional representations is the loss of the individual variate values. To overcome this, the low-dimensional data overviews need to be integrated with other techniques to recover more of the multivariate information. Unfortunately most visualisation and analysis techniques are limited in their practical use by both the dimensionality and the amount of the multivariate data (Wong and Bergeron, 1997). For example, a scatterplot matrix (an array of panels presenting pairwise adjacent scatterplots) of all 59 triplets of a synonymous codon usage would require 1711 plots.

In almost all low-dimensional plots presented in this work, multidimensional scaling (also known as *principal coordinate analysis*; Gower, 1966) was chosen for presenting the results because it generally had a higher amount of variation explained with less axes, making it particularly useful for two-dimensional data representation. Additionally, MDS preserves the Euclidean distances, which were chosen as the dissimilarity measure between codon usages. In fact, in the plots produced after multidimensional scaling the distances between the points in the plot reproduce the dissimilarities between the points in the high-dimensional space. In other words, the larger the dissimilarity between two points in the high-dimensional space, the farther apart they should be in the low-dimensional representation.

The values appearing at the tick marks on the axes of a multidimensional scaling plot represent the variation along the two (in the case of two-dimensional plots) principal coordinates. Unless no principal coordinate can be found (as in the case of random distributions), the x-axis has a higher range of values, indicating a greater separation of the data along that axis. This is the case for all the multidimensional scaling plots appearing in this work, where the first principal coordinate is related to G+C content.

The *R* statistical computing environment (Ihaka and Gentleman, 1996) was used to perform these multivariate analyses, which are implemented by the functions *cmdscale* (library *MVA*) for multidimensional scaling and *ca* (library *multiv*) for correspondence analysis.

D RESULTS AND DISCUSSION

D.1 Completed prokaryotic genomes

The transcript sequences of all the completely sequenced archaeal and bacterial genomes were compared in terms of codon composition with the average codon bias of their genome. The distribution of the codon similarity values (measured using Euclidean distance in the codon profile vector space) was analysed for each genome in order to determine the amount of variability in codon composition. Figure III-1 and Figure III-2 show these distributions for archaea and bacteria, respectively, in boxplot (*box-and-whiskers plot*) representation.

All the distributions share a similar range and shape: they are all skewed toward lower values and they exhibit a long tail for the higher values. In almost every case, the 75% of the distances, the $x_{0.75}$ quantile, falls below 1.5, and all minima are around 0.4.

The long tails of the distribution show that all genomes have some transcripts (less than 5%) with highly atypical codon usage.

If codon similarity is computed with synonymous codon usage (CSYN) vectors, the distributions are equivalent but shifted in scale. This is due to the fact that CSYN Euclidean distances are generally lower than codon profile (CPRO) Euclidean distances of the same data, because of the combined contributions of triplets in CPRO which leads to enhanced differences (as explained in II D.1.3).

Comparing the Euclidean distances of all transcripts shows that CPRO distances are, on average, around 4.5% (slightly more, 4.68%, for bacteria than for archaea, 4.31%) greater than CSYN ones. This is a constant trend for all genomes (see Figure III-4). Although the differences between the distances in the two vector schemes are restricted to between two and six percentage points of difference for the majority of transcripts, there are several transcripts in which the treatment of sixfold degenerate amino acids results in a CPRO Euclidean distance as much as 30% greater than the corresponding CSYN distance.

There are also negative differences: in some cases CPRO has lower similarity values (up to 10% lower) than CSYN. These are those cases in which the synonymous triplets are not differentiated in CPRO, because contributing to the same nucleotide contents, but differently in CSYN (as in the extreme case shown in II D.1.1).

In fact, the base orientated approach does not equally consider the synonymous triplets in the 6-fold degenerate sets but combines them according to their contribution to the genomic nucleotidic composition (see the previous chapter for a discussion on the different treatment of 6-fold degenerate codon information between CPRO and CSYN).

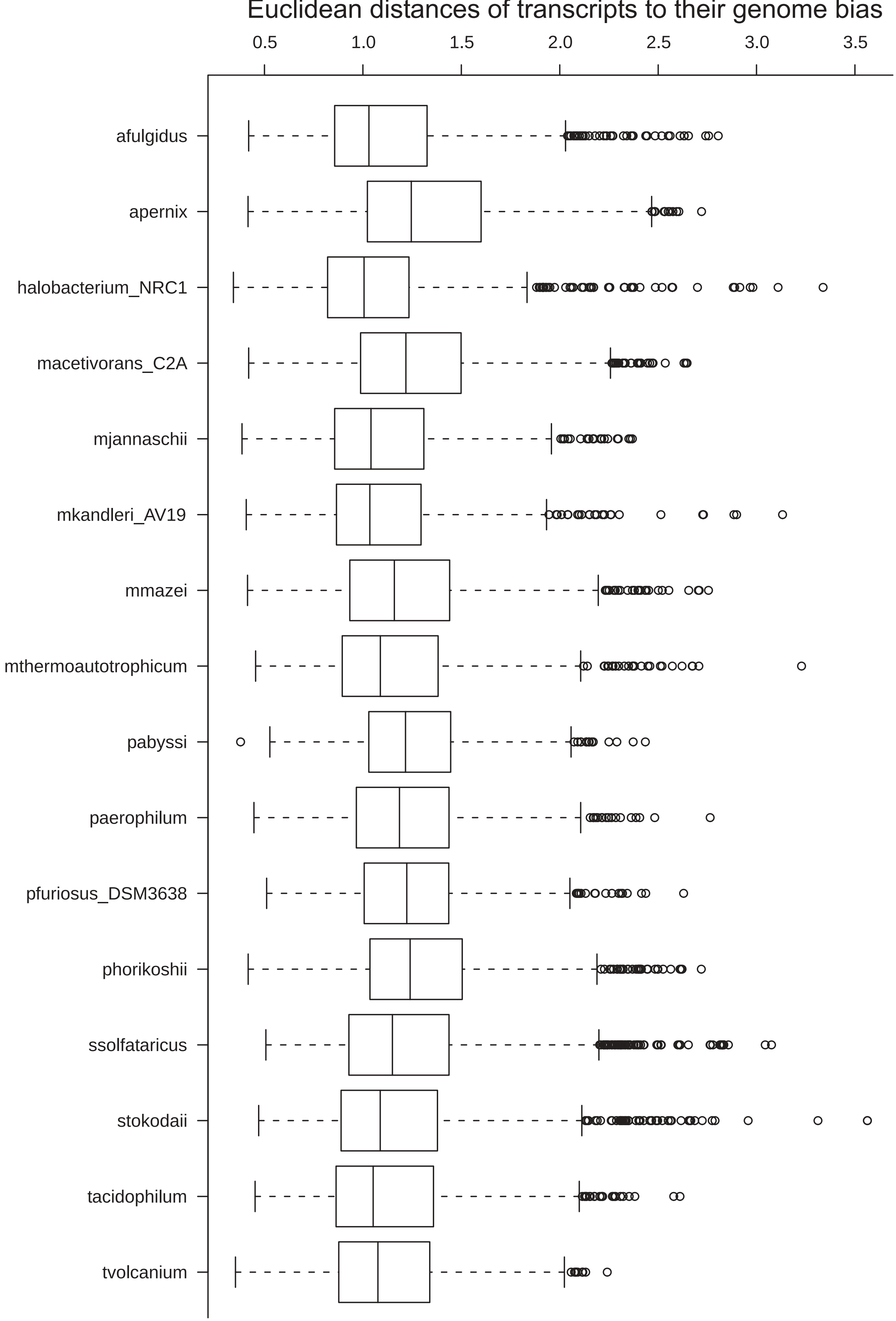


Figure III-1*: Distributions of the codon similarity of transcripts to the genome biases for archaea, in codon profile Euclidean distances (list of complete genome names and accession numbers in appendix F.1).*

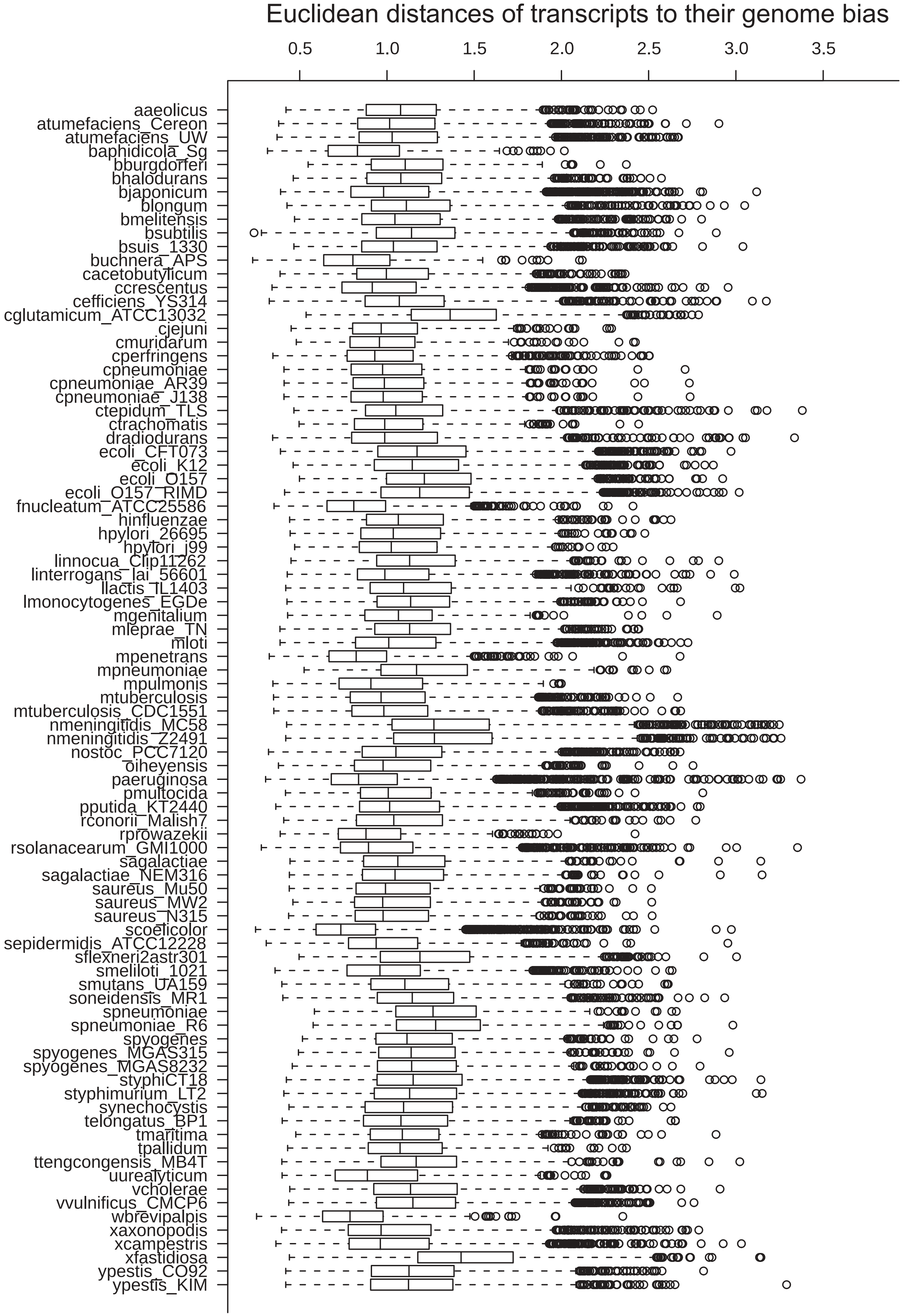


***Figure III-2**: Distributions of the codon similarity of transcripts to the genome biases for bacteria, in codon profile Euclidean distances (list of complete genome names and accession numbers in appendix F.1).*

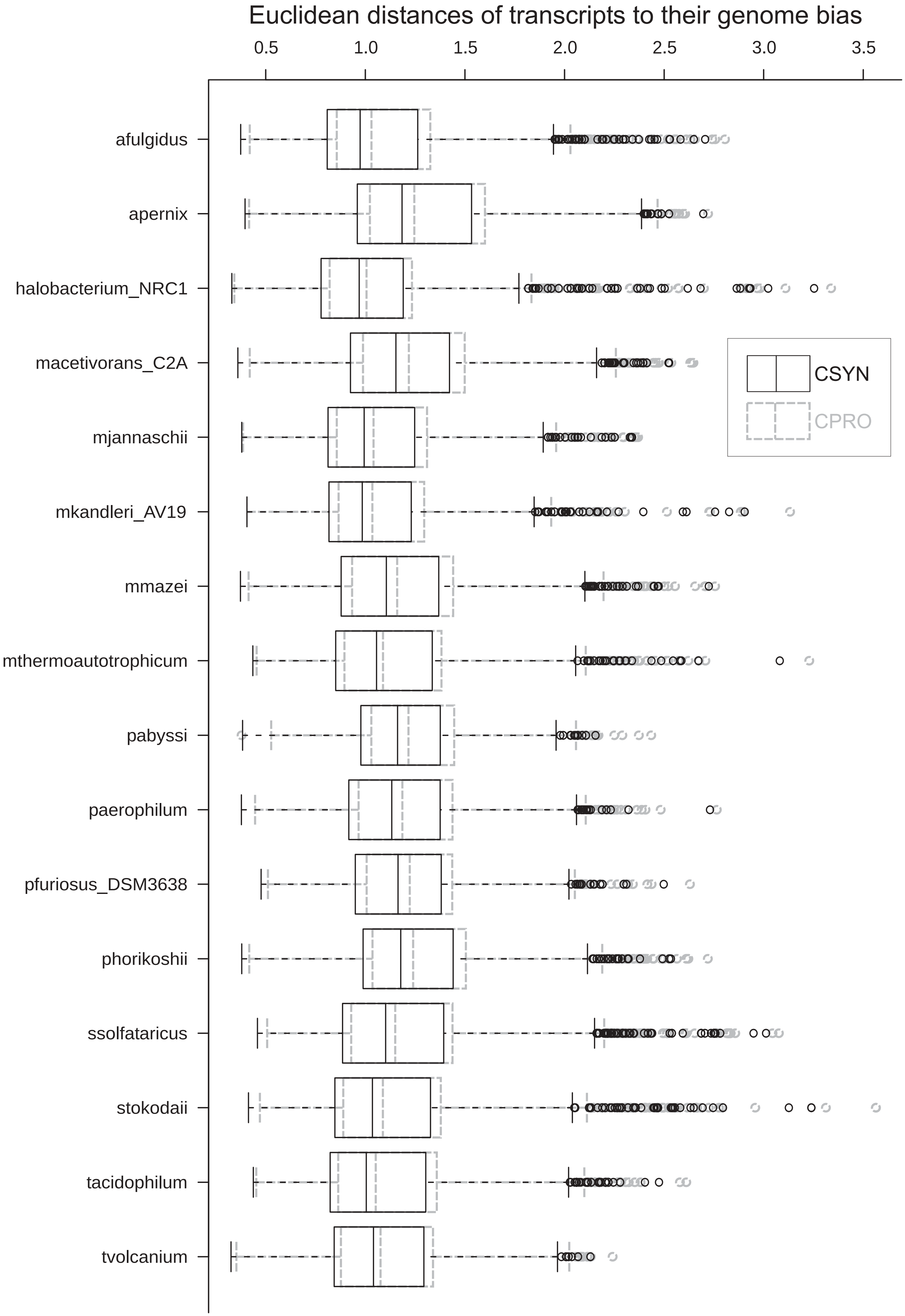


***Figure III-3**: Distributions of the codon similarity of transcripts to the genome biases for archaea, in synonymous codon usage Euclidean distances (in black) compared to the codon profile Euclidean distances (underneath, in gray) from Figure III-1.*

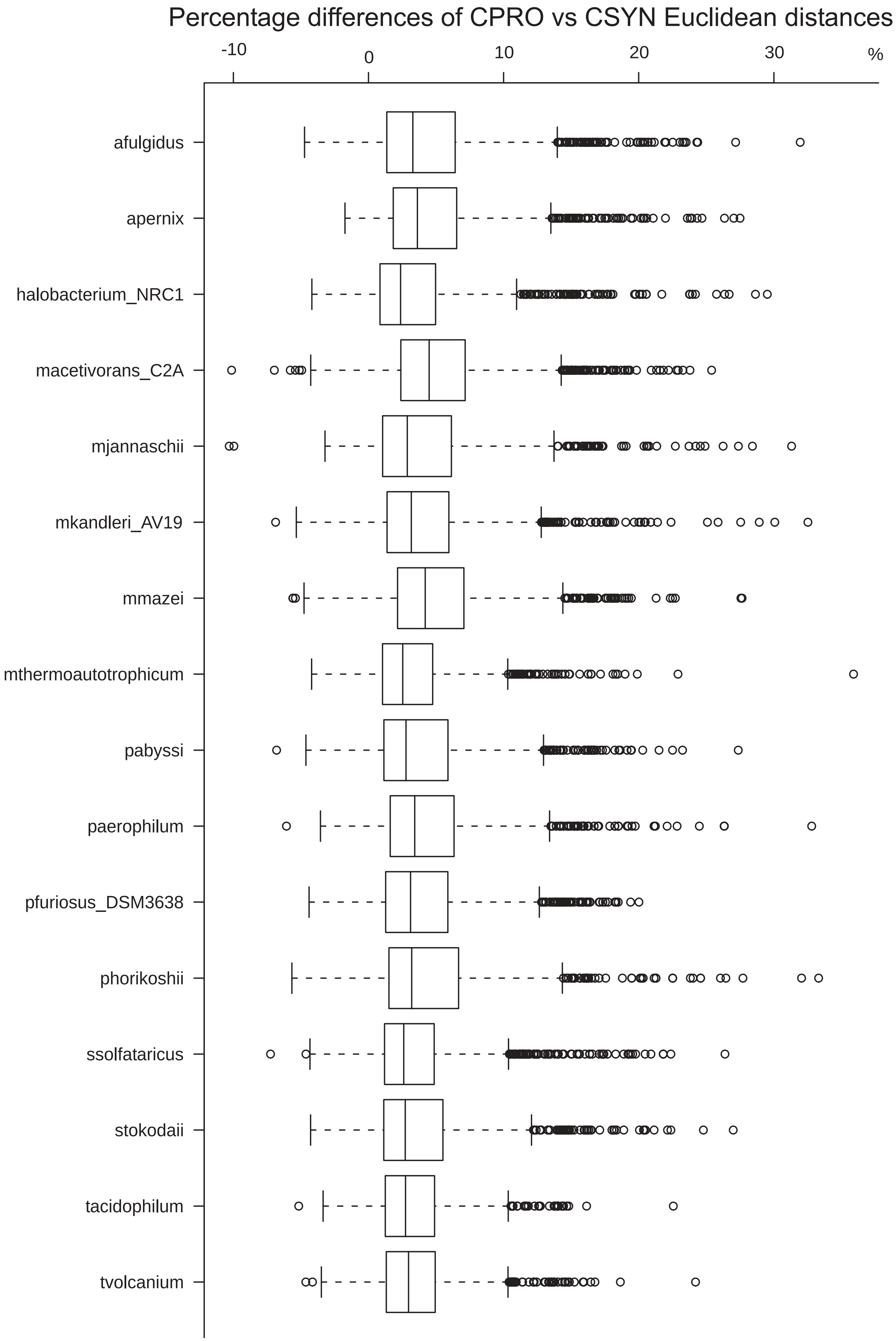


Figure III-4: *Comparison between CPRO and CSYN Euclidean distances of transcripts to genome biases for archaea. The percentage differences between CPRO and CSYN vectors oscillate around 4.3% for the majority of transcripts but for a number of them they can result as much as 30% greater under the CPRO triplet combination scheme.*

D.2 Human infecting viruses

Many of the human infecting viruses that have been completely sequenced were analysed for their intra-genomic heterogeneity and additionally their codon usage was compared to that of their host, *viz. Homo sapiens*.

For each transcript, the distance to *self* and to *human* were computed. The former refers to the difference in codon usage between the single viral transcript and the viral average codon bias; the latter is the difference between the viral transcript and the human codon bias (a short distance indicates similar codon usage to the human average one).

A comparison of the distributions of these distances shows that the distances to human are always higher than those to self. What had been repeatedly observed for whole genomes is now confirmed at the level of single transcripts: the majority of the viruses have a codon usage quite different from that of the host they infect, with distributions of distances to human almost completely above the maximum of the corresponding distributions of distances to self (Figure III-5). There are almost no viral transcripts with a codon similarity to the human codon bias lower than 1.0 units, with the distances to human mainly centred at 1.5 or 2.0 units of distance.

The prominent exceptions are represented by the *adenoviruses*, in particular *adenovirus type B*, some *herpes* viruses and *papilloma type 2a*. In these viruses the distances to the host genome are comparable and in some cases shorter than the distances to self.

The majority of *self* Euclidean distances are distributed between 0.5 and 1.5 units. The analysed viruses appear hence not dissimilar as a whole from the bacteria or the archaea, but they are characterised by higher inter-genomic variability (represented by the large spread in their distance distributions): some viruses encode transcripts with a very homogeneous codon set (*e.g. hepatitis E*, *parainfluenza*) while other viruses employ very diverse codon usages in their transcripts (*e.g. coronavirus*, *herpesvirus 5*, *immunodeficiency* viruses).

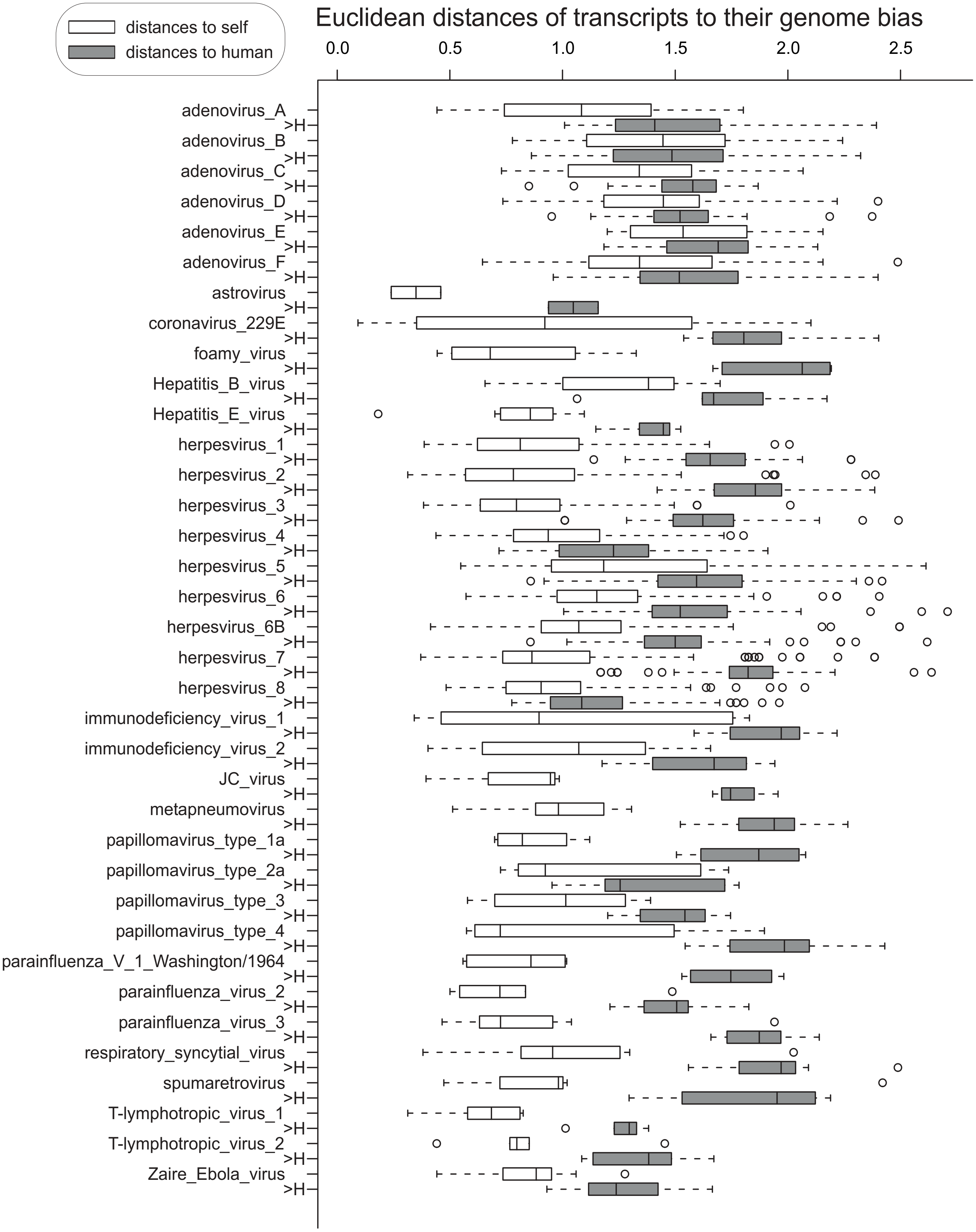


Figure III-5: Comparison of codon profile Euclidean distances of viral transcripts to their own genome bias and to the human one. ">H" indicates distances to the human bias.

D.3 Eukaryotic genomes from the Ensembl project

For each genome, boxplots and multidimensional scaling plots of the protein family clusters were created on the basis of synonymous codon usage vectors and codon profile vectors. In multidimensional scaling plots the major component of separation along the x-axis can be ascribed to G+C content (with principal contributions from GC3 and GC1). The shown plots refer to codon profile vectors. Unless stated otherwise in the text, the results are the same (same separation and topology of the distributions) if synonymous codon usage vectors are used. The analyses expose the intra-genomic variability and allow the identification of the most deviant families or the possibility of inter-genomic classification.

D.3.1 *Homo sapiens* and *Takifugu rubripes*

Figure III-6 shows the results of multidimensional scaling on the codon profile vectors corresponding to human and pufferfish transcripts clustered by the *Tribes* protein family classification algorithm. The pufferfish genome appears more compact than the human one, reflecting lower heterogeneity in the codon usage of its transcripts and transcript families. This is also made obvious in the boxplot comparisons of section D.3.4 below (*q.v.* Figure III-11).

The plotted distribution of GC3 shows that coding sequences have mainly values of 60–80% (Figure III-7), with G+C content being around 54%, much higher than the genome-wide G+C content reported in the recent pufferfish genome paper (Aparicio *et al.*, 2002), where it is shown to be around 43–44% with a very narrow distribution. This indicates that the intra/inter-genic regions of the pufferfish have a much lower G+C content than the coding sequences, and this would balance the observed high 54% G+C content of the coding part of the genome to the reported genomic 43–44%.

The distribution of codon similarity for human families is here reported in histogram representation (Figure III-8). It also appears in boxplot representation in Figure III-11, where it is compared to the distributions for single transcripts and unfiltered families for the human genome and for the other Ensembl genomes analysed. Human protein families were analysed in relation to human infecting viruses and the results are presented in chapter five (V D.1).

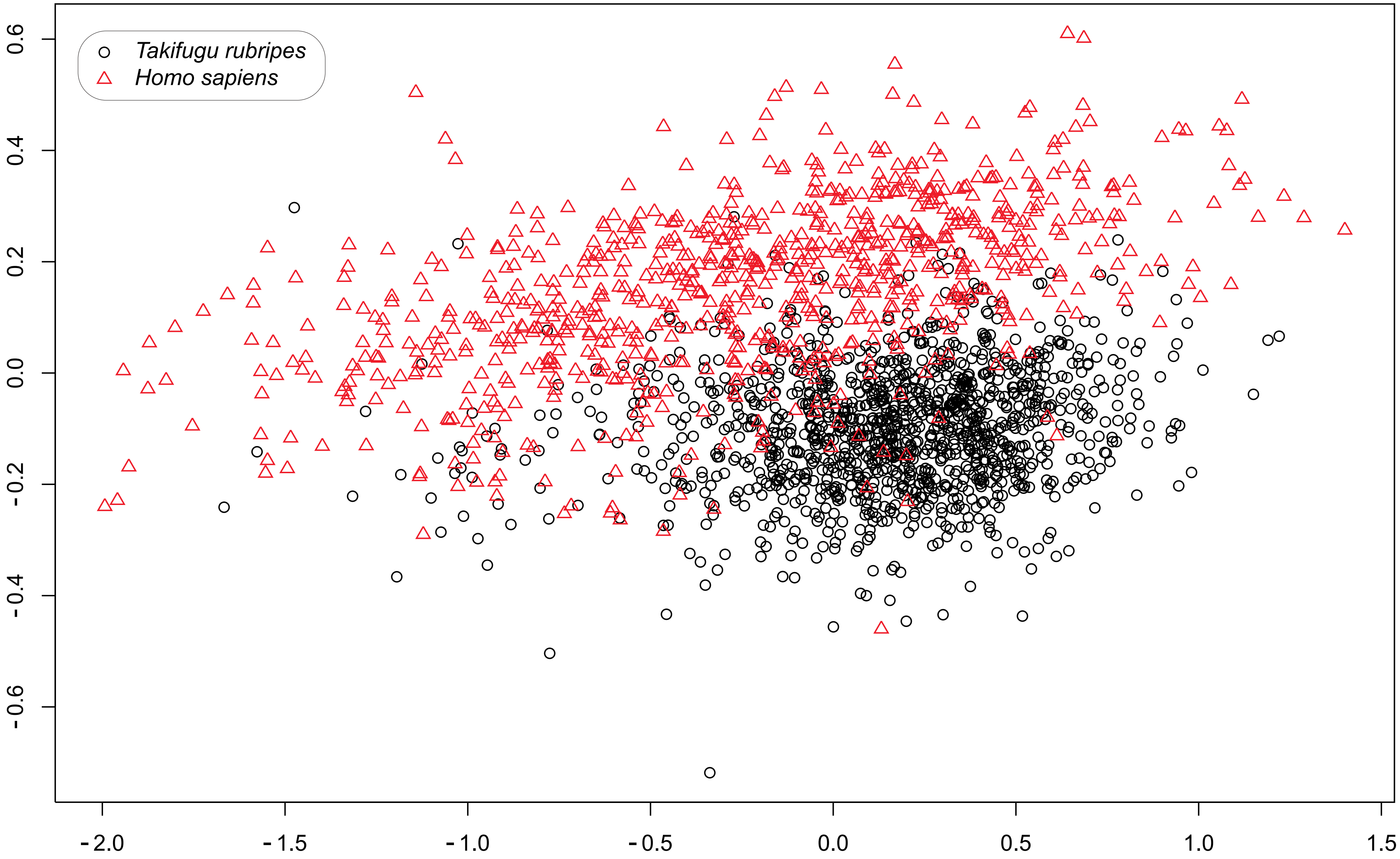


Figure III-6*: Multidimensional scaling plot of human and pufferfish transcript families.*

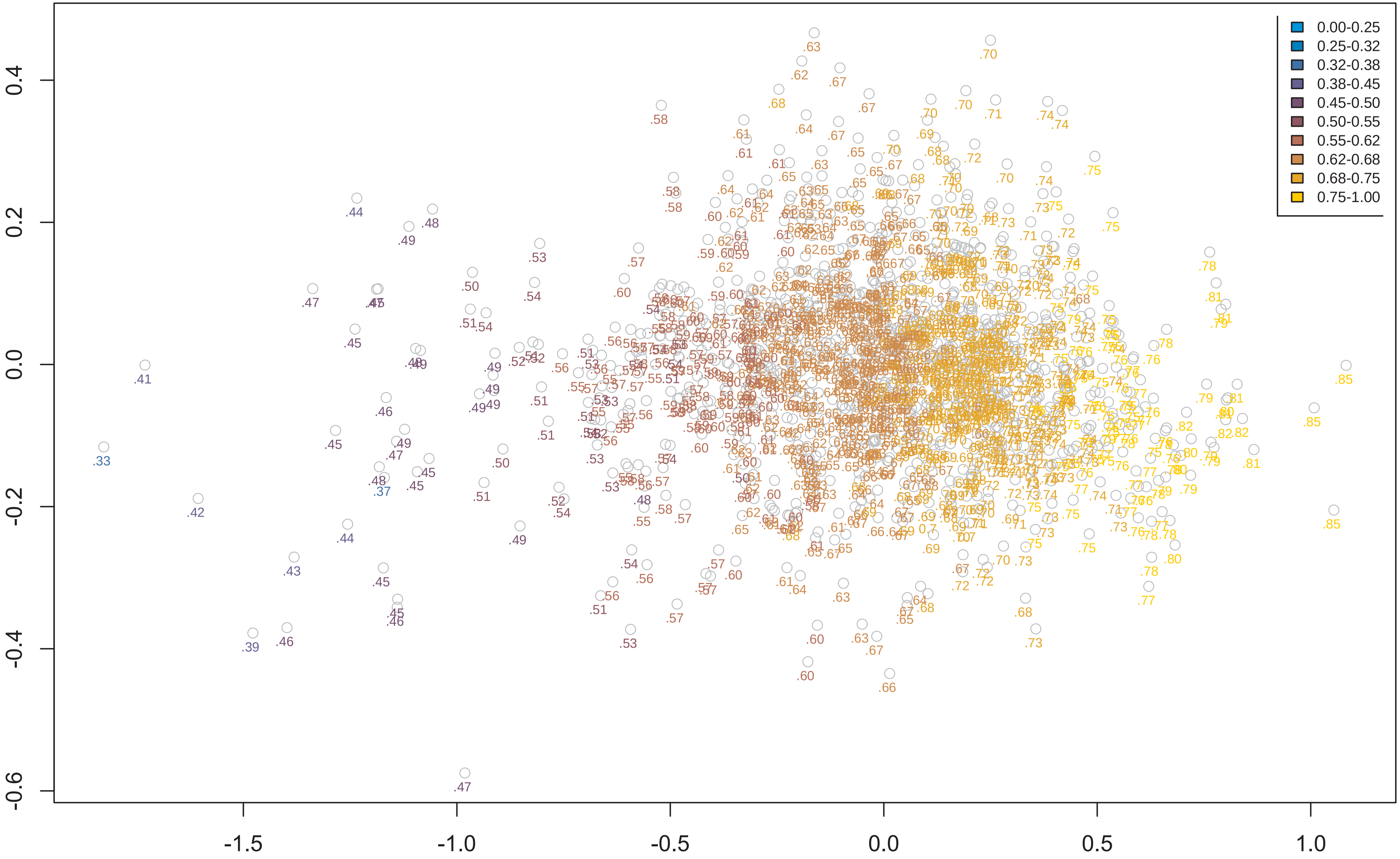


***Figure III-7**: Multidimensional scaling of pufferfish transcript families with indicated values of GC3 content.*

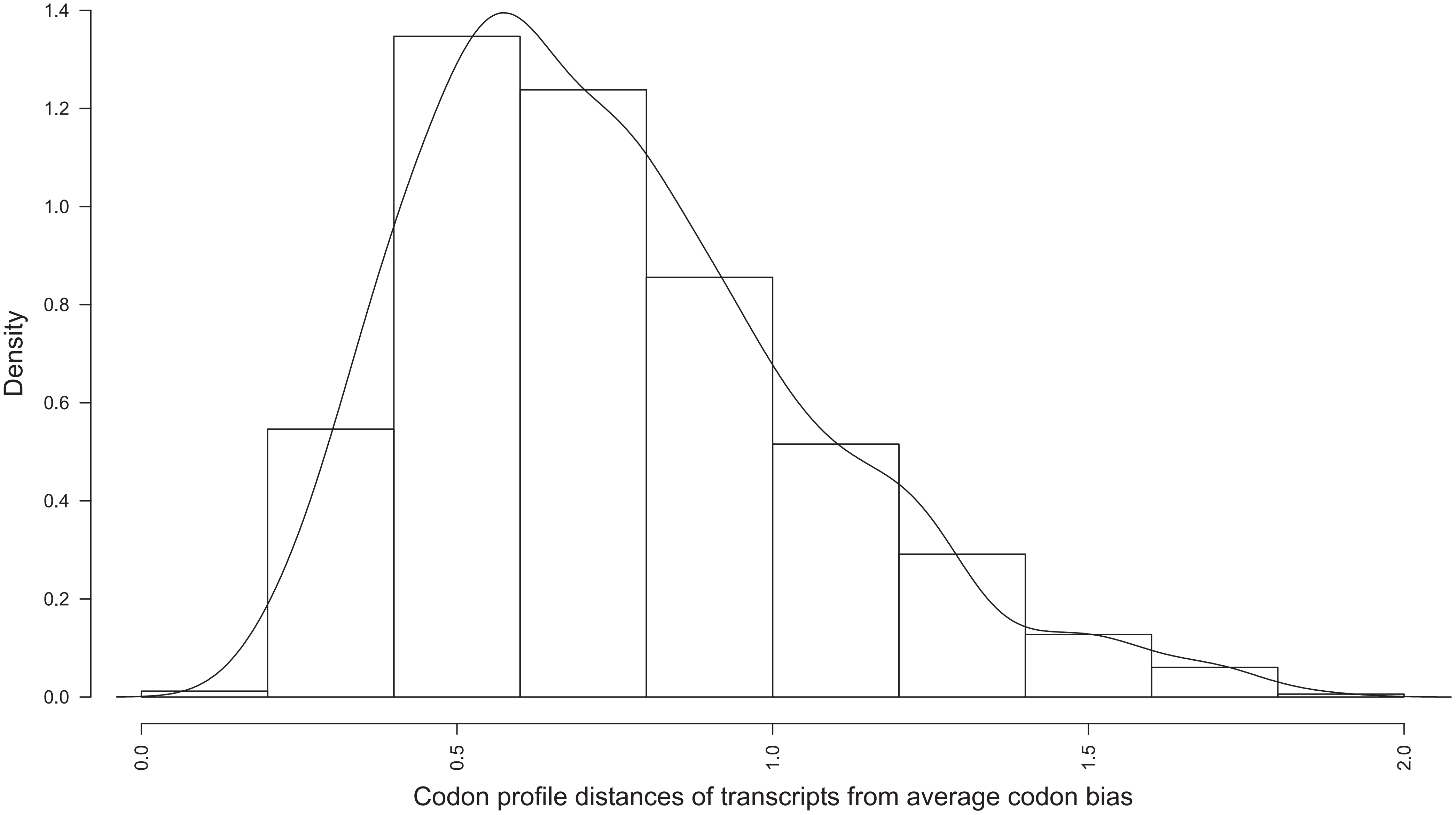


Figure III-8*: Histogram representation of the distribution of Euclidean distances of human transcript families from human codon bias.*

D.3.2 *Mus musculus*

As the chimpanzee genome sequencing project nears completion (the sequencing began in January 2003), the closest Eukaryote to human which has been sequenced and is available from the Ensembl project is *Mus musculus*, the mouse.

A comparison of mouse families with human families shows that they are very similar in codon usage and in the intra-genomic heterogeneity (although the distribution of codon usages from mouse is more compact than the one from the human genome). There is very little possibility to discern human transcripts from mouse ones according to codon usage (Figure III-9 *a*). In fact, the average genome biases of these two species differ by only 0.146 Euclidean distance units and the average distance between all human and all mouse families is 0.968 units. Neither multidimensional scaling (MDS) nor correspondence analysis (CA) can discriminate between the transcript families from the two genomes. The first component of separation between the points is mainly due to G+C content and accounts for over 50% of the total variation (58% in MDS, 54% in CA). The second axis (accounting for 5%) is particularly associated to Arginine codons (with higher usages of AGR codons, over 60%, for the points towards the bottom side of the plot) and to the CG3 content of two-fold degenerate amino acids, while the third axis (4% relative weight; figure not shown) separates according to the usage of Cysteine triplets.

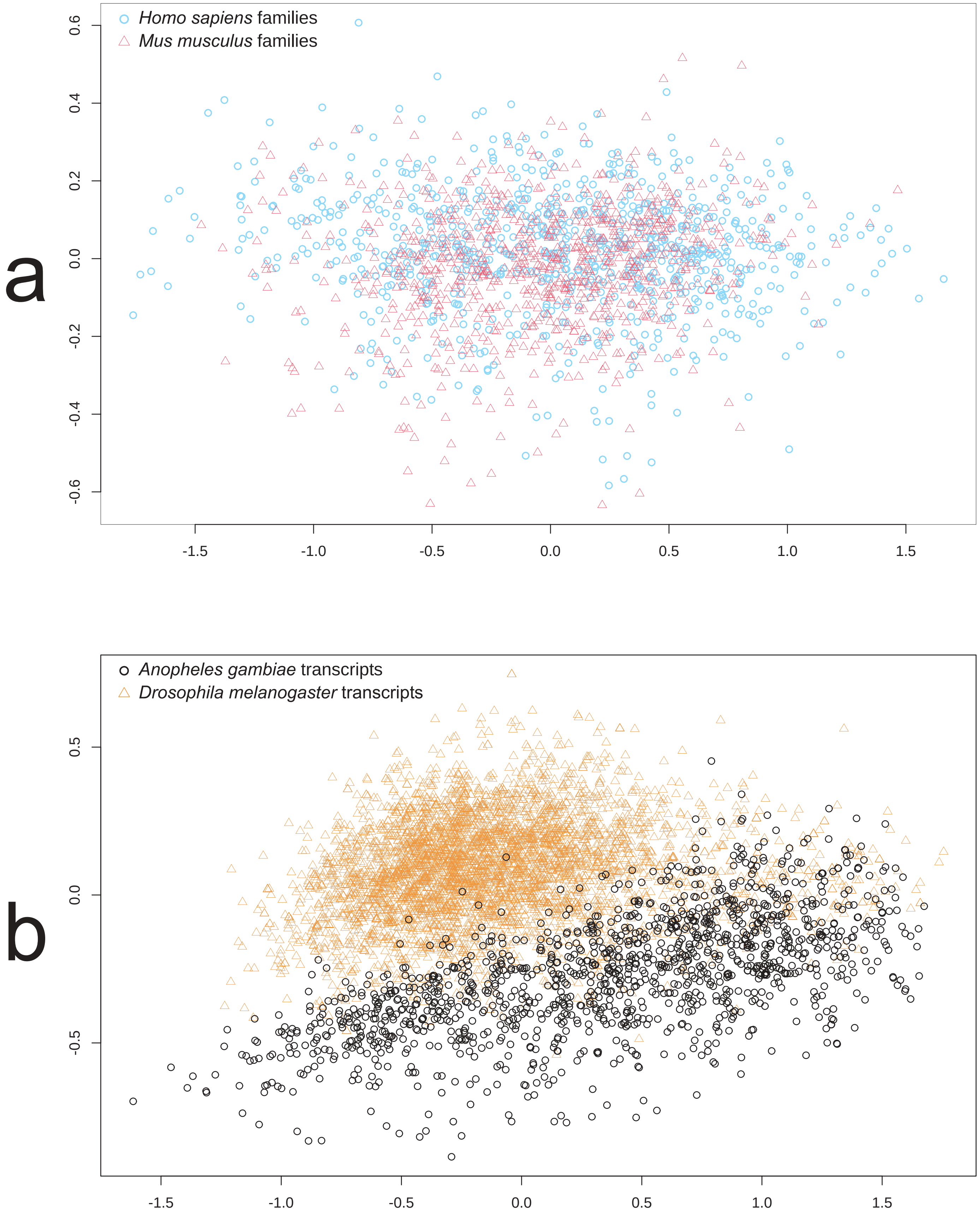


***Figure III-9**: (a) Multidimensional scaling plot of transcript families from the human and mouse genomes. (b) Multidimensional scaling plot of CSYN-filtered transcripts from fly and mosquito genomes.*

D.3.3 *Anopheles gambiae*

A comparison between CSYN-filtered transcripts from the mosquito genome and from the fly genome (*Drosophila melanogaster*) confirms that the latter is less heterogeneous than the mosquito genome. Although the codon similarity between the two genomes is quite high, with an Euclidean distance of 0.54 units between the respective genome biases, multidimensional scaling is able to separate the majority of the transcripts of the two species in two clusters (Figure III-9 *b*). The first component of separation (accounting for 43% of the total variation) is G+C content, with high total G+C on the left side of the map and low G+C transcripts on the right side. The second and weaker component (7% of the total variation), which effectively separates the transcripts of the two genomes, is mainly accounted by a differential preference for NAT codons over NAC ones (N = any nucleotide), *i.e.* the T-ending codons for Aspartate, Asparagine, Tyrosine and (to a lower extent) Histidine have higher relative frequencies in fly while mosquito transcripts preferentially use the synonymous C-ending triplets. These principal components identified are the same if correspondence analysis is used instead of multidimensional scaling.

D.3.4 Together

When plotted in the same map, the invertebrate genome clearly stands out (Figure III-10 *a*, bottom side of the plot) and compresses the other genomes because of the great difference. The relative compactness of the pufferfish genome is also noticeable in this map, revealing its narrow spread in G+C contents and homogeneity of codon usage. The major component of separation (the x-axis) can once more be equated to G+C or GC3 content. For the y-axis, along which the vertebrate genomes are separated from the mosquito, one of the main contributions is the content of Arg_A1 (A-beginning codons for Arginine, which are almost absent in the *Anopheles* genome; Figure III-10 *b*). Another important component to the separation is the content of the CCG codon for Proline (Pro_G3) which has a very high relative content in mosquito (for many clusters it is the most abundant synonymous triplet for Proline, with more than 50% usage and even reaching values of 80%) while it is mostly under-represented in vertebrates (see also the analysis of vertebrate codon space, chapter VI section D.3.1).

The intra-genomic codon heterogeneity of the analysed Ensembl genomes is lower than the one observed in the prokaryotes, with the exception of *Anopheles gambiae* which has a distribution of transcript distances comparable with archaea and bacteria (Figure III-11). For the vertebrate genomes, the $x_{0.75}$ quantile is under 1.2 distance units (and even under 1 for the pufferfish, whose codon usage is very homogeneous for the majority of transcripts).

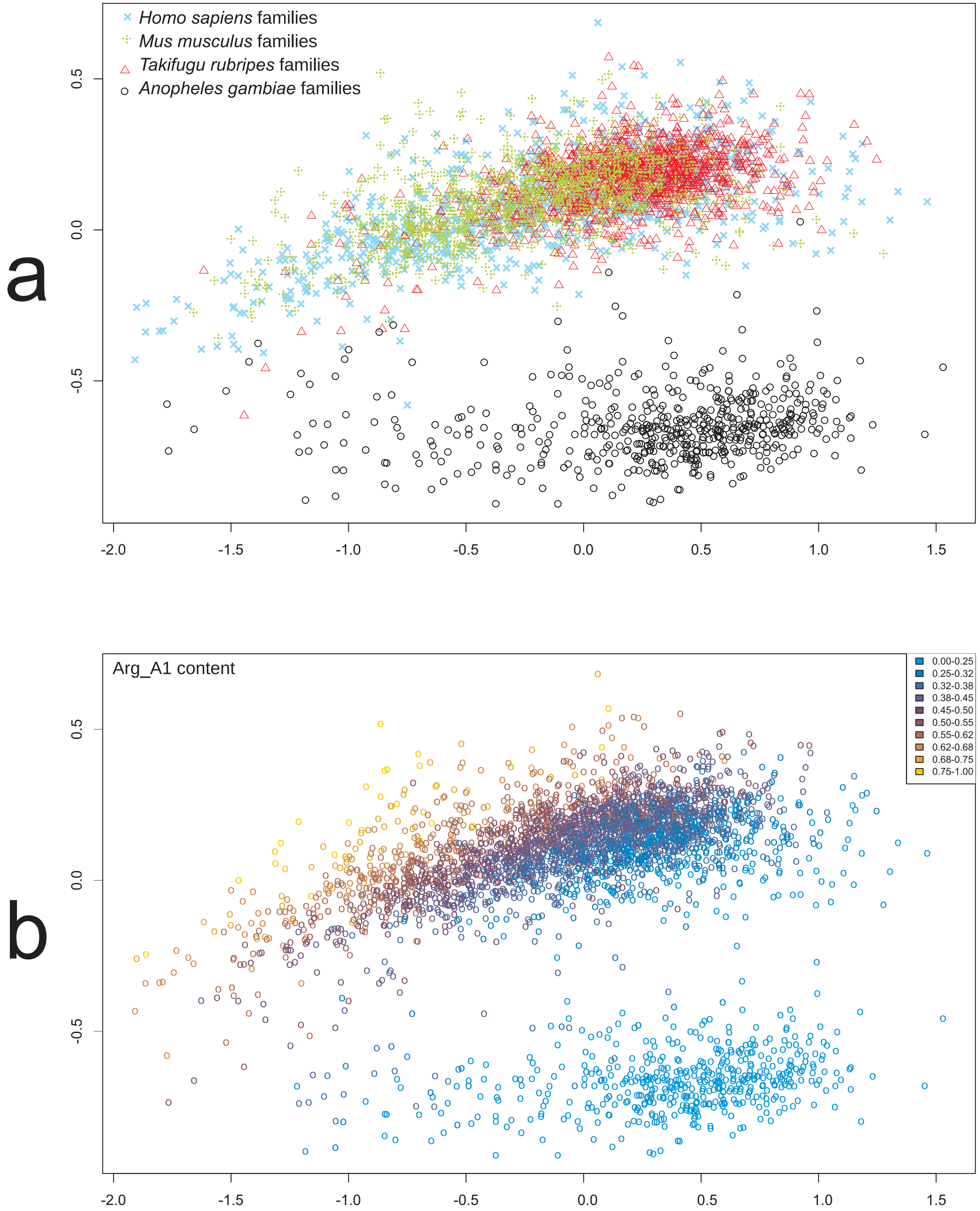


Figure III-10: *(a) The four eukaryotic genomes on the same multidimensional scaling plot. (b) The major contribution to the separation of the mosquito genome from the vertebrate ones comes from Arginine AGR (AGA or AGG) codons: values shown are for the Arg_A1 codon profile dimension.*

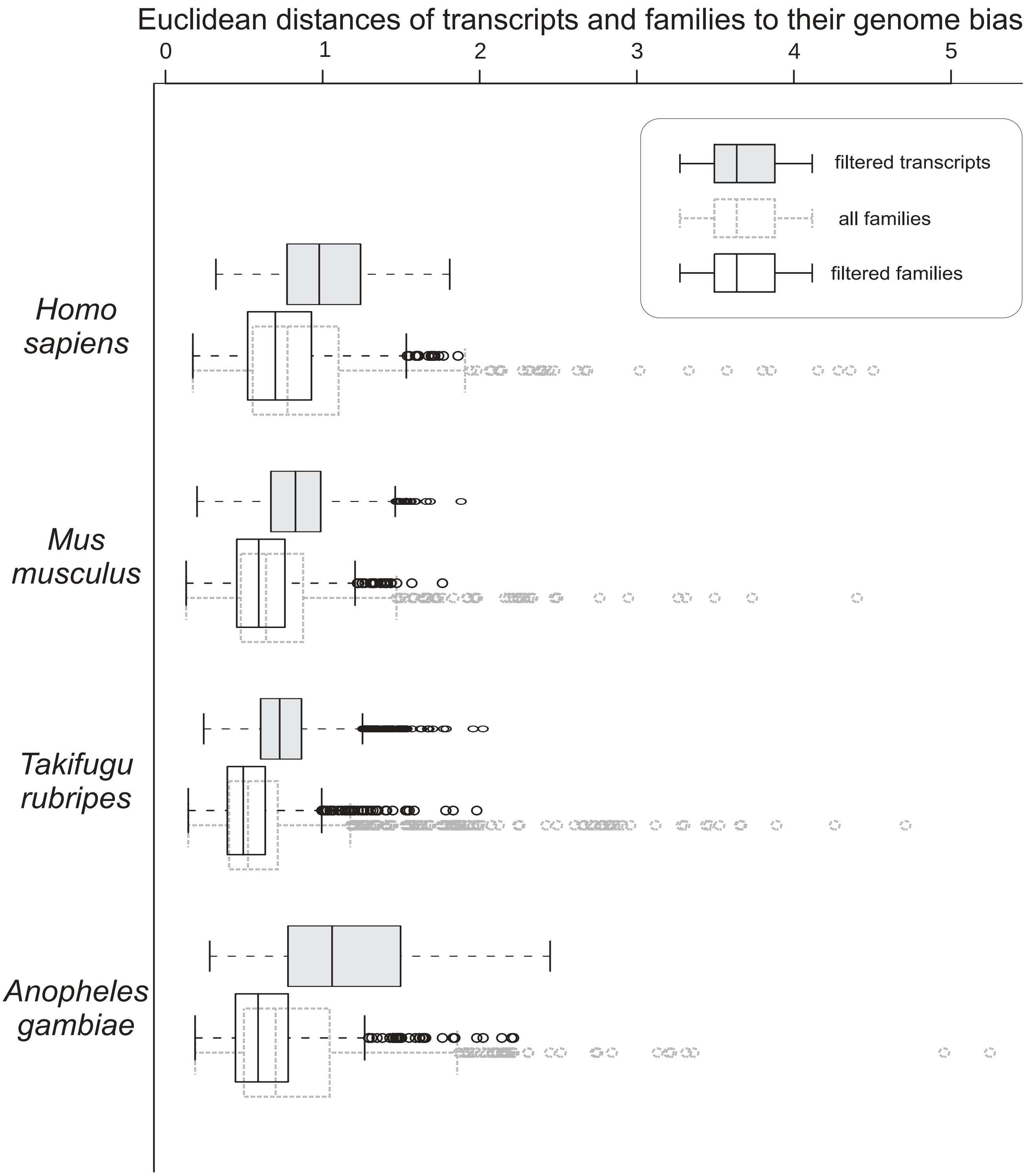


Figure III-11: *The intra-genomic codon heterogeneity of the four Ensembl genomes analysed. The distributions of distances from the genomic codon bias are plotted with boxplot representation. The distributions of filtered transcripts, filtered protein families and unfiltered (total) protein families are shown, for each genome. The boxplots relative to all the families are shown in gray dashed lines, with the boxplots of the CSYN-filtered families superimposed.*

D.4 Ranges, definition of atypicality

The results of the above described analyses reveal the relatively consistent trend in intra-genomic heterogeneity for several different genome types. The distributions of codon similarity always fall between nearly equivalent ranges and have the same shape.

It is hence possible to quantify the amount of codon similarity and codon atypicality using the Euclidean vector distance between a transcript and the average genomic bias. A transcript whose distance from the average codon bias is more than 1.5 units can rightly be called atypical in any genome, being more diverse than the 75% of the transcripts (a bit more diverse under Eukaryota, a bit less in Archaea). A transcript whose distance is 2.0 is definitely characterised by a very exceptional codon usage (in its genomic context), an outlier in the genomic distribution.

The possibility to define a quantitative measure of codon similarity in a consistent way was applied to the detection of Horizontal Gene Transfers and the identification of donor genomes (chapter four) and to the isolation and characterisation of groups of transcripts with atypical codon usage (chapter five).

E CONCLUSIONS

The intra-genomic heterogeneity was represented as the distribution of codon usage distances from the average genomic bias. The heterogeneity was then compared across the species. The intra-species heterogeneity and the *genome hypothesis* – according to which each genome holds a specific codon usage signature – are evaluated and in particular the complementarity of the two views is reaffirmed. Even if a considerable codon usage diversity exists between transcripts in the same genome, it is still possible to observe a clearly limited and well defined pattern in the variability, with codon similarity among transcripts coherently bound between comparable ranges. The determination of consistent ranges for similarity and atypicality allows clustering studies based on this information and provides them the necessary statistical background.

F APPENDIX

F.1 List of analysed prokaryotic genomes

Genomes analysed with their abbreviation, accession number and scientific name.

Archaea: afulgidus (AE000782: Archaeoglobus fulgidus), apernix (BA000002: Aeropyrum pernix), halobacterium_NRC1 (AE004437: Halobacterium sp. NRC-1), macetivorans_C2A (AE010299: Methanosarcina acetivorans C2A), mjannaschii (L77117: Methanocaldococcus jannaschii), mkandleri_AV19 (AE009439: Methanopyrus kandleri AV19), mmazei (AE008384: Methanosarcina mazei Goe1), mthermoautotrophicum (AE000666: Methanothermobacter thermautotrophicus str. Delta H), pabyssi (AL096836: Pyrococcus abyssi), paerophilum (AE009441: Pyrobaculum aerophilum), pfuriosus_DSM3638 (AE009950: Pyrococcus furiosus DSM 3638), phorikoshii (BA000001: Pyrococcus horikoshii), ssolfataricus (AE006641: Sulfolobus solfataricus), stokodaii (BA000023: Sulfolobus tokodaii), tacidophilum (AL139299: Thermoplasma acidophilum), tvolcanium (BA000011: Thermoplasma volcanium)

Bacteria: aaeolicus (AE000657: Aquifex aeolicus), atumefaciens_Cereon (AE007869, AE007870: Agrobacterium tumefaciens str. C58), atumefaciens_UW (AE008688, AE008689: Agrobacterium tumefaciens str. C58), baphidicola_Sg (AE013218: Buchnera aphidicola str. Sg), bburgdorferi (AE000783: Borrelia burgdorferi), bhalodurans (BA000004: Bacillus halodurans), bjaponicum (BA000040: Bradyrhizobium japonicum), blongum (AE014295: Bifidobacterium longum NCC2705), bmelitensis (AE008917, AE008918: Brucella melitensis), bsubtilis (AL009126: Bacillus subtilis), bsuis_1330 (AE014291, AE014292: Brucella suis 1330), buchnera_APS (BA000003: Buchnera sp. APS), cacetobutylicum (AE001437: Clostridium acetobutylicum), ccrescentus (AE005673: Caulobacter crescentus CB15), cefficiens_YS314 (BA000035: Corynebacterium efficiens YS-314), cglutamicum (AX114121: Corynebacterium glutamicum), cglutamicum_ATCC13032 (BA000036: Corynebacterium glutamicum ATCC 13032), cjejuni (AL111168: Campylobacter jejuni subsp. jejuni NCTC 11168), cmuridarum (AE002160: Chlamydia muridarum), cperfringens (BA000016: Clostridium perfringens str. 13), cpneumoniae (AE001363: Chlamydophila pneumoniae CWL029), cpneumoniae_AR39 (AE002161: Chlamydophila pneumoniae AR39), cpneumoniae_J138 (BA000008: Chlamydophila pneumoniae J138), ctepidum_TLS (AE006470: Chlorobium tepidum TLS), ctrachomatis (AE001273: Chlamydia trachomatis), dradiodurans (AE000513, AE001825: Deinococcus radiodurans), ecoli_CFT073 (AE014075: Escherichia coli CFT073), ecoli_K12 (U00096: Escherichia coli), ecoli_O157 (AE005174: Escherichia coli O157:H7 EDL933), ecoli_O157_RIMD (BA000007: Escherichia coli O157:H7), fnucleatum_ATCC25586 (AE009951: Fusobacterium nucleatum subsp. nucleatum), hinfluenzae (L42023: Haemophilus influenzae Rd), hpylori_26695 (AE000511: Helicobacter pylori 26695), hpylori_j99 (AE001439: Helicobacter pylori J99), linnocua_Clip11262 (AL592022: Listeria innocua), linterrogans_lai_56601 (AE010300, AE010301: Leptospira interrogans serovar lai str. 56601), linterrogans, llactis_IL1403 (AE005176: Lactococcus lactis subsp. lactis), lmonocytogenes_EGDe (AL591824: Listeria monocytogenes), mgenitalium (L43967: Mycoplasma genitalium), mleprae_TN (AL450380: Mycobacterium leprae), mloti (BA000012: Mesorhizobium loti), mpenetrans (BA000026: Mycoplasma penetrans), mpneumoniae (U00089: Mycoplasma pneumoniae), mpulmonis (AL445566: Mycoplasma pulmonis), mtuberculosis (AL123456: Mycobacterium tuberculosis H37Rv), mtuberculosis_CDC1551 (AE000516: Mycobacterium tuberculosis CDC1551), nmeningitidis_MC58 (AE002098: Neisseria meningitidis MC58), nmeningitidis_Z2491 (AL157959: Neisseria meningitidis Z2491), nostoc_PCC7120 (BA000019: Nostoc sp. PCC 7120), oiheyensis (BA000028: Oceanobacillus iheyensis), paeruginosa (AE004091: Pseudomonas aeruginosa PA01), pmultocida (AE004439: Pasteurella multocida), pputida_KT2440 (AE015451: Pseudomonas putida KT2440), rconorii_Malish7 (AE006914: Rickettsia conorii), rprowazekii (AJ235269: Rickettsia prowazekii), rsolanacearum_GMI1000 (AL646052: Ralstonia solanacearum), sagalactiae (AE009948: Streptococcus agalactiae 2603V/R), sagalactiae_NEM316 (AL732656: Streptococcus agalactiae NEM316), saureus_MW2 (BA000033: Staphylococcus aureus subsp. aureus MW2), saureus_Mu50 (BA000017: Staphylococcus aureus subsp. aureus Mu50), saureus_N315 (BA000018: Staphylococcus aureus subsp. aureus N315), scoelicolor (AL645882: Streptomyces coelicolor), sepidermidis_ATCC12228 (AE015929: Staphylococcus epidermidis ATCC 12228), sflexneri2astr301 (AE005674: Shigella flexneri 2a str. 301), smeliloti_1021 (AL591688: Sinorhizobium meliloti), smutans_UA159 (AE014133: Streptococcus mutans UA159), soneidensis_MR1 (AE014299: Shewanella oneidensis MR-1), spneumoniae (AE005672: Streptococcus pneumoniae TIGR4), spneumoniae_R6 (AE007317: Streptococcus pneumoniae R6), spyogenes (AE004092: Streptococcus pyogenes M1 GAS), spyogenes_MGAS315 (AE014074: Streptococcus pyogenes MGAS315), spyogenes_MGAS8232 (AE009949: Streptococcus pyogenes MGAS8232), styphiCT18 (AL513382: Salmonella enterica subsp. enterica serovar Typhi), styphimurium_LT2 (AE006468: Salmonella typhimurium LT2), synechocystis (BA000022: Synechocystis sp. PCC 6803), telongatus_BP1 (BA000039: Thermosynechococcus elongatus BP-1), tmaritima (AE000512: Thermotoga maritima), tpallidum (AE000520: Treponema pallidum), ttengcongensis_MB4T (AE008691: Thermoanaerobacter tengcongensis), uurealyticum (AF222894: Ureaplasma urealyticum), vcholerae (AE003852, AE003853: Vibrio cholerae), vvulnificus_CMCP6 (AE016795, AE016796: Vibrio vulnificus CMCP6), wbrevipalpis (BA000021: Wigglesworthia brevipalpis), xaxonopodis (AE008923: Xanthomonas axonopodis pv. citri str. 306), xcampestris (AE008922: Xanthomonas campestris pv. campestris str. ATCC 33913), xfastidiosa (AE003849: Xylella fastidiosa 9a5c), ypestis_CO92 (AL590842: Yersinia pestis CO92), ypestis_KIM (AE009952: Yersinia pestis KIM)

F.2 List of analysed viral genomes

Human infecting viruses analysed (with genome identifier and GenBank accession number): adenovirus A (10190 - NC_001460), adenovirus B (246 - NC_004001), adenovirus C (10108 - NC_001405), adenovirus D (15049 - NC_002067), adenovirus E (15868 - NC_003266), adenovirus F (10182 - NC_001454), astrovirus (13969 - NC_001943), coronavirus 229E (15577 - NC_002645), foamy virus (11546 - NC_001736), Hepatitis B virus (16449 - NC_003977), Hepatitis E virus (10157 - NC_001434), herpesvirus 1 (12187 - NC_001806), herpesvirus 2 (12163 - NC_001798), herpesvirus 3 (10044 - NC_001348), herpesvirus 4 (10040 - NC_001345), herpesvirus 5 (10043 - NC_001347), herpesvirus 6 (10586 - NC_001664), herpesvirus 6B (15112 - NC_000898), herpesvirus 7 (10884 - NC_001716), herpesvirus 8 (15951 - NC_003409), immunodeficiency virus 1 (12171 - NC_001802), immunodeficiency virus 2 (10902 - NC_001722), JC virus (10684 - NC_001699), metapneumovirus (16593 - NC_004148), papillomavirus type 1a (10055 - NC_001356), papillomavirus type 2a (10051 - NC_001352), papillomavirus type 3 (10440 - NC_001588), papillomavirus type 4 (10187 - NC_001457), parainfluenza virus 1 Washington/1964 (15991 - NC_003461), parainfluenza virus 2 (15975 - NC_003443), parainfluenza virus 3 (12158 - NC_001796), respiratory syncytial virus (11728 - NC_001781), spumaretrovirus (12157 - NC_001795), T-lymphotropic virus 1 (10159 - NC_001436), T-lymphotropic virus 2 (10260 - NC_001488), Zaire Ebola virus (15507 - NC_002549).

IV *Applications of codon profiling I: HGT detection*

A ABSTRACT

A computationally inexpensive procedure to discover potential horizontal transfer events, and to identify the donor species, was developed and tested on complete bacterial and archaeal genomes.

Comparing the codon usage of all the transcripts which are atypical in their own genomic context against the codon biases of all the genomes identifies a number of transcripts which could be the result of horizontal gene transfer events. Only those that resulted in similar codon usage as some other genome were considered. By retrieving their location on the chromosomes it was possible to predict potentially transferred regions and donor genomes.

These predictions were tested with an automatic sequence search, multiple alignment and construction of phylogenetic trees, hence combining a compositional approach with a phylogenetic one.

B INTRODUCTION

There is growing evidence (Jain *et al.*, 2002; Dutta and Pan, 2002) that natural exchange of genetic information is an essential mechanism for genetic plasticity in archaeal and bacterial genomes. The ability to thrive in new environments, metabolize new substrates or resist to new compounds, most often results from the rapid acquisition of new genes through horizontal transfer rather than by gradual alteration of the existing gene functions by mutations. Horizontal gene transfer (HGT) is the transfer of genes across species, including those belonging to different kingdoms of life.

Anomalous nucleotide or codon composition has been widely used to detect horizontally transferred genes (Garcia-Vallvé *et al.*, 2000; Mrázek and Karlin, 1999; Lawrence and Ochman, 1997; Lawrence and Ochman, 1998; Karlin, 1998; Médigue *et al.*, 1991; Koonin *et al.*, 1997; Ragan, 2001; Xie *et al.*, 2003). Those genes which present sequence composition significantly differing from the average one of their genome are considered probable lateral acquisitions. Nevertheless, the likely origin of these genes can rarely be identified (Koonin *et al.*, 2001).

Comparison of phylogenetic trees among individual genes allows identification of those with unusual origins (Smith *et al.*, 1992; Nelson *et al.*, 1999; Nesbo *et al.*, 2001). The phylogenetic methods are very powerful but require extensive sequence information and rigorous manual analysis. Furthermore, they are computationally challenging and sensitive to database sequence sampling (Ragan 2001; Lawrence and Ochman 2001; Koski *et al.*, 2001).

This chapter presents a fast and multifaceted procedure to predict the donor genome (the source of the horizontally transferred genes) or its higher taxon through analysis of codon profiles (see chapter II for a discussion on the codon profile scheme) in completely sequenced genomes. It combines the compositional and the phylogenetic approaches to circumvent the limits of both.

The number of sequenced species is extremely limited compared to the huge number of prokaryotic species in nature. The methodology presented will hence inevitably produce many false positive signals and few true positive ones, but it is bound to improve with the steadily growing number of genomic sequences being determined. Since the procedure is scalable and does not require high computing power, it can deal with very large data sets.

C METHODS

C.1 HGT detection: atypical to self, similar to other

The codon profile vectors from all completely sequenced archaeal and bacterial genomes were used in this analysis. They were computed as outlined in section III C.1 of the previous chapter.

Only transcripts that encode at least one of each amino acid species were analysed (*AA-filtering*, see II C.4). The Euclidean distances between the removed transcripts (which lack triplets for some amino acid) and the whole genomes are very high. They would hence always appear atypical and never similar to any genome: the removed genes would not be predicted as possible results of HGT. The filtering prevents unnecessary calculations: even if the amount of dropped transcripts is high, these would not be transcripts valuable to the presented HGT detection methodology.

Codon similarity was measured by the Euclidean distance in the codon profile vector space and computed between each transcript and each genome.

The distributions of distances for all the prokaryotic genomes were previously plotted (chapter III, section D.1) and compared to determine what values of codon similarity are to be considered normal and what values are to be considered atypical. The histogram and boxplot representations were used to investigate shapes and ranges of the distributions. The transcripts representing the upper outliers of those distributions (from a distance of 1.7–1.8 upwards, see Figure III-1 and Figure III-2) are taken to be sufficiently different to be analysed against other genomes.

The codon similarity of each deviant transcript to all other genome averages was then computed, isolating those transcripts which presented anomalous codon usage: different from self but very similar to some other genome.

Different levels of similarity and deviance were tried in order to determine which distance values, in codon profile space, between transcripts and genomes would set the limit for the identification of possible HGT (see for example Figure IV-4 in section D.6).

C.2 Chromosome localisation

Retrieving the genome location of these transcripts, some regions were predicted as being potentially originating from HGT events: the finding of consecutive transcripts which are all atypical in relationship to their own genome bias, and all similar to another genome, increases the feasibility of the transfer hypothesis, since they can represent transferred (and positively selected) operons. This is based on the fact that there is little conservation of gene order between distantly related genomes and the presence of three or more genes in the same order in two such genomes has been determined very unlikely unless the genes are part of an operon (Wolf *et al.*, 2001).

A cut-off distance can be specified for the definition of region, as the maximum distance, in base pairs, between beginning and end of transcripts for these to be linked in the same region. For the presented results, the distance limit of 2500 base pairs was chosen. The actual average distance between the matching transcripts was found to be

512 *bp*, with 40% of the distances under one hundred *bp* and only 10% more than 1000 *bp*. Considering that the average length of the prokaryotic transcripts in the analysis was 1122 base pairs (with a standard deviation of 696), one or two unmatched genes are allowed in the region definition. This might be considered a restrictive constraint (since there could be inter-genic sequences longer than 2500 bp between genes from the same HGT event) that could be relaxed in order to detect more regions.

The delineation of possibly transferred regions enabled fine-tuning of the thresholds for codon similarity and atypicality: if limits for similarity were set too low, almost no region would appear (since there is deviation among the similarities between any assembly of consecutive transcripts); the same would hold true if limits for atypicality were set too high. Conversely, if limits for atypicality are set too low, or limits for similarity too high, then too many transcripts would be predicted as being the result of horizontal gene transfer.

A codon profile distance threshold of 2 was judged too restrictive, as it yields only 4,412 atypical transcripts. On the other hand a threshold of 1.6 marks 16,501 transcripts as atypical. After repeated testing, the threshold of 1.8 was chosen for atypicality. 9,273 bacterial transcripts and 1,426 archaeal ones satisfied this constraint. The extension and diversity of the codon usages of these atypical transcripts is represented in section D.3.2 of chapter VI, where the average genomic biases and the codon usages of the atypical transcripts are plotted on a multidimensional scaling map (Figure VI-10).

As for the lower threshold, the limit of similarity, it was set to 1.25 units (lower values could be used when analysing closely related species, as outlined in section D.5.2 below). With these thresholds, 1,548 atypical transcripts showed codon similarity to one or more genomes. On average, each of these transcripts had nine matching genomes. See Figure IV-4 in section D.6 for the correspondence between several similarity thresholds and the number of matches between atypical transcripts and genomes.

The longest regions (or multi-regions) were chosen as the best predictions of the composition-based detection methodology, like in the example case reported in Table IV-1.

Locations: Genomes:	1045281..1046744	1046741..1048285	1048384..1050231	1051095..1051637	1051646..1052614
blongum		1	2		3
cefficiens_YS314	1	2	3		4
ctepidum_TLS		1	2		3
mtuberculosis	1	2	3		4
mtuberculosisCDC1551	1	2	3		4
pputida_KT2440	1	2	3	4	5
xaxonopodis		1	2	3	4
xcampestris		1	2	3	4

***Table IV-1**: The possible donor genomes for a region detected in the S.oneidensis genome (genes nuoN to nuoH) and the selection of the best match as the one with the linked region containing more transcripts: P.putida. The transcripts belonging to the regions are indicated by their location in the EMBL file (note that the "complement" keyword, indicating a transcript encoded on reverse strand, was removed).*

C.3 Semi-automated phylogenetic verification

As for the comparative genomics procedure, all transcripts were translated to proteins and the resulting data set was searched against NR, the non-redundant protein database (February 2003), using PSI-BLAST (Altschul *et al.*, 1997). Only matches with an E-value lower than $5{\cdot}10^{-4}$ and a sequence identity higher than 25% were included. If the number of BLAST hits for any target protein turned out to be lower than 25, constraints were relaxed to a maximum E-value of $4{\cdot}10^{-3}$ and a minimum sequence identity of 15%.

The scripts for automated large-scale PSI-BLAST analyses and the computational resources to run them were generously provided by Park Jong Hwa.

Multiple alignments of the matches were generated with Clustalw (Thompson *et al.*, 1994) with calculation of neighbour-joining trees (Saitou and Nei, 1987) which were plotted using the *drawgram* program from the PHYLIP package (Felsenstein, 1989). The cladogram-like trees were assessed for evidence of horizontal gene transfer.

C.3.1 Assessment of generated trees

Probable HGT was considered where proteins from the same species were consistently found (for the majority of the transcripts belonging to a region) in the closest clades to the target sequence, in positions higher than that of some other species accepted as taxonomically closer to the target species.

If the probable donors found by this procedure matched the ones predicted on the basis of codon profile similarity and genome location, the prediction was considered positive.

To cope with the low sampling of all the existing genomes represented by the available sequenced ones, predictions were also considered positive if the phylogenetic procedure identified as a probable donor a species belonging to the same lineage (in the same genus or family) as the one being predicted by the compositional approach. These cases are marked as *v?* in Table IV-2.

In many cases no consensus could be found among the resulting trees: no single taxon would consistently appear as probable donor. That is, proteins for the same taxon would not be found in close proximity to the target sequence for at least half of the transcripts belonging to the analysed regions. For example in a region consisting of four genes, the best hits are all from different genomes and even considering hits with lower similarity, there would be no consensus. These cases are marked as *x?* in Table IV-2.

D RESULTS AND DISCUSSION

D.1 Overview: the multifaceted and lightweight approach

Several authors (in particular Koski *et al.*, 2001) advocate the necessity for combined approaches to the detection of HGT events, stressing the requirement of a phylogenetic approach for the main purpose of avoiding to predict as HGT those vertically evolved genes with atypical composition. Furthermore, the phylogenetic methods are computationally challenging and very time consuming in the analysis of results, while compositional methods are very fast and easy to automate.

The methodology here presented combines a very fast compositional method based on codon information with a slower and computationally expensive phylogenetic method.

The compositional detection phase can be run in only three hours on a modest 400 Mhz PC for all the completed bacterial genomes (enabling repeated testing with different threshold settings).

By comparison, the phylogenetic data is obtained after BLAST searches, sequence extraction (from the constantly growing public databases), multiple alignments and tree generation, with much higher computational requirements. The analysis of the results (visual inspection of the phylogenetic trees) is the most time consuming part, especially when the number of trees becomes very large.

The combination of the two approaches increases the significance of the results and eliminates the high number of false positives that the compositional method alone would produce. But equally importantly, it restricts the use of the phylogenetic approach to a reduced and filtered set, thereby making the procedure practical and more efficient.

D.2 The predicted regions

Each transcript showing sufficient difference from its genomic codon composition was compared to the average codon profile of every other genome to identify transcripts showing codon similarity to one or more genomes.

To select transcripts with a high probability of having been acquired through HGT, it was assumed that several genes would be transferred in one event. A transferred region was defined as a sequential array of at least three transcripts with a codon profile similar to the same genome. 134 transcripts in 28 transferred regions (Table IV-2) were identified as highly probable HGT and chosen to validate the detection methodology.

These transcripts were further tested by performing automatic BLAST searches on protein sequence space, multiple alignments of all hits and visual scrutiny of the generated cladogram-like trees. It was required that most or all transcripts in a transferred region formed a clade with proteins from the same candidate donor species.

The clade had to be tight enough to exclude matches to any other species in the same genus as the recipient species.

Sequence similarity clustering confirmed the codon profile prediction in seven studied regions. Two other regions fulfilled the phylogenetic criteria for an HGT event, but the donor species differed from the codon profile prediction (regions 10 and 20 in Table IV-2). Figures Figure IV-1 and Figure IV-2 report some of the phylogenetic trees. The remaining candidate regions were taken to be false positive predictions of the codon profile method, ruled out by the phylogenetic approach. Table IV-3 reports the Euclidean distances of each region to its own genome bias and to the predicted donor genomes. Often, but not always, the confirmed donor is equivalent to the one with the shortest Euclidean distance to the region (marked with bold typeface in the table). That table also includes the distances between the genomic codon bias of the recipient and the donors.

	Genome (predicted recipient)	region's limits(1)	gene names(2)	size (3)	predicted donor genomes	probable donor, if any, after verification	notes(4)	(5)
1	*A.aeolicus*	273-280	*aq_378-aq_386*	4	*P.furiosus, S.tokodaii, S.solfataricus*	*?*		*x?*
2	*A.aeolicus*	370-375	*aq_509-mtfC*	3	*C.acetobutylicum, C.perfringens*	*?*		*x?*
3	*A.pernix*	1224-1235	*APE1182-APE1193*	7	*S.solfataricus, S.tokodaii*	*?*		*x?*
4	*B.longum*	192-195	*BL0206-BL0209*	3	*C.pneumoniae, B.halodurans, B.subtilis, Nostoc sp., S.pneumoniae*	*?*		*x?*
5	*B.longum*	209-214	*BL0230-cps2F*	6		*?*		*x?*
6	*E.coli_K12*	534-536	*ybcK-ybcM*	3	*C.muridarum, O.iheyensis, R.conorii, S.agalactiae*	*?*	maybe the donor is another strain of *E.coli*	*x?*
7	*E.coli_K12*	1991-1997	*wbbK-rfbC*	6	*B.burgdoferi, R.conorii, L.interrogans, S.pyogenes, S.pneumoniae*	*S.pneumoniae*	(4/6)	*v*
8	*E.coli_K12*	3545-3551	*rfaK-rfaS*	3	*B.burgdoferi, B.aphidicola, L.lactis, O.iheyensis, R.conorii, S.agalactiae*	*?*		*x?*
9	*E.coli_O157*	2860-2866	*Z3198-wbdN*	6	*R.conorii, S.tokodaii*	*?*		*x?*
10	*E.coli_O157_RIMD*	3507-3512	*ECs3507-ECs3512*	4	*B.burgdoferi, C.acetobutylicum, R.conorii, S.tokodaii*	*P.multocida ?*	too similar in CPRO (3/4)	*x!*
11	*M.thermoautotrophicum*	329-334	*MTH332-MTH337*	4	*C.acetobutylicum, T.tengcongensis, S.solfataricus, S.tokodaii*	*?*		*x?*
12	*N.meningitidis_MC58*	1887-1890	*NMB2008-NMB2013*	3	*L.interrogans, Nostoc sp., R.prowazekii*	*?*		*x?*
13	*N.meningitidis_MC58*	675-677	*NMB0725-NMB0727*	3	*Nostoc sp., H.influenzae, L.innocua, S.galactiae*	*H.parainfluenzae* or *H.paragallinarum*	same genus as *H.influenzae* but complete sequence not available (3/3)	*v?*
14	*P.aeruginosa*	1369-1371	*PA1370-PA1372*	3	*T.volcanium*	*?*		*x?*
15	*P.aeruginosa*	2222-2226	*PA2224-PA2228*	3	*T.volcanium*	*?*		*x?*
16	*P.aeruginosa*	3143-3149	*wbpL-hisF2*	7	*X.fastidiosa, Y.pestis*	*?*		*x?*
17	*P.putida_KT2440*	4402-4408	*PP4461-PP4467*	4	*X.fastidiosa*	*?*		*x?*
18	*S.flexneri2astr301*	2093-2100	*SF2093-rfbE*	5	*O.iheyensis, R.conorii*	*?*	(2/5)	*x?*
19	*S.oneidensis_MR1*	995-1001	*nuoN-nuoH*	5	*P.putida*	*?*		*x?*
20	*S.typhiCT18*	2110-2118	*rfbP-rfbS*	7	*O.iheyensis, R.conorii, S.tokodaii*	*Y.pseudotuberculosis* or *Y.enterocolitica*	too similar in CPRO (4/7)	*x!*
21	*S.typhiCT18*	2166-2173	*STY2350-STY2358*	3	*R.conorii*	*0*		*x0*
22	*S.typhiCT18*	4483-4485	*STY4822-STY4824*	3	*R.conorii, S.tokodaii, L.innocua, L.monocytogenes, Oiheyensis*	*?*		*x?*
23	*X.axonopodis*	1472-1487	*orf2-XAC1509*	8	*B.subtilis*	*?*		*x?*
24	*X.fastidiosa*	1714-1723	*XF1718-XF1727*	5	*P.aeruginosa, P.putida, R.solanacearum*	*P.aeruginosa* or *P.putida*	(3/5)	*v*

25	*X.fastidiosa*	1724-1734	*XF1728-XF1738*	7	*P.aeruginosa,* *P.putida,* *R.solanacearum*	*A.vinelandii*	*A.vinelandii* is in *Pseudomonadaceae* family but its complete sequence is not available (6/7)	*v?*
26	*X.fastidiosa*	1738-1748	*XF1742-XF1752*	8	*P.aeruginosa,* *P.putida,* *R.solanacearum*	*R.solanacearum* or one *Pseudomonas*	(5/8)	*v*
27	*X.fastidiosa*	1754-1764	*XF1758-XF1768*	7	*P.aeruginosa,* *P.putida,* *R.solanacearum*	*P.aeruginosa* or other *Pseudomonas*	(5/7)	*v*
28	*X.fastidiosa*	1769-1778	*XF1773-XF1783*	4	*P.aeruginosa,* *P.putida,* *R.solanacearum*	*P.aeruginosa*	(4/4)	*v*

***Table IV-2**: probable HGT identified by the methodology. Predictions are in general not unique since there is usually more than one genome with codon usage similar to that of the atypical transcripts. Regions 24 and 25 are contiguous but separated in this table because of different results from the philogenetic verification procedure.*

(1): the boundaries of the regions are indicated by the sequential number of the transcript coding sequence as it appears in the deposited sequence - e.g. the first region is the one between the 273th and the 280th CDS appearing in its genome sequence file (AE000657). See Table IV-5 in appendix F.1 for a complete list of the coding sequences.

(2): the gene names of the boundary transcripts - e.g. the first region, is the one between the genes aq_378 and aq_386 in the sequenced genome of A.aeolicus

(3): size of region in number of transcripts (some transcripts in between the regions limits are not included in the region either because they were filtered out - too short or lacking codons for certain amino acids - or because their codon profile is similar to that of the host)

(4): a question mark indicates that there is no consensus among blast hits; a zero indicates that there are no xenologous blast hits; the numbers between brackets indicate the consistency index of the probable donor, in other words in how many transcripts (out of the total in the region) that genome can be found among the top blast hits

(5): a very condensed symbolic representation of the results: v=positive x=negative, ?=no consensus among blast hits, 0=no xenologous blast hits, !=missed

	Region average to recipient genome (self)	**Region average to predicted donors**	**Recipient genome to predicted or verified donor genomes**
1	*A.aeolicus: 1.985*	*P.furiosus: 1.238, S.tokodaii:* **0.579**, *S.solfataricus: 0.825*	*C.acetobutylicum: 2.086, C.perfringens: 2.432, P.furiosus: 1.125, S.solfataricus: 1.625, S.tokodaii: 1.923*
2	*A.aeolicus: 2.007*	*C.acetobutylicum:* **0.617**	
3	*A.pernix: 2.074*	*S.solfataricus:* **0.470**, *S.tokodaii: 0.623*	*S.solfataricus: 2.155, S.tokodaii: 2.525*
4	*B.longum: 2.073*	*C.pneumoniae: 0.761, B.halodurans:* **0.708**, *B.subtilis: 0.860, Nostoc sp.: 0.927, S.pneumoniae: 0.739*	*B.halodurans: 2.183* *B.subtilis: 1.993* *C.pneumoniae: 2.409* *Nostoc sp.: 2.495* *S.pneumoniae: 2.411*
5	*B.longum: 2.508*	*C.pneumoniae:* **0.594**, *B.halodurans: 0.849, B.subtilis: 0.854, Nostoc sp.: 0.878, S.pneumoniae: 0.769*	
6	*E.coli_K12: 1.898*	*C.muridarum: 0.940, O.iheyensis: 0.802, R.conorii:* **0.722**, *S.agalactiae: 0.878*	*B.aphidicola: 2.391, B.burgdoferi: 2.287 C.muridarum: 1.563, L.interrogans: 1.766 L.lactis: 1.788, O.iheyensis: 1.854, R.conorii: 1.941, S.agalactiae: 1.737, S.pneumoniae: 1.329 S.pyogenes: 1.447*
7	*E.coli_K12: 1.729*	*B.burgdoferi: 0.775, R.conorii:* **0.664**, *L.interrogans: 0.769, S.pyogenes: 0.840, S.pneumoniae: 0.912*	
8	*E.coli_K12: 1.891*	*B.burgdoferi: 0.948, B.aphidicola: 0.899, L.lactis: 0.728, O.iheyensis:* **0.602**, *R.conorii: 0.747, S.agalactiae: 0.628*	
9	*E.coli_O157: 1.776*	*R.conorii:* **0.555**, *S.tokodaii: 0.805*	*R.conorii: 1.815, S.tokodaii: 2.119*
10	*E.coli_O157_RIMD: 2.061*	*B.burgdoferi: 0.708, C.acetobutylicum: 0.748, R.conorii: 0.918, S.tokodaii:* **0.551**	*C.acetobutylicum: 2.309, P.multocida: 1.423 R.conorii: 1.895, S.tokodaii: 2.236*
11	*M.thermoautotrophicum: 1.874*	*C.acetobutylicum: 0.780, T.tengcongensis: 0.806, S.solfataricus: 0.750, S.tokodaii:* **0.735**	*C.acetobutylicum: 2.418, S.solfataricus: 1.958, S.tokodaii: 2.266, T.tengcongensis: 1.778*
12	*N.meningitidis_MC58: 2.392*	*L.interrogans: 0.871, Nostoc sp.: 0.946, R.prowazekii:* **0.579**	*H.influenzae: 1.879, L.innocua: 0.561, L.interrogans: 2.003, Nostoc sp.: 1.618, R.prowazekii: 2.505, S.galactiae: 2.085*
13	*N.meningitidis_MC58: 2.397*	*H.influenzae: 0.859, L.innocua: 0.934, Nostoc sp.: 0.984, R.prowazekii:* **0.579**, *S.galactiae: 0.754*	
14	*P.aeruginosa: 2.488*	*T.volcanium: 0.707*	*T.volcanium: 2.824, X.fastidiosa: 1.919, Y.pestis: 2.436*
15	*P.aeruginosa: 2.462*	*T.volcanium: 0.902*	
16	*P.aeruginosa: 2.151*	*X.fastidiosa:* **0.671**, *Y.pestis: 0.782*	
17	*P.putida_KT2440: 1.731*	*X.fastidiosa: 0.762*	*X.fastidiosa: 1.351*
18	*S.flexneri2astr301: 1.895*	*O.iheyensis:* **0.573**, *R.conorii: 0.591*	*O.iheyensis: 1.830, R.conorii: 1.914, P.aeruginosa: 2.077*
19	*S.oneidensis_MR1: 1.968*	*P.putida: 0.802*	*P.putida: 2.007*
20	*S.typhiCT18: 1.867*	*O.iheyensis:* **0.555**; *R.conorii: 0.602; S.tokodaii: 0.930*	*L.innocua: 1.668, L.monocytogenes: 1.615, O.iheyensis: 1.932, R.conorii: 1.998, S.tokodaii: 2.336*
21	*S.typhiCT18: 1.909*	*R.conorii:* **0.670**	
22	*S.typhiCT18: 2.006*	*R.conorii:* **0.653**, *S.tokodaii: 0.737, L.innocua: 0.969, L.monocytogenes: 1.016, O.iheyensis: 0.807*	
23	*X.axonopodis: 1.961*	*B.subtilis:* **0.554**	*B.subtilis: 2.098*
24	*X.fastidiosa: 1.857*	*P.aeruginosa:* **0.463**, *P.putida: 0.694, R.solanacearum: 0.514*	*P.aeruginosa: 1.919, P.putida: 1.351, R.solanacearum: 1.828*
25	*X.fastidiosa: 1.964*	*P.aeruginosa: 0.548, P.putida: 0.835, R.solanacearum:* **0.368**	
26	*X.fastidiosa: 2.054*	*P.aeruginosa: 0.569, P.putida: 0.874, R.solanacearum:* **0.435**	
27	*X.fastidiosa: 1.705*	*P.aeruginosa: 0.579, P.putida: 0.644, R.solanacearum:* **0.452**	
28	*X.fastidiosa: 1.806*	*P.aeruginosa: 0.683, P.putida: 0.774, R.solanacearum:* **0.426**	

Table IV-3*: Euclidean distances of the identified HGT regions. The distances of the region to its own genome and to the predicted genomes are indicated. Furthermore, the distances between donor and acceptor genomes are reported. A bold typeface marks the shortest region-donor distances.*

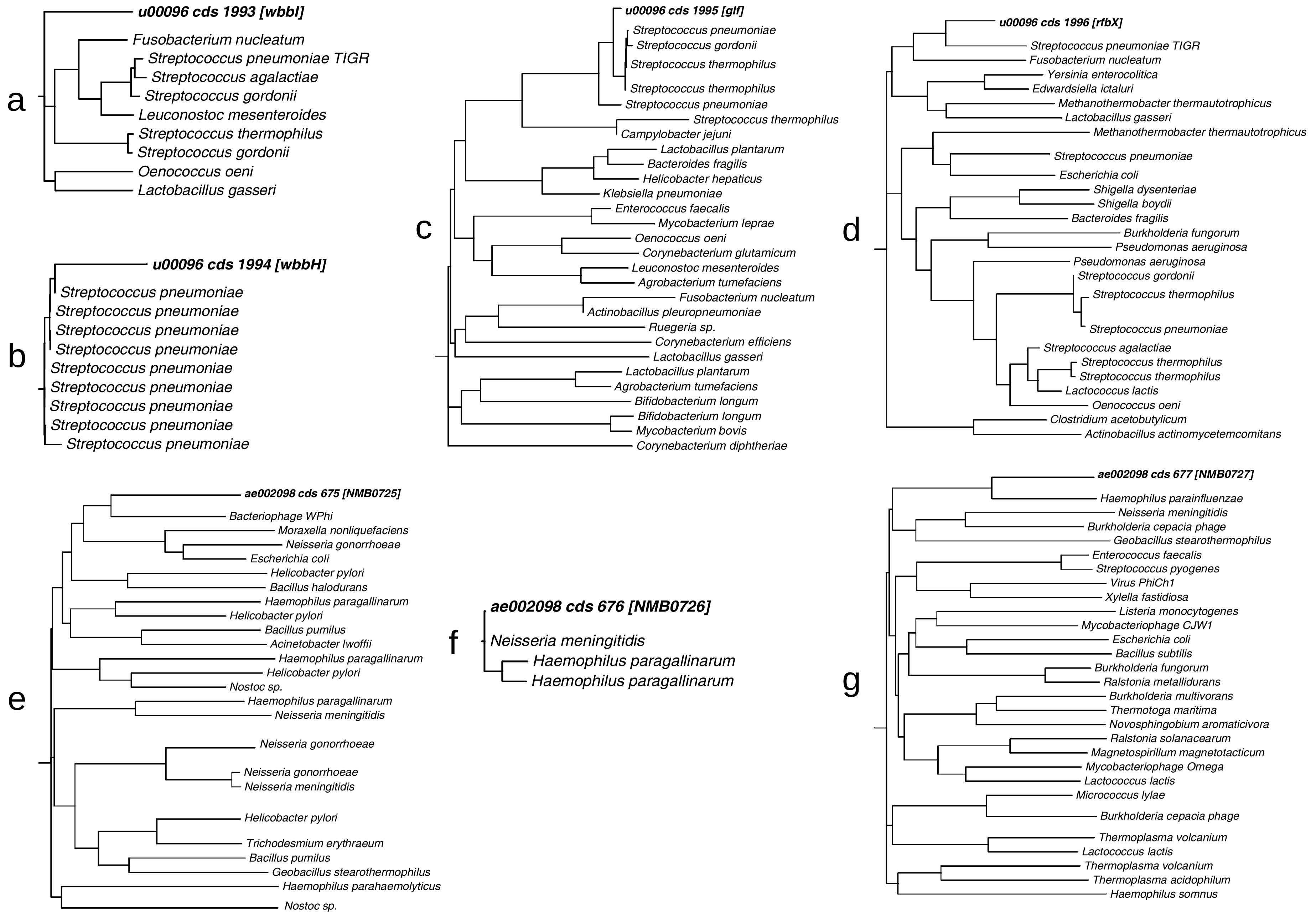
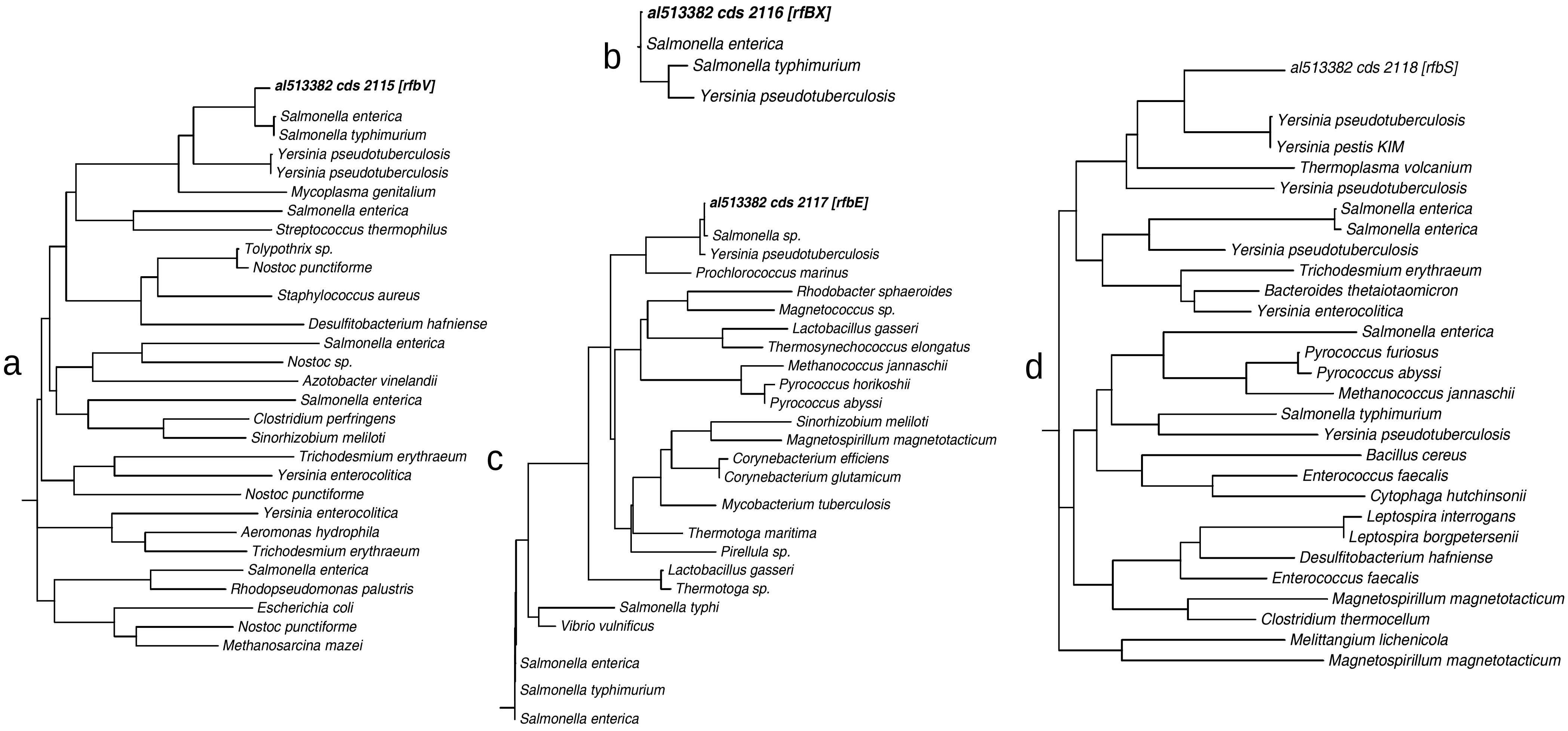


Figure IV-1: *Phylogenetic verification trees for region 7, Escherichia coli: (a) wbbI (b) wbbH (c) glf (d) rfbX and for region 13, Neisseria meningitidis: (e) NMB0725 (f) NMB0726 (g) NMB0727.*



***FigureIV-2**: Phylogenetic verification trees for region 20, Salmonella typhi: (a) rfbV (b) rfbX (c) rfbE (d) rfbS.*

D.3 The *X.fastidiosa/P.aeruginosa* case

The most striking case of identified HGT comprises several regions of *Xylella fastidiosa*, which were detected as originating from *Pseudomonas aeruginosa* (or from another species in that lineage). As they are all relatively close in the genome (62.6 kilobases between the first and the last transcript), they could all be the result of the same HGT event.

The transcripts belonging to these regions in *X.fastidiosa* have a very atypical codon profile in the *X.fastidiosa* genomic context (Figure IV-3 *a*) while being practically identical to the one of *P.aeruginosa* (Figure IV-3 *b*). Since *P.aeruginosa* has one of the most extreme codon usages (Grocock and Sharp, 2002) – 2.34 distance units from the codon profile vector averaged in all its dimensions – it is even more striking to find such a close correspondence of codon usage (only 0.5 units of distance, practically equal) within the *X.fastidiosa* genome.

Further evidence for the hypothesised HGT comes from a nucleotide alignment search of the *X.fastidiosa* region which identified zones of extremely high sequence identity with *P.aeruginosa*.

The first gene (*XF1718*) in the first predicted transferred region is annotated as *phage-related integrase*, with a 91% sequence identity with *Int-B13*, a recombinase of the bacteriophage P4 integrase family responsible for HGT of the clc element (containing chlorocatechol degradative genes) in genus *Pseudomonas* (Ravatn *et al.*, 1998). The *clc* element integrates in various bacterial recipients with a Glycine tRNA structural gene; a tRNA-Gly lies 227 bp upstream of *XF1718*.

The other genes are annotated as hypothetical proteins identified through sequence similarity (with matches to - among others - *B.subtilis*, *E.coli*, *S.coelicor*, *A.vinelandii* and *P.aeruginosa*) and are mainly ketoreductases/dehydrogenases and transcriptional regulators.

The 67kb region encompassing the predicted transfer had been identified as GI_2 (Genomic Island 2) by Nunes *et al.* (2002) while it was identified as a cryptic prophage by Bhattacharyya *et al.* (2002).

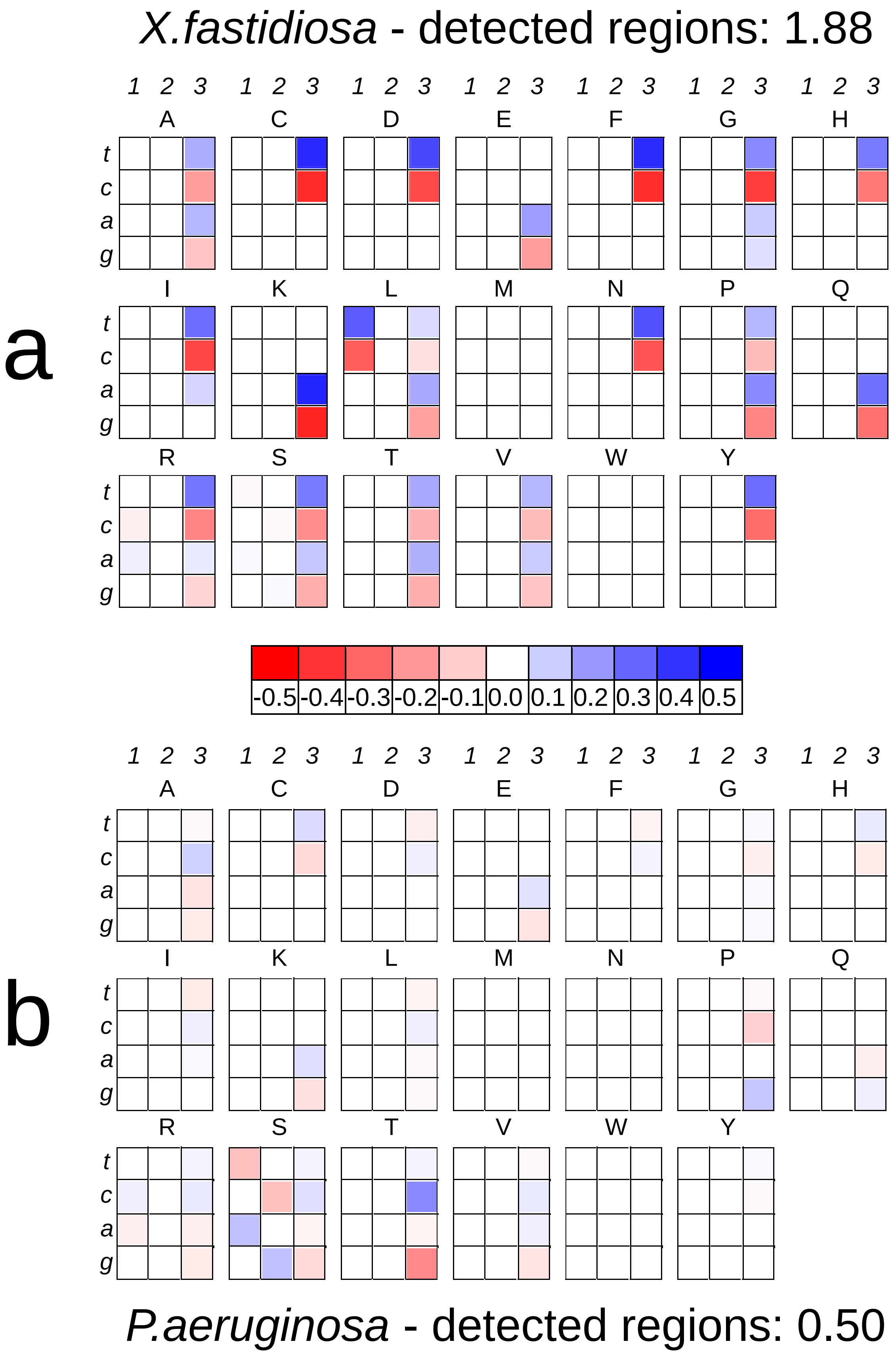


__Figure IV-3__: (a) Codon profile difference matrices and Euclidean distances (a) between the predicted HGT regions in X.fastidiosa and the genome average of X.fastidiosa (b) between the predicted HGT regions in X.fastidiosa and the genome average of P.aeruginosa, the hypothesised donor.

D.4 Sensitivity and selectivity

Since the available collection of completely sequenced genomes is only a tiny fraction of all the existing genomes, the probability of finding the exact sequence match is low; nevertheless, the methods presented could allow narrowing down the suspects for the donor genome to genus or family level.

The HGT detection methodology based on codon usage information is scalable, extremely fast and computationally inexpensive. It only requires calculation of the codon profile of each sequence (an operation based solely on counting and normalising) and measures of distance between vectors. Furthermore, the distances are computed only between atypical transcripts and the genome averages.

Considering the kind of information used and the very low number of bacterial genomes of known sequence, this method yields a considerable number of positive predictions (25%; 17% if the *X.fastidiosa* regions are considered – as they probably are – result of a single transfer event).

The number of false predictions is nevertheless very high and these need to be ruled out by a verification procedure (employing a phylogenetic approach) after the detection. The verification can be a computationally expensive procedure, but since it is applied to an already small and filtered set of cases it will not excessively affect the performance of the analysis.

D.4.1 Causes of error

The adopted phylogenetic procedure identified some possible HGT events which the compositional method failed to detect (for example region 20, *S.typhi*, which could be originating from *Y.pseudotubercolosis*; see Figure IV-2). The failure in the detection is due to two main causes: 1) high similarity in codon usage between donor and recipient genome; 2) lack of annotated genomic sequence data.

As for the former cause, high codon similarity, this is an intrinsic limit of the methodology which could only be avoided at the price of a steep increase in the number of false predictions. If the donor and receiving genomes have very similar codon usage, the transcripts in the recipient genome which are very different from their genome average bias will also be different from the average bias of the donor genome.

These transcripts (if they are really the result of a horizontal gene transfer and not atypical because of other causes: like selection acting on translational efficiency, compositional symmetry of leading versus lagging strand, and others) do not share the characteristic codon usage of either the donor or the recipient and hence would not be detected by this methodology.

Some example genomes, which have similar codon usage and between which the methodology would not have been able to predict HGT events, are: *V.cholerae* and *E.coli* (codon profile distance between the two genomes of 0.69 units), *R.solanacearum* and *P.aeruginosa* (distance of 0.62), *Y.pestis* and *S.typhi* (0.71). See D.5.2 below for a possible solution to the detection of intra-family transfers.

The other main cause, lack of annotated genomic sequence data, will be less and less relevant as more genome sequences are deposited in the public databases. As more genomes get sequenced, the methodology will yield better predictions without a substantial loss of computational performance. Fine tuning with more restrictive codon similarity thresholds (including thresholds on a genome-per-genome basis, see below) could be used, leading to a higher sensitivity during detection.

D.4.2 Undistinguishable codon usage of ameliorated genes

There is another class of false negatives to which all the compositional prediction methods are susceptible. These procedures cannot detect fully ameliorated genes whose sequences adjusted to the base composition and codon usage of the resident genome to become undistinguishable from ancestral sequences (Lawrence and Ochman, 1997). Such methods are hence restricted to discovery of relatively recent acquisitions (few millions of years, depending on the extent and rate of the amelioration process).

At the time of introduction, HGT genes have the codon usage pattern of the donor genome and will progressively accumulate substitutions (due to the mutational processes in the recipient genome) to eventually reflect the codon bias of the recipient genome. This process of amelioration was estimated to produce divergence with a rate of 0.47% and 0.0195% (for synonymous and nonsynonymous sites, respectively) per million years (Myr) (in *E.coli* when compared to *S.enterica*; Lawrence and Ochman, 1998). The average age of an HGT gene in *E.coli* was estimated in the same work as being 6.7 Myr.

The elaborated methodology is inherently biased (as all statistical compositional procedures) towards recently transferred genes (*e.g.* under 10 Myr) which have not undergone an extensive amelioration process. Comparative genomics, analysis of sequences and of phylogenetic trees are the requirements to possibly identify the ameliorated Horizontally Transferred genes.

D.4.3 Distinct codon usage for highly translated genes

For various bacterial species there is considerable evidence that intra-genomic codon usage can distinguish several classes in which the genes can be clustered, with a class of highly translated genes employing an optimised codon usage. This physiological codon bias could then be perceived as an impediment to the presented HGT detection methodology.

Without a specific study for each species, it is not trivial to differentiate the two phenomena (although some authors indicate that in general the highly expressed genes do not deviate in G+C content from the mean values of the genome; Garcia-Vallvé *et al.*, 2000). This is particularly relevant for those studies which aim at estimating the amount of horizontally transferred genes in the genomes.

The described methodology has a different scope, namely the identification of donor genomes, which had not been computationally done before, especially in this very general and automated way (Kanaya and co-workers identified matches for seven *E.coli O157* genes in *V.cholerae* according to proximity on a bacterial Self-Organising Map, comparisons with a similarity measure and BLASTP searches; 2001b).

This work does not attempt to give an estimate on the extent of cross-species transfer but to identify precise transfers. To this end, the thresholds for codon atypicality (to self) and codon similarity (to alien genome) have been set very high. The first part of the procedure, selecting atypical genes, could select genes which are atypical from the genome average not because of HGT origin but because of high-translational efficiency (*e.g.* ribosomal genes).

But the second part, the search for codon similarity matches between these atypical genes and other genomes, would remove many, if not all, of those false positives. A cross-species transfer is a more parsimonious explanation for an observed very high codon similarity between a gene in species A and a very different genome bias of species B, rather than the coincidence between the codon usage of high-translated genes in species A and the normal codon usage of the genes in genome B, especially considering the extremely high number of possible codon usages (*q.v.* chapter VI for a discussion on the size of the codon usage space and the number of codon usages for specified levels of similarity).

D.5 Possible improvements to the methodology

D.5.1 Genome specific thresholds

With the continuously increasing number of sequenced genomes, the success rate can only improve (because the donor species is more likely to be present in the data corpus) and tighter requirements of codon similarity can be set, to detect the donor directly at species level (lowering the number of multiple matches).

One logical step in this direction is represented by the possibility of setting genome specific thresholds. Although the intra-genomic heterogeneity is coherently bounded between certain ranges (as shown in chapter three), there is still a certain amount of variability which could be tapped in order to raise the specificity of the detection procedure. Instead of choosing a constant atypicality threshold of, for example, 1.8 units, a specific atypicality threshold could be defined for each genome: for instance the bacterium *Buchnera aphidicola* has lower intra-genomic heterogeneity (*q.v.* Figure III-2) and hence its atypicality threshold could be set to about 1.5 units.

Genome specific atypicality thresholds could be easily computed from the distributions of intra-genomic codon similarity. Additionally, the threshold specificity could be extended to all possible pairs of donor/acceptor genomes (n^2-n couples, where n is the number of sequenced genomes) with an automatic procedure to determine correct thresholds on the basis of the number and size of the detected regions.

D.5.2 Detection of intra-family transfers

As noted above, one of the intrinsic limits of the methodology is due to the high codon similarity between certain related genomes. Discrimination would decrease as donor and recipient genomes are more compositionally similar.

Furthermore, HGT is expected to be more likely between closely related bacteria, although HGT events have also been observed between distant species. The proposed methodology uses conservative thresholds to lower the number of many false positives but in so doing it prevents detection of intra-family transfers.

To circumvent this problem, and detect transfers between closely related taxa without generating too many false positives, their genomes would need to be analysed separately: for example transfers between *Enterobacteriaceae* could be detected, applying the procedure only on the sequences of genomes belonging to this family and setting lower and finer thresholds (in particular, the threshold of atypicality would need to be greatly lowered).

D.6 HGT detection performed on synonymous codon usage vectors.

The compositional approach, resulting in the detection of the transcripts alien in their own genomic context but similar in codon usage to another genome, was performed at various thresholds of similarity and using both codon profile (CPRO) and synonymous codon usage (CSYN) vectors for the computation of distances.

CSYN distances are generally lower in scale than CPRO ones, as explained in II D.1.3 and as observed in the study of heterogeneity of prokaryotic genomes in section III D.1, Figure III-3 and Figure III-4. CSYN distances are around 4–5% lower than the corresponding CPRO ones. For this reason, the thresholds of similarity and atypicality for CSYN were reduced to 95% of their values when used with CPRO vectors. The confirmation that the scale adjustment is appropriate comes from the number and distribution of regions identified with the two schemes and the respective thresholds (Figure IV-4): the number of matches is highly comparable between the two schemes for all the examined thresholds.

The results presented in this chapter were obtained with CPRO vectors and distance settings of 1.80 for codon atypicality and 1.25 for codon similarity (**Methods** section C.1). The corresponding thresholds for CSYN are 1.71 and 1.20 for atypicality and similarity distances, respectively. These settings produce a number of total region matches (and a distribution of their sizes) very comparable to CPRO, with actually slightly more matches (Table IV-4).

Region size	CPRO 1.8 to 1.25	CSYN 1.71 to 1.20
12	1	1
11	5	6
10	2	3
9	3	3
8	6	4
7	6	9
6	44	29
5	57	93
4	113	127
3	263	229
2	1036	984
1	7515	7605

Table IV-4*: Distribution of detection matches (linked in regions) for the two schemes CPRO and CSYN.*

Nevertheless, when unique regions are selected (the numbers in the previous table refer to the multiple matches) to choose the best regions (those that contain more matches), a lower number of long unique regions are found using CSYN vectors compared to the results obtained using CPRO ones. To recover all the regions (as in Table IV-2) and obtain the same results with both schemes, the CSYN similarity threshold needs to be raised to 1.25. Thus CSYN obtains the same results of CPRO with atypicality thresholds lowered to 95% and similarity thresholds kept at 100% of the corresponding values used in the CPRO analysis.

The similarity threshold has hence to be set to a more permissive value for CSYN vectors, increasing the total number of matches (and hence also the number of false positives). This might indicate a higher sensitivity, although slight in nature, of the codon profile scheme.

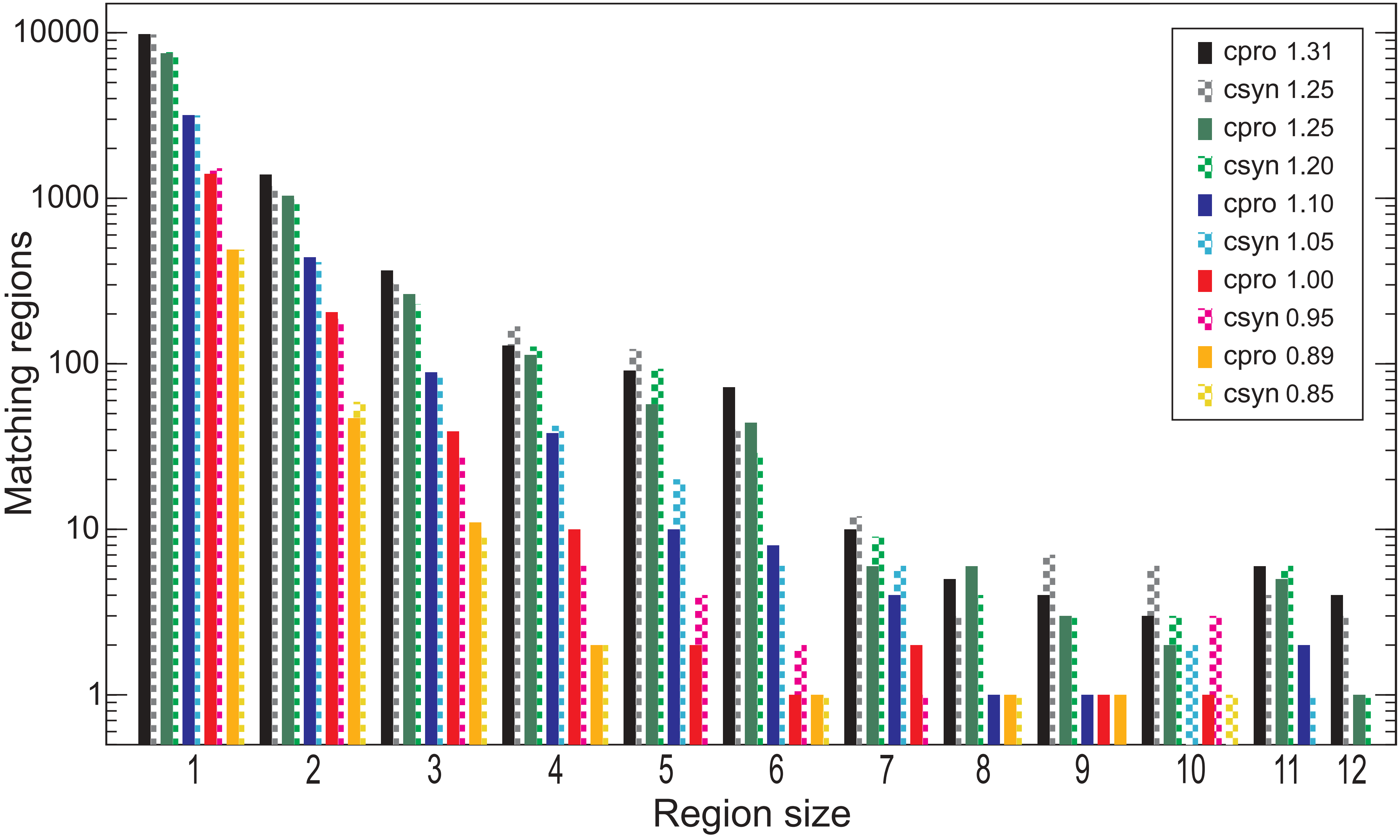


***Figure IV-4**: Number of transcript-to-genome matches for several thresholds of codon similarity. There is a correspondence between the results obtained setting CSYN similarity levels to 95% of CPRO similarity levels (each CSYN bar appears, in chequered pattern fill, to the right of the corresponding CPRO bar, full coloured). The thresholds for codon atypicality are set to 1.8 and to 1.71 for CPRO and CSYN vectors, respectively.*

E CONCLUSIONS

The elaborated methodology is computationally inexpensive and the codon profiling appears sensitive enough to successfully identify the donor genomes of the predicted HGT (something existing compositional methods do not provide). The combination of the phylogenetic approach to the compositional detection removes the many false positives that a method of compositional detection would yield if used alone. Furthermore, the sequence database searches and the assessment of phylogenetic trees, which usually require a great amount of time and resources, are restricted to a small and filtered set. The compositional detection is automated and very fast: it can be run in three hours on a 400 Mhz PC for all bacterial genomes.

Comparing with the statistical procedure – based on G+C content and codon usage – developed by Garcia-Vallvé *et al.* (2000), the number of predictions of the presented methodology is very low. This is due to the fact that HGT cases are proposed only when a probable donor genome can be identified. Of all the atypical transcripts, only those with codon usage very similar to some other genome are considered. This ensures that the detected transcripts are atypical only to their genomic context and not absolutely anomalous, thus eliminating many false positives. Out of 10,699 possible atypical transcripts, only 1,548 have a clear similarity to another genome. This might be due to the low number of sequences which have been determined (with comparison to the enormous number of species). Alternatively, the codon usage heterogeneity of those anomalous transcripts is not to be found in HGT origin.

F APPENDIX

F.1 Coding sequences belonging to the predicted regions

Region	Genome	Sequential numbers for the coding sequences as they appear in the deposited sequence of the genomes
1	A.aeolicus	273 275 276 280
2	A.aeolicus	370 374 375
3	A.pernix	1224 1225 1226 1228 1229 1230 1235
4	B.longum	192 193 195
5	B.longum	209 210 211 212 213 214
6	E.coli_K12	534 535 536
7	E.coli_K12	1991 1993 1994 1995 1996 1997
8	E.coli_K12	3545 3548 3551
9	E.coli_O157	2860 2861 2862 2863 2865 2866
10	E.coli_O157_RIMD	3507 3510 3511 3512
11	M.thermoautotrophicum	329 331 332 334
12	N.meningitidis_MC58	1887 1889 1890
13	N.meningitidis_MC58	675 676 677
14	P.aeruginosa	1369 1370 1371
15	P.aeruginosa	2222 2225 2226
16	P.aeruginosa	3143 3144 3145 3146 3147 3148 3149
17	P.putida_KT2440	4402 4403 4406 4408
18	S.flexneri2astr301	2093 2097 2098 2099 2100
19	S.oneidensis_MR1	995 996 997 1000 1001
20	S.typhiCT18	2110 2113 2114 2115 2116 2117 2118
21	S.typhiCT18	2166 2170 2173
22	S.typhiCT18	4483 4484 4485
23	X.axonopodis	1472 1474 1476 1479 1481 1484 1485 1487
24	X.fastidiosa	1714 1719 1720 1722 1723
25	X.fastidiosa	1724 1725 1726 1730 1731 1733 1734
26	X.fastidiosa	1738 1739 1741 1742 1745 1746 1747 1748
27	X.fastidiosa	1754 1757 1758 1759 1761 1762 1764
28	X.fastidiosa	1769 1771 1775 1778

Table IV-5: *Coding sequences belonging to the predicted regions (those that satisfy the filtering procedure and that represent matches of the compositional detection methodology).*

V *Applications of codon profiling II: Investigation of atypicality*

A ABSTRACT

Another application of the codon profiling technique is a procedure for the detailed analysis of those elements (genomes, transcripts or protein families) that present a codon usage atypical in a specified context.

Atypical codon usages can first be identified by multivariate analysis, and their dissimilarity to the codon usage context (for example to a genomic average bias) can be presented by a single measure, namely the Euclidean distance between the codon vectors. Subsequently, the contributions to the observed distances can be decomposed and displayed as difference matrices. Finally, the usage of the codons which contribute the most to the dissimilarity can be analysed, with the additional possibility of recovering and graphically representing the positional information (the sequential distribution along the coding sequences).

A complete methodology for the identification and study of genes with highly heterogeneous codon usage is presented in this chapter and exemplified by a real case analysis of human infecting viruses in the context of the human genome.

B INTRODUCTION

B.1 Levels of approximation and averaging effects

Transcripts with atypical codon usage can be identified at first approximation by their high Euclidean distance from the average codon bias of their genome. In this way their dissimilarity is summarised in one single measure that comprises all the contributions. This index of (dis)similarity is by its very nature extremely coarse and concise, combining the high dimensionality of the codon information into a single scalar value. A large Euclidean distance between two codon vectors could be due to the sum of many relatively small differences, or conversely to few but very significant differences in the usage of specific codons. To differentiate between such cases and properly identify the causes of codon usage atypicality, other instruments of analysis are used, for either automatic or manual investigation.

The suitable integration of complementary techniques and instruments at different approximation levels provides the researcher with the possibility of, firstly, rapid convergence to the more interesting data in the domain analysed and, secondly, extensive in-detail study of those aspects.

The Euclidean distance can be thought of as the average of the contributions of the individual codon preferences, a summarisation, which inevitably masks some information while revealing general trends. Also the synonymous frequencies are a summarisation (of the absolute codon occurrences) which reveal the trend (the preference for some codons) while masking the information about quantity, about absolute abundances. Additionally, the information on the position of codons in the gene is lost when either codon occurrences or frequencies are computed.

Similarly to these averaging effects in the methodology used, the averaging effects in the data can be recognised and opportunely exploited or circumvented. Compositional measures can be applied to data sets of different magnitudes and to different levels of biological detail.

For example the total G+C content or codon bias of whole genomes can be computed. This proves very useful in species-to-species comparisons but at the same time hides the variability inside the genomes. Alternatively, the same measures can be conducted on chromosomes or chromosomal regions (for example with a sliding window), on protein families or on single genes, with different possible aspects being investigated in each of these biological entities.

B.2 Viruses and hosts

Viruses are taxonomically classified into more than sixty families according to their genome type (like RNA or DNA based, circular or linear, double or single strand) and to their gene content. Human infecting viruses are very diverse, with genome sizes spanning from little more than a kilobase to hundreds of thousands of bases, characterised by different life-styles and different transmission routes.

As first reported in the pioneering work of Grantham *et al*. (1980, 1981) and later by Ikemura (1992) and others, the codon usage of many viruses is often quite different from the codon usage of the host organisms they infect. An exception to this trend is found in many bacteriophages which usually exhibit the codon usage of their bacterial hosts, unless they carry their own polymerase and are hence subject to different mutational pressure (Kunisawa *et al.*, 1998). In higher Eukaryotes, factors like polymerase replication errors and translational efficiency seem to play a less important role (as outlined in section I C.1.5) and this would probably also be reflected in the genome of the viruses infecting them. The codon usage of viral genomes is mostly influenced by the coexistence of many overlapping biological messages (*q.v.* section I C.3.3).

An analysis of the heterogeneity in the codon usage of viral transcripts for human infecting viruses is reported in chapter three, showing that not only the average genomic bias, but also the codon usage of the individual transcripts is significantly different from the human codon bias, with few exceptions (III D.2). But in that and in similar analyses the human genome was treated as a coherent whole, with a single average measure of codon preference, which would conceal the intra-genomic heterogeneity. It was hence decided to analyse human transcripts, comparing the codon usages of clusters of them, instead of averaging over all of them in a single genomic bias. The clusters analysed enclose transcripts which share similarity in the sequence or structure of the proteins they encode (and are hence related in the function).

The aim was to investigate possible similarities between the codon usage of viruses and of human transcripts, and to explore the possibility of shaping forces inside the viral genomes which might be influenced by the human genome. In other words, to find out whether there are niches of human codon biases towards which the viral codon usages would tend (by way of selection). The codon usage of a virus could be different from the human average bias but similar to the codon usage of particular classes of human transcripts.

C METHODS

C.1 Human protein families

Human transcripts were obtained from the *Ensembl* genome annotation project (Hubbard *et al.*, 2002; http://www.ensembl.org/). Releases 100, 110 and 120 have been subsequently used to obtain human transcripts, comparing the results and observing their consistency across the releases. Perl scripts were written to retrieve and manipulate the sequences, either through mySQL direct access to Ensembl servers or by parsing the information retrieved from *EnsMart* (Ensembl data retrieval web interface).

The transcripts are clustered according to protein sequence similarity, using the Tribes protein classification (Enright *et al.*, 2002; their database of protein family resources is accessible at http://www.ebi.ac.uk/research/cgg/tribes/) or according to the SCOP-HMM structural classification (Gough *et al.*, 2001; website available at http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/). In this way the transcripts analysed are grouped in functional classes according to the proteins they encode. 85 families were analysed from the Tribes and 130 from the SCOP classification: those with low codon profile deviation (average of intra-family distances lower than 1.7 distance units), to prevent including families containing transcripts too diverse in codon usage.

C.2 Human infecting viruses

The codon profile vectors for viral genomes (56 of them) were computed from the entries in the CUTG database (Nakamura *et al.*, 2000), which stores the codon usage information for all species that have been (even partly) sequenced. The codon usage tabulated for a species is an average of the codon usages of the individual sequences that have been determined for that species. The latest release of this database encompasses more than sixteen thousand species and strains. On one hand, this database contains redundant copies of the genes, and is thus avoided by several authors because this redundancy might introduce some bias. For the same reason it can be preferred in some studies: for example the presence of data from multiple strains of the same bacterium or virus can favour the discernment of inter-species codon usage differences over intra-species ones.

A non-redundant source for viral codon usage data is represented by the completely sequenced viral genomes, which can be obtained from the GenBank (Benson *et al.*, 2000) database. Appendix III F.2 lists the analysed viruses and their accession numbers.

C.3 Clustering algorithms

In addition to the multidimensional scaling (*q.v.* III C.4), two clustering algorithms based on unsupervised learning were jointly employed for this work. A clustering algorithm attempts to find natural groups of points based on some general criteria.

Unsupervised classification methods are used to automatically find clusters in the input data without *a priori* knowledge. They do not need to be told the number of classes in which to divide the data and they do not need a training set (supervised learning, on the contrary, implies the generation of class descriptions from labelled examples).

The Self-Organising Map (Kohonen *et al.*, 1996) is an unsupervised neural network algorithm that maps high-dimensional data to a lower dimensional grid (usually two-dimensional), with a nonlinear projection. The grid can be considered an elastic surface which iteratively updates its nodes, with the goal of preserving the structure of the high-dimensional space. The SOM_PAK program package was used in the analyses.

AutoClass (Cheeseman and Stutz, 1996) is a Bayesian classifier that finds a set of classes (with the goal of finding the most probable one) to which the data elements can be assigned, adopting a trade-off between the fit to the data and the complexity of the class descriptions. The trade-off prevents extreme (and practically useless) over-fitting, where each element would be assigned to single case classes. AutoClass searches both by trying alternative class models and by re-assigning the elements across the different classes.

D RESULTS AND DISCUSSION

D.1 Human families and human infecting viruses

For simplicity, the term *protein family* will also be used to refer to the groups of transcripts clustered according to the Tribes algorithm (transcripts clustered on the basis of the amino acid sequence they encode). This terminology may seem inappropriate, but could be partly excused since the results shown for the Tribes clusters are highly comparable with those obtained using transcripts belonging to real protein families – as defined in the SCOP (Murzin *et al.*, 1995) structural domain database.

D.1.1 Clustering of Tribes families

Among the clusters analysed, the protein family 13122 (marked with a *green pointer* in the figures) was used as a *control*, as the representative of the average human codon bias. Its annotation is “Cytochrome P450” and groups 37 transcripts. It was chosen as representative because it has a confirmed annotation and because its codon usage is almost the same as the average one for *Homo sapiens*.

The codon profile vectors relative to the human protein families and to the human-infecting viruses have been clustered combining two different multivariate analysis methods: a neural-net based clustering (SOM), and an unsupervised classifier (AutoClass).

The clustered map of protein families and human-infecting viruses (Figure V-1) shows that the majority of human families occupy a well defined (although broad) space, not too far from the average human bias and, as expected, distant from the majority of the viruses.

AutoClass suggests fives classes in which the data can be sorted, marked with a different colouring on the SOM of Figure V-1.

Classes identified by the colours yellow, magenta and violet are mainly populated by human protein families. Some Herpes and Adenovirus (Adenovirus types 2, 5 and 17; Herpes virus types 1, 2, 4 and 5) fall in these classes, their codon usage being similar to the average human bias, possibly an indication of adaptation to the human codon usage. This result is consistent with the study on viral transcripts presented in the previous chapter (*q.v.* III D.2). The yellow class also includes three families of histone transcripts, which are generally considered among the genes with the highest expression levels. They have very biased codon usages which might be due to selection for rapid translation of mRNA (Wells *et al.*, 1986; DeBry and Marzluff, 1994; but see Kanaya *et al.*, 2001a).

The green-coloured class groups viruses whose codon usage is most dissimilar from that of human proteins. Among these are the Papilloma viruses, *rotavirus* and *torovirus*. The last remaining class (coloured in light blue) can be considered an *interface* class, comprising several viruses and human protein families with atypical codon usage. Some of these families are very close (in the clustering and hence in terms of codon usage) to a group of viruses, in particular to the RNA viruses parainfluenza and Human Immunodeficiency Virus – HIV.

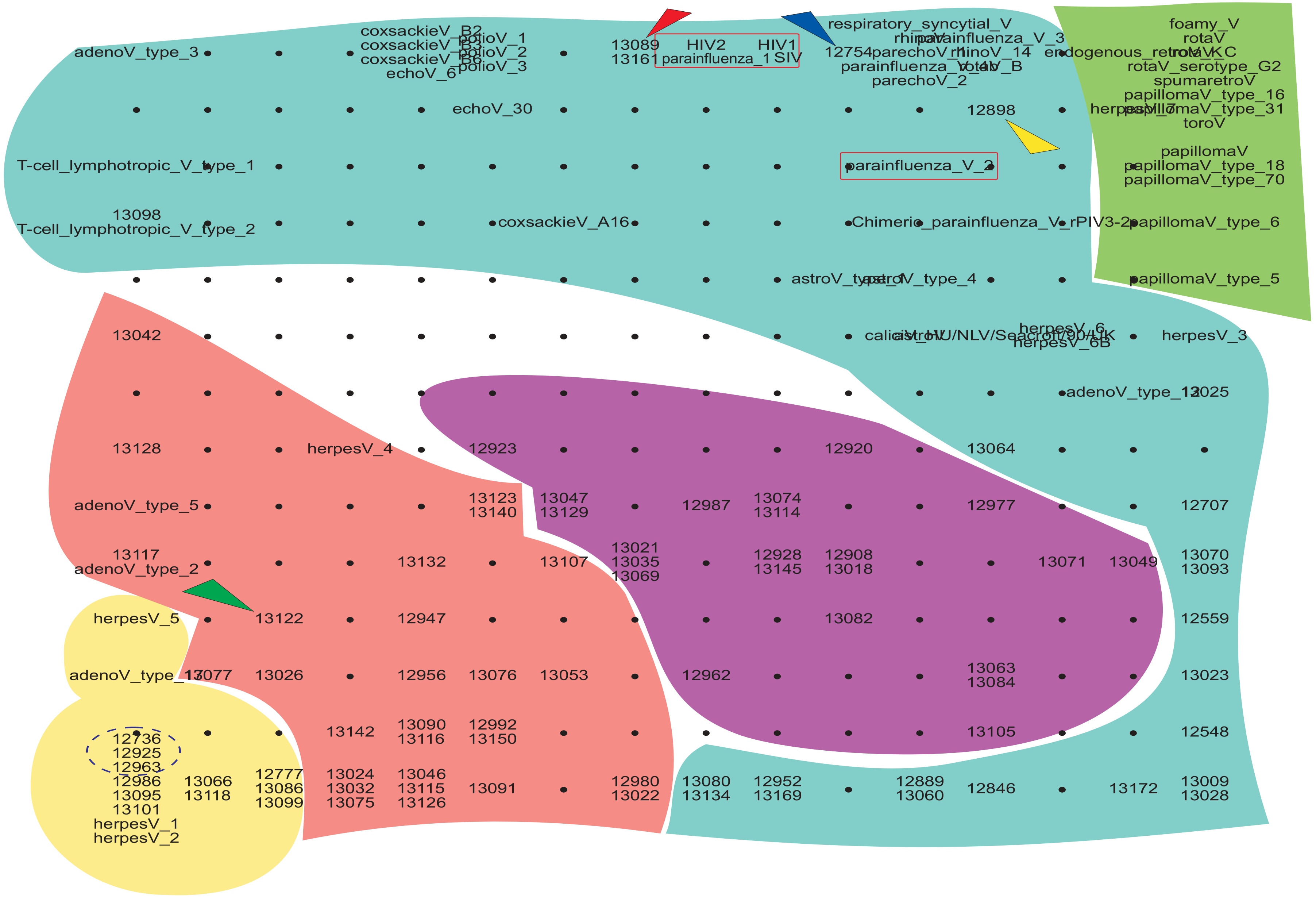


***Figure V-1**: Codon profile Self-Organising Map with Tribes human protein families (the labels correspond to the Ensembl family identifier, ENSF) and human infecting viruses. The coloured areas indicate classes identified by AutoClass. Red rectangle (top centre of the map): Immunodeficiency and parainfluenza 1 viruses - Green pointer (bottom left): 13122 (Cytochrome P450) - Blue pointer (top centre): 12754 (hnrnp) - Red pointer (top centre): 13089 (L1 reverse transcriptase) - Yellow pointer (top centre): 12898 (RNA binding protein) - Blue ellipse (lower left corner): histones - Map information: average distance between data points 1.311 (with standard deviation 0.624); maximum distance 3.836 (between Rotavirus in the top right corner and family 12925 in the bottom left corner).*

Table V-1 shows the human protein families that are closer to these viruses than to the average human codon bias. The transcripts in these families all show a very atypical (when compared to the human codon bias) codon usage. The most atypical is 12898 (annotated as *RNA binding protein*), which has an Euclidean distance from the average human codon bias of 1.680, almost three standard deviations further away than the average distance of all the human families from the human codon bias. The values for these atypical transcript clusters can be compared to the histogram showing the distribution of distances for all the human protein families (Figure III-8, from section III D.3.1): over 75% of human protein families have a distance from the human bias less than 1 and the average for all families is 0.657.

All these families that have a codon usage similar to that of parainfluenza and HIV are annotated as being functionally related to RNA and DNA: RNA binding, retrotranscriptase, DNAse I, RNA polymerase.

	Distance from family					Average distance from all families	Deviation of distances from all families	Maximum distance from all families
	control 13122	12898	13089	13161	12754			
Homo sapiens	0.574	1.680	1.589	1.307	1.554	0.657	0.368	1.680
parainfluenza1	1.723	1.204	0.959	1.055	1.376	1.424	0.387	2.575
parainfluenza2	1.998	0.828	1.269	1.186	1.259	1.633	0.466	2.862
HIV-1	2.005	1.075	1.309	1.092	1.102	1.671	0.432	2.929
HIV-2	1.687	1.184	1.012	0.912	1.311	1.409	0.386	2.573
Total transcripts	37	32	74	463	21	3837 (in 85 families)		

***Table V-1**: Codon profile Euclidean distances between Tribes human protein families and the genomic biases of human, parainfluenza and HIV. Consensus annotations: 13122: Cytochrome P450; 12898: RNA binding protein; 13089: LINE1 reverse transcriptase; 13161: LINE1 retrotransposon; 12754: Heterogeneous nuclear A1 helix destabilising protein single strand binding protein HNRNP core protein.*

13089, 13161: LINE1s (Long INterspersed Element 1) are long (6-8kb) GC-poor transposable sequences (accounting for 15% of the human genome) encoding an RNA binding protein and a reverse transcriptase/endonuclease (Smit, 1996).

12898: The genes coding for the transcripts in the RNA binding protein cluster are located in the Y chromosome and encode a nuclear protein implicated in spermatogenesis. It has been proposed that these genes arose from transposition of an ancestral autosomal gene, hnRNPG (Chai *et al.*, 1998).

12754: The heterogeneous nuclear ribonucleoproteins (HNRNP) have a general role in processing, packaging and transport of RNA but some of them display also sequence-specific binding (Krecic and Swanson, 1999; Shan et al., 2000).

D.1.2 Clustering of SCOP superfamilies

The results obtained on the Tribes families were compared to those obtained from the analysis performed on transcripts grouped according to the SCOP-HMM structural classification (Gough *et al.*, 2001). In this way two different classification schemes were used to group the transcripts, to assess to what extent the observed correspondences are dependent on the family clustering. The resulting SOM map (Figure V-2) has a very similar topology, but since the clustering of the transcripts is different (based on Hidden Markov Models of structural domains rather than on sequence similarity), it is not quite identical. AutoClass identifies in this data set one additional class which comprises *coxsackie* and *polio* viruses (grouped in the light blue class by the previous analysis).

The majority of the transcripts that *Tribes* classification groups in the families 12898 and 12754 (RNA binding proteins and HNRNP) are part of the SCOP superfamily d.58.7 (Annotated as "RNA-binding domain, RBD").

Other SCOP protein superfamilies close to HIV codon profiles are shown in Table V-2. Being superfamilies, they contain in general more transcripts than Tribes families – the average family size is twice as large – and because of this fact their average distance to the human bias is lower.

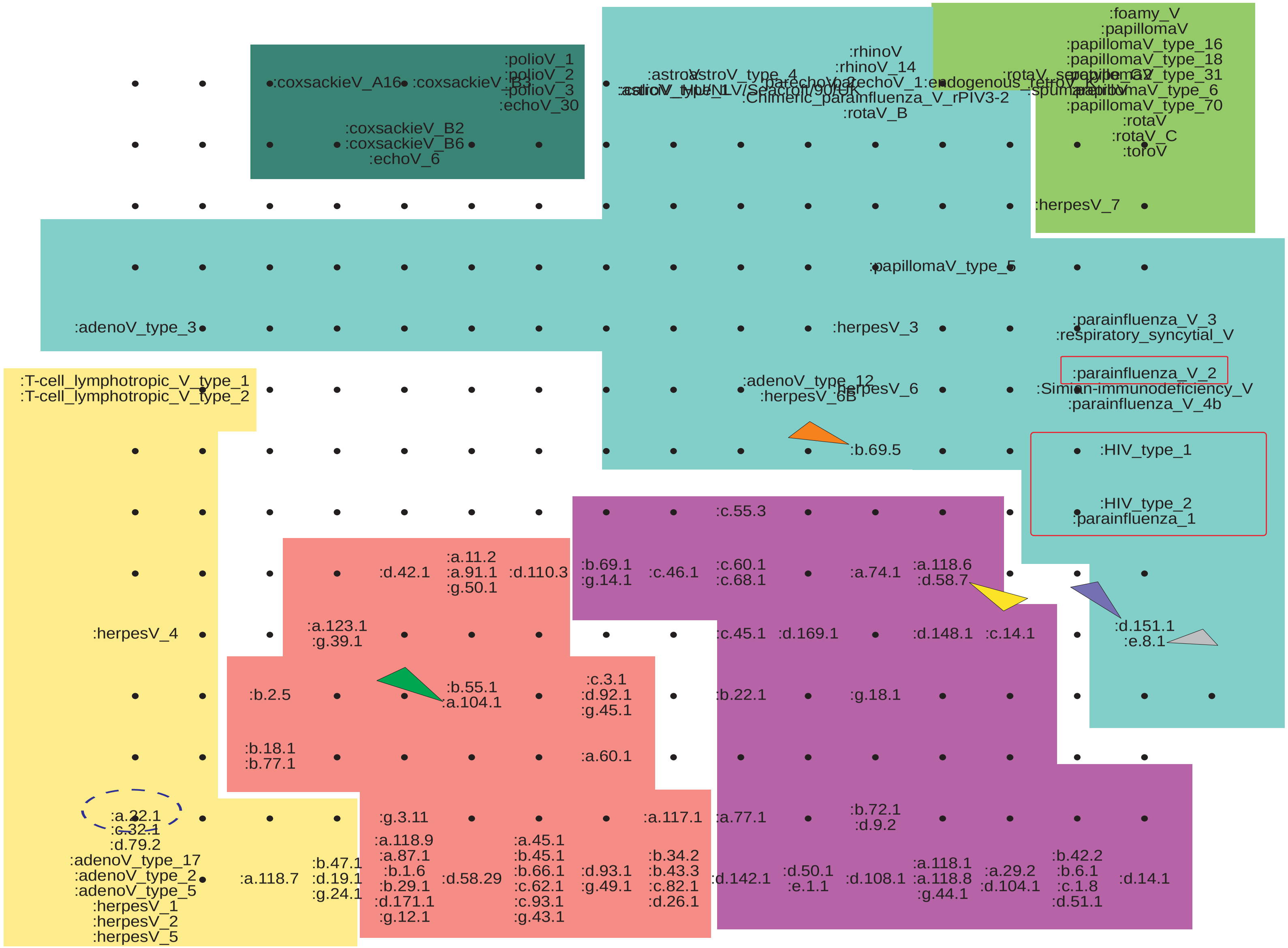


Figure V-2: *Codon profile Self-Organising Map with SCOP human protein families (the labels correspond to the structural superfamily code) and human infecting viruses. The coloured areas indicate classes identified by AutoClass. Red rectangle (right centre of the map): Immunodeficiency and parainfluenza 1 viruses - Green pointer (bottom left): a.104.1 (Cytochrome P450) - Yellow pointer (right centre): d.58.7 (RNA binding domain) - Gray pointer (right centre): e.8.1 (DNA/RNA polymerases) - Violet pointer (right centre): d.151.1 (DNase I-like) - Orange pointer (right centre): b.69.5 (Regulator of chromosome condensation) - Blue ellipse (lower left corner): histones - Map information: average distance between data points 1.179 (with standard deviation 0.643); maximum distance 3.918 (between Rotavirus in the top right corner and herpes 2 in the bottom left corner).*

	Distance from family					Average distance from all families	Deviation of distance from all families	Maximum distance from all families
	control a.104.1	e.8.1	d.151.1	d.58.7	b.69.5			
Homo sapiens	0.342	1.324	1.258	0.985	0.917	0.359	0.221	1.324
parainfluenza1	1.536	0.801	0.823	1.008	0.952	1.312	0.237	1.996
parainfluenza2	1.788	1.128	1.170	1.067	0.911	1.516	0.292	2.289
HIV-1	1.788	1.173	1.202	1.164	1.027	1.557	0.262	2.313
HIV-2	1.491	0.834	0.863	0.959	0.943	1.321	0.233	1.997
Total transcripts	81	80	90	273	20	12093 (in 130 families)		

***Table V-2**: Codon profile Euclidean distances between SCOP-HMM human protein superfamilies and the genomic biases of human, parainfluenza and HIV. Superfamily descriptions: a.104.1: Cytochrome P450; e.8.1: DNA/RNA polymerases; d.151.1: DNase I-like; d.58.7: RNA-binding domain; b.69.5: Regulator of chromosome condensation RCC1*

D.1.3 Clustering repeated together with the pufferfish genome

To verify the hypothesised similarities between human genome families and human infecting viruses, the clustering was repeated in conjunction with families from the recently sequenced genome of the pufferfish, *Takifugu rubripes* (Figure V-3). In this case multidimensional scaling (MDS) was used, a multivariate ordination procedure that preserves Euclidean distances in the low-dimensional plot (III C.4).

Apart from some exceptions – like the protein family with identifier *f33518* (annotated as "reverse transcriptase/ribonuclease H") that clusters near some *polio* and *coxsackie* viruses – pufferfish transcripts generally have a different codon usage from that of the human infecting viruses; human families are located in between the regions of the map occupied by the codon usages of the pufferfish and of the viruses.

This new clustering confirms that the observed correspondences are not methodological artefacts: human and pufferfish families are separated like in Figure III-6 from section III D.3.1, while the topological arrangement of human infecting viruses (with respect to human protein families) is highly comparable with that found in the previously presented analyses.

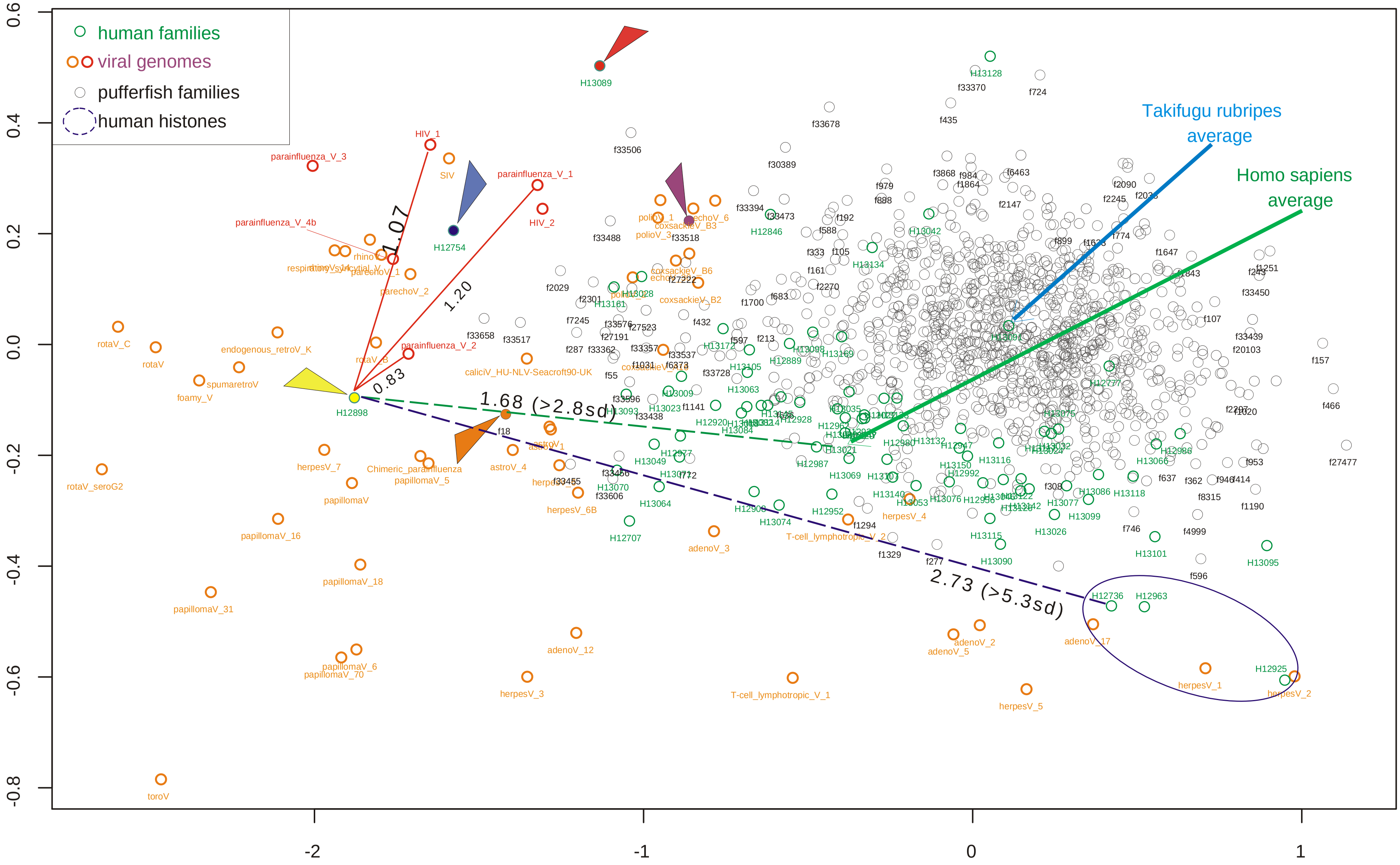


__Figure V-3__: Multidimensional scaling map of human families (in green), pufferfish families (in dark gray) and human infecting viruses (in orange colour, parainfluenza and immunodeficiency viruses marked in red). The Euclidean distances of RNA binding protein family from the human codon bias and from HIV 1 are indicated. - Blue pointer (left centre): 12754 (hnrnp) - Red pointer (top centre): 13089 (L1 reverse transcriptase) - Yellow pointer (left centre): 12898 (RNA binding protein) - Orange pointer (left centre): f18 (L1 reverse transcriptase, T.rubripes) - Violet pointer (top centre): f33518 (reverse transcriptase/ribonuclease, T.rubripes) - Blue ellipse (lower right corner): histones. sd=standard deviations.

D.2 The human RNA binding protein

The cluster of RNA binding transcripts was found to have one of the most atypical codon usages in both SOM and MDS multivariate analyses. There are in fact significant differences between the codon usage of the RNA binding protein family (family identifier 12898 in the Tribes analysis) and the average human codon bias, while its differences with the HIV or with parainfluenza are minor (as shown in Table V-1). By contrast, Cytochrome 450 (13122) shows almost no difference from the human bias, while its difference from retroviral genomes is very pronounced. These differences, which were summarised by the Euclidean distances in the previous sections, are more accurately shown by the codon difference matrices (II C.3), which reveal the exact causes of the codon usage dissimilarities. The single scalar value represented by the dissimilarity measure is thus decomposed into its multivariate components.

The major cause accounting for the observed atypicality of the RNA binding protein family is AT-richness, with some synonymous sets particularly contributing to it, as detailed in Figure V-4. That figure also shows a comparison between the codon usage of the RNA binding transcripts and those of parainfluenza and HIV.

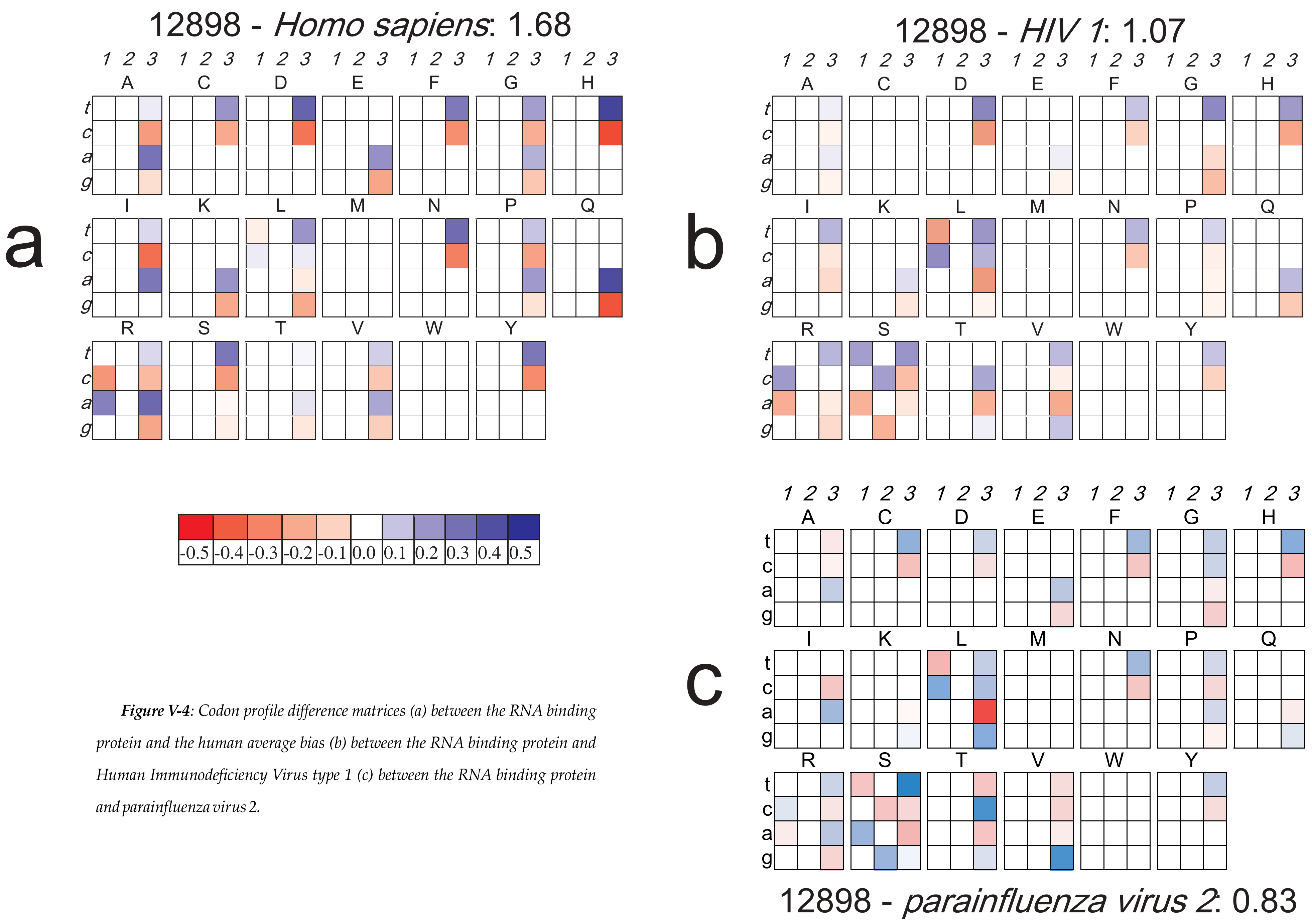


Figure V-4: *Codon profile difference matrices (a) between the RNA binding protein and the human average bias (b) between the RNA binding protein and Human Immunodeficiency Virus type 1 (c) between the RNA binding protein and parainfluenza virus 2.*

D.3 Coloured codon analysis of RNA binding proteins

The investigation has now progressed from the analysis of whole genomes to the analysis of protein families, and from the presentation of scalar measures of dissimilarities to their decomposition into synonymous frequencies, at each step increasing the level of detail and decreasing the averaging effects. The final logical step is to analyse the single transcripts individually, and to recover the information on the absolute codon occurrences and on the location of the codons along the sequences. These informations are both ignored when the codon usage is analysed with relative frequencies.

Two sets of synonymous codons that are major contributors to the atypicality (in the context of the human codon usage) of the RNA binding protein are presented in this section by the individual codon occurrences along the single transcripts.

In the following figures (Figure V-5 and Figure V-6), every line represents a sequence belonging to the analysed protein families and only the relative usage of the triplets coding for a single amino acid is shown. For each transcript the synonymous codons for the analysed amino acids were isolated and plotted using the corresponding symbol (a coloured codon; *q.v.* II C.5.1). The symbols used, their correspondence to the codon triplets and the average codon bias of these triplets in *Homo sapiens*, are reported in the figure legends. This representation of the individual triplets helps the visual inspection of sequences, and in particular allows the observation of patterns in the usage of the triplets along a sequence. The recovery of positional information would reveal, for example, the presence of a rarely used codon at the beginning of the transcript (a fact believed to play regulatory effects in the genes of the bacterium *Escherichia coli*; Ohno *et al.*, 2001) or would make the presence of different usages in different regions of the genes visible.

In all transcripts of the RNA binding protein (12898) it is possible to observe an almost exclusive preference, 87%, for a synonymous codon for Histidine (CAT) whose average codon bias in *Homo sapiens* is tabulated as being 41%. The Cytochrome P450 transcripts instead show a distribution closer to the average, with the relative frequency of the CAT triplet being 32% (Figure V-5).

As for the amino acid Arginine, the RNA binding protein transcripts have a very high preference for the AGR codons (represented as squares with black borders, accounting for 65% of total Arginine codons) and in particular for AGA (represented as a yellow square with black border, accounting for 53%) whose usual distribution would be the one appearing in the Cytochrome P450 transcripts, since its average codon usage in the human genome is tabulated as 21% (Figure V-6).

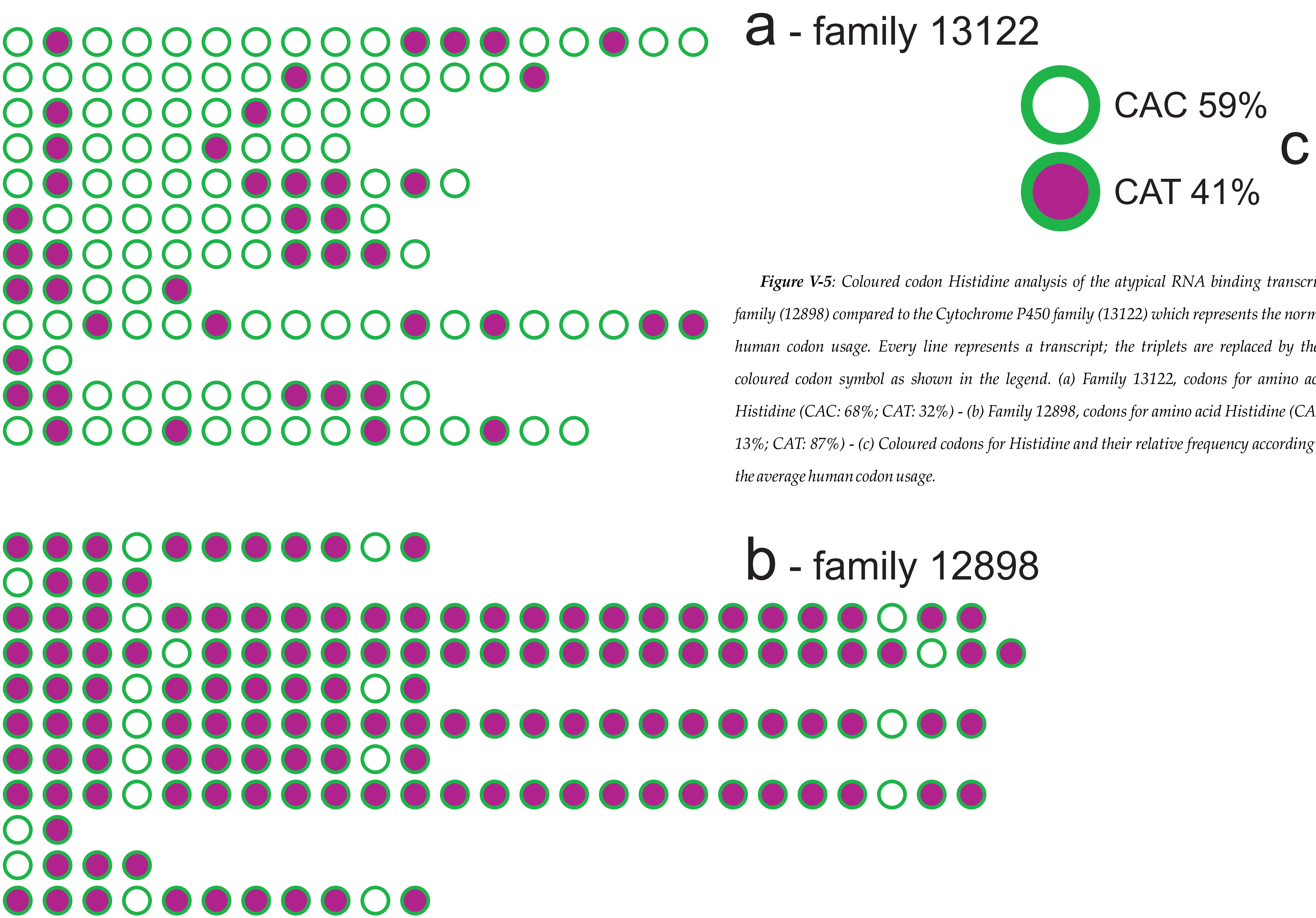


Figure V-5: Coloured codon Histidine analysis of the atypical RNA binding transcript family (12898) compared to the Cytochrome P450 family (13122) which represents the normal human codon usage. Every line represents a transcript; the triplets are replaced by their coloured codon symbol as shown in the legend. (a) Family 13122, codons for amino acid Histidine (CAC: 68%; CAT: 32%) - (b) Family 12898, codons for amino acid Histidine (CAC: 13%; CAT: 87%) - (c) Coloured codons for Histidine and their relative frequency according to the average human codon usage.

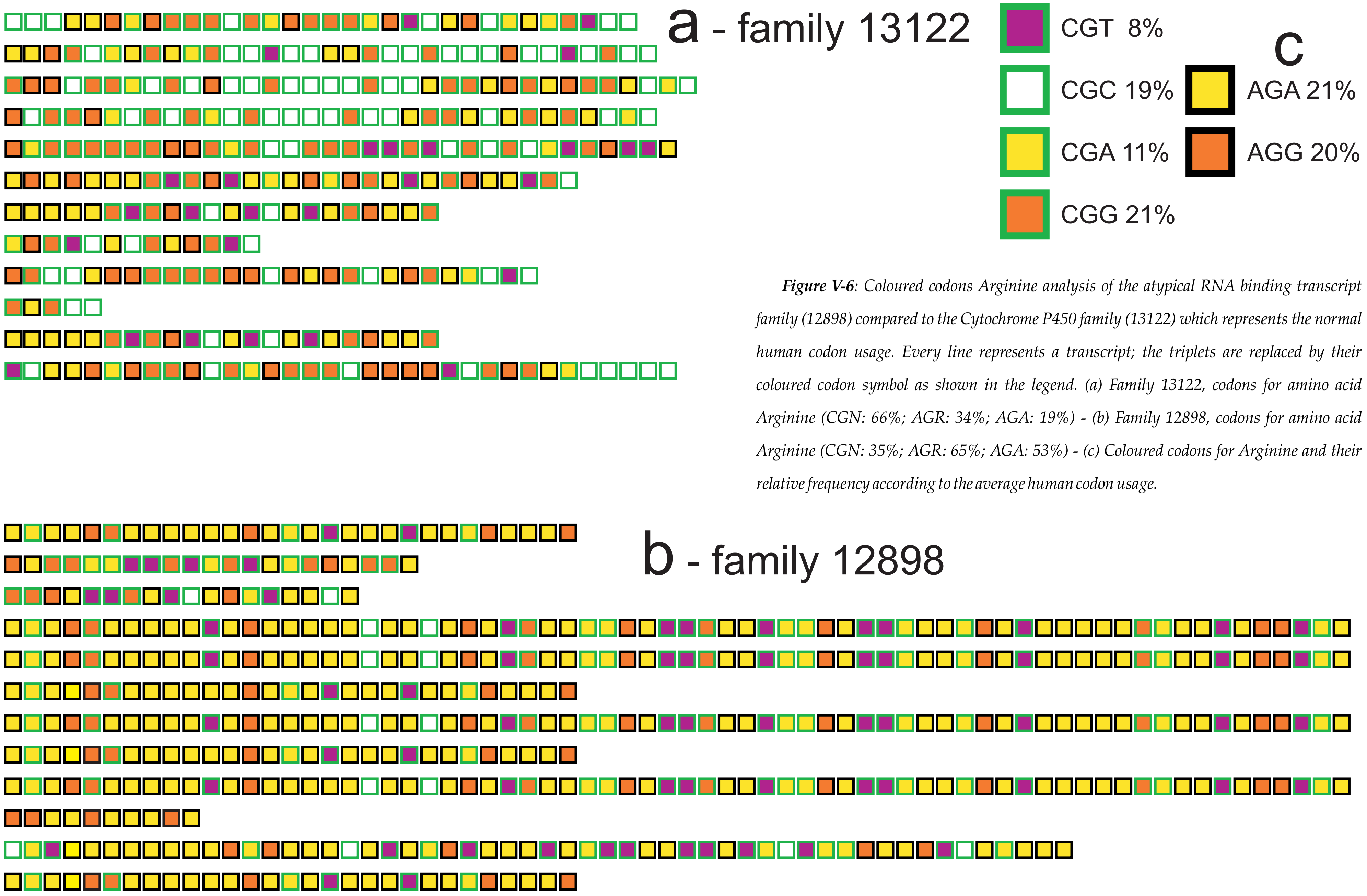


Figure V-6: *Coloured codons Arginine analysis of the atypical RNA binding transcript family (12898) compared to the Cytochrome P450 family (13122) which represents the normal human codon usage. Every line represents a transcript; the triplets are replaced by their coloured codon symbol as shown in the legend. (a) Family 13122, codons for amino acid Arginine (CGN: 66%; AGR: 34%; AGA: 19%) - (b) Family 12898, codons for amino acid Arginine (CGN: 35%; AGR: 65%; AGA: 53%) - (c) Coloured codons for Arginine and their relative frequency according to the average human codon usage.*

D.4 Hypotheses to explain the observed similarities

The exact reasons that account for the similarity between the codon usage of the human RNA binding proteins and that of the parainfluenza and HIV genomes are currently unknown and would require further investigation. Three main hypotheses can be outlined as follows:

First of all, the simplest and most probable hypothesis is that the similarity is just coincidental, the result of the overlap between the viral codon usages and the codon usage of those significantly atypical protein families. The reasons for the atypicality of these protein families are most probably not related to the forces shaping the codon usage of the viruses.

The second hypothesis is related to translational efficiency. Both the human RNA binding protein families and most of the viruses have a codon usage which is AT-rich and very different from that of the histones. This could indicate a suboptimal (from the point of view of expression levels) codon usage. By contrast, several herpes viruses and adenoviruses have very similar codon usage to that of the histones (Table V-3).

	Distances from histone families		
	Tribes 12736	Tribes 12925	Tribes 12963
Homo sapiens average	1.078	1.553	1.204
RNA binding protein	2.512	3.031	2.647
Parainfluenza 1	2.164	2.575	2.202
Parainfluenza 2	2.404	2.862	2.495
HIV-1	2.363	2.929	2.531
HIV-2	2.121	2.573	2.252
Herpes 1	0.940	0.976	0.779
Herpes 2	1.117	0.933	0.913
Adenovirus 17	0.768	1.065	0.836
Total transcripts	21	25	25

***Table V-3**: Codon profile Euclidean distances of the RNA binding protein family (Tribes 12898) and of the genomic biases of human and several viruses from the human histone families. Consensus annotations: 12736: Histone H3, 12925: Histone H2B, 12963: Histone H2A.*

Inefficient translation due to codon usage was in fact reported for HIV transcripts, and several groups studying vaccine approaches against this virus engineered HIV transcripts with optimised codon usage (Haas *et al.*, 1996; zur Megede *et al.*, 2000; Deml *et al.*, 2001). For the viruses, low levels of expression could be a way of evading detection by the immune system or it could be the result of stronger constraints on their codon usage that would prevent adaptation towards higher translational efficiency. In fact, the HIV codon usage is one of the most constrained: more than 90% of its sequence is coding and certain regions code simultaneously for two or even three genes (each one following one of the three possible codon reading frames). For the identified transcripts with atypical codon usage, a lower translational efficiency could be one of the ways to regulate their expression levels.

A third possibility (related and complementary to the previous one) would be the necessity of preserving particular mRNA secondary structures in these transcripts (for example in relation to their stability or, conversely, propensity for degradation; I C.3.3). Recently published studies by Peleg and co-workers (2002; 2003) seem to be pointing in this direction. Their work analysed the sequence conservation and the possible RNA structure of HIV transcripts and confirmed the presence of highly conserved RNA folds in the coding regions for the proteins Env (envelope glycoprotein) and Nef (whose function is not completely understood but which has been observed as being involved, among others, in down-regulation, alteration of cellular signalling and RNA binding; Geyer *et al.* 2001; Echarri *et al.*, 1996).

The transcripts of the identified protein families (such as RNA binding proteins, Line1 retrotranscriptase and HNRNP) are very rich in AT, with GC3 values between 30% and 40%, while the average GC3 for the coding part of the human genome is 58%. They could contain inhibitory sequences that reduce mRNA stability and inhibit translation.

A characterisation with molecular biology techniques (for example assaying the degradation of these mRNA transcripts) would be the next step to evaluate these hypotheses, integrating the complementary disciplines of bioinformatics and molecular biology.

E CONCLUSIONS

A complete procedure to identify and characterise genes with anomalous codon usage was presented. The atypicality can be observed and analysed at different scales (different sample groups), outlining codon usage patterns for whole genomes, for transcripts taken in clusters or for single sequences. Furthermore, the multivariate codon information can be summarised in a single measure or observed in all its components. The coarser levels of detail enable easy observation of general trends and faster convergence to the most interesting domains, which can then be extensively explored, recovering the information that had been ignored or concealed by computation of average measures.

A real case scenario was investigated with this procedure, employing different techniques and data sets, identifying some human protein families whose codon usage is very atypical in the human genome context but similar to that of the viruses parainfluenza and HIV. Conversely, several adenoviruses and herpes viruses were shown to have codon usage patterns similar to those of the human histones.

VI Codon usage space

A Abstract

The *codon usage space* is the multidimensional space of all the possible synonymous codon distributions. An investigation of this space was conducted, with particular focus towards the examination of the portion represented by currently available biological sequences. The interest lies in the characterisation of correlations between non-synonymous triplets, hence evaluating the non-randomness of the codon usages, and in the discovery of universal constraints.

A binning algorithm was used to create a model of the codon space at a desired level of granularity, reducing the data set to a lower number of codon usages, each representing a populated region of the entire space. In this way the continuum of the codon space is modelled by a discrete grid of codon vectors, uniformly spaced, in which a large part of the sequencing sampling bias (more data available for specific taxonomic groups) has been removed.

This enabled the identification of the major components in the variation among the codon usages, those possibilities for variability which have been most 'explored' by extant codon usages. It hence outlined the order and the characteristics of the global constraints to the possible variability.

The biological and the theoretical space were compared, revealing the high degree of correlation between synonymous triplets in the whole populated space, estimating its coverage and compactness, and showing the very confined region of the theoretical space in which codon usages can be found.

B Introduction

B.1 Defining the codon space

The *codon usage space* (from here on referred simply as *codon space*) is here defined as the set of all possible codon usages, all the possible relative utilisations of the synonymous triplets in coding sequences. The concept of codon space parallels the one of *protein fold space* which encompasses all possible protein folds (Holm and Sander, 1996; Holm and Sander, 1998; Zhang and DeLisi, 1998). It was first analysed by Rowe and co-workers in 1984, using nucleotide frequencies at the three codon positions (Rowe *et al.*, 1984; Rowe, 1985).

A total *codon profile space* would include all possible codon profile vectors deriving from all possible genetic codes, while the codon space presented in this chapter is restricted to codon usages from the Standard code.

An *incomplete codon usage* is a codon usage vector missing information for particular amino acids or triplets. A total codon space also encompasses all the possible incomplete vectors, but the analyses will be mostly restricted to filtered spaces containing only complete codon usages.

The codon space can be analysed as the multidimensional space of vectors whose components are the relative frequencies for synonymous triplets, with each vector in this space representing a codon usage. A measure like the Euclidean distance between any two vectors (II C.2) can be used to assess the degree of similarity between the codon usages associated to the vectors. With this distance function defined on it, the codon space is a metric space, as it satisfies the three properties of non-negativity, symmetry and triangle inequality.

B.2 Measuring the size of the codon space

B.2.1 The number of vertices

How large is a codon space? In particular, narrowing the focus of interest to a better-defined subset of the total codon space, in how many ways can the codons be used in a coding sequence, so that there is at least one triplet coding for each amino acid? This subset of the space takes the name of *AA-filtered space* (see II C.4 for vector filtering schemes).

One possible answer to this question is given by calculating the number of *minimal codes* (all possible set of codons with only one triplet present for each amino acid species) for the Standard genetic code. The product of the number of synonymous triplets for each amino acid computes the total number of minimal codes as amounting to roughly 340 million:

$$4\cdot2\cdot2\cdot2\cdot2\cdot4\cdot2\cdot3\cdot2\cdot6\cdot2\cdot4\cdot2\cdot6\cdot6\cdot4\cdot4\cdot2=339{,}738{,}624$$

These represent the vertices of the AA-filtered space, the extreme perimeter. For a better understanding of this concept, a two-dimensional space can be considered, where the vectors have components restricted to the values between 0 and 1. All the possible vectors can be represented by the points inside the square ABCD, where A=(0,0), B=(0,1), C=(1,1) and D=(1,0). There are infinite points in this space but all are restricted to the region inside the square, with the extremes being the four vertex points A, B, C and D.

In the multidimensional space of AA-filtered codon usages, there are 340 million extreme vertices, defining the boundaries between which all the other data points can be found (all the other codon usages with more than one triplet for each amino acid).

Two other subsets of the codon space, the CPRO-filtered and CSYN-filtered space (II C.4), lie inside the AA-filtered one, at a certain non-zero distance from those extremes, but with limits approaching those of the AA-filtered one: when for example one synonymous triplet for Isoleucine has a relative usage of 96% and the other two only 2% (like in the case of the *Streptomyces coelicolor* genome), it approaches the extreme case of a minimal code data point (which would be 100% - 0% - 0%).

B.2.2 The number of combinations

Another possible way of estimating the extension of the codon space is by restricting the estimation to the Standard code CSYN-filtered space (all synonymous triplets present, ignoring the terminator/STOP ones), with a minimum frequency of 0.1 (*i.e.* 10%) for each triplet, and considering frequencies only in multiples of 0.1 (0.1, 0.2, 0.3, …, 1.0). This is equivalent to setting a granularity in the space, where only these discrete values are allowed, representing the continuum of real values in between.

With these boundary conditions it is possible to calculate how many possible distributions of frequencies for each set of synonymous triplets exist (*e.g.* for the four triplets coding for Alanine) and hence (by multiplication of the distributions for each amino acid) how many possible codon usages exist.

The frequency distributions of two synonymous triplets are nine (0.9:0.1, 0.8:0.2, 0.7:0.3, 0.6:0.4, 0.5:0.5, 0.4:0.6, 0.3:0.7, 0.2:0.8, 0.1:0.9) and there are nine amino acids coded by two triplets. For three synonymous triplets (the case of the Isoleucine amino acid) there are 36 possible distributions of frequencies. There are 84 possible distributions for four synonyms (and there are five amino acids coded by four codons) and 126 in the case of six synonymous triplets (amino acids Leucine, Arginine and Serine in the Standard code).

The total number of relative synonymous codon frequencies for the entire codon usage, with these boundary conditions, would hence be: $126^3 \cdot 84^5 \cdot 36 \cdot 9^9 = 1.17 \cdot 10^{26}$, clearly an unmanageable number of possibilities to analyse by exhaustive search methods. Since some combinations of distributions in the cases of six synonymous triplets are equivalent from the point of view of codon profiles (as explained in II D.1.1), the number of codon profile vectors is slightly smaller: $6.752 \cdot 10^{25}$.

The shortest Euclidean distance between two data points in this space amounts to 0.141 units (the square root of $0.1^2+0.1^2$). Two codon usages with an inverted extreme bias for a synonymous couple (for example one having 90% usage of Phe_TTT and the other one having 90% usage of Phe_TTC) have vectors whose distance in the codon space is 1.13 units ($\sqrt{(0.9-0.1)^2+(0.1-0.9)^2}$). The longest possible distance is the one between two codon usages with completely inverted biases in the codon frequencies for all amino acids. It amounts to 4.28 units. This is the measure of the longest diagonal of this multidimensional space.

If only the most extreme codon usages were considered (such as 0.9:0.1 distributions for synonymous couples or 0.7:0.1:0.1:0.1 in the case of four synonymous triplets), we return to the number of the *minimal codes* previously computed: $6^3 \cdot 4^5 \cdot 3 \cdot 2^9 = 339{,}738{,}624$.

B.3 Granularity of the codon space

Setting *coarse* frequencies (multiples of 0.1) can be thought of as the equivalent to setting the *granularity* in the codon space, with elements existing as discrete entities rather than as a continuum. The 10^{25} elements calculated above for the codon space are a huge but finite number of representatives for the infinite codon usages. Each of those elements stands for all the possible codon usages inside a hypersphere with the radius 0.141 (√2/10): a grid of points uniformly spaced at 0.141. This process is usually defined as *binning* of the data points.

Choosing a higher granularity setting – *i.e.* choosing a longer radius for the hypersphere whose centre is taken as representative for all the points inside its radius –

lowers the number of total points of the codon space grid. In other words, the number of total possible theoretical codon usages that exist at a higher granularity radius is less.

The following Table VI-1 shows the numbers of grid points (in orders of magnitude) for each granularity level (the radius of the representative spheres).

granularity radius	*number of points*
0.283	10^{18}
0.424	10^{14}
0.566	10^{11}
0.707	10^{9}
0.849	10^{7}
0.990	10^{6}
1.414	10^{4}

Table VI-1*: Estimated number of grid points – bins – for each granularity level of the CSYN-filtered codon space, in orders of magnitude. Estimations were obtained computing the number of points that could be found at the given distance (the granularity radius) in exhaustive subsets of the codon vectors (all the possible distributions of coarse frequencies for a subset of the total codon dimensions). These estimations are in agreement with those obtained from randomly generated codon usage vectors used to sample the codon space, although the sampling is feasible only for longer granularity radii (q.v. C.2).*

Longer radii are best used to explore the codon space. The coarser granularities enable dealing with the vast amount of possibilities, and attempt to chart the codon space at a certain level of detail, which could, if necessary, be increased – analogous to changing from a large scale map to one of a smaller scale. In addition to the requirement of feasibility, the granularity radius needs to be sufficiently large to effectively sample the space and to group similar codon usages, without resulting in bins containing single elements. As will be shown in the **Methods** (C.1.1), there is a definite range of radii that should be used for useful and coherent binning.

B.4 Exploring the codon space

In order to chart the codon space – to create a map which represents the codon usages employed by genomes – the concept of *representative hyperspheres* was applied. When analysing all codon usages coming from all the coding sequences determined so far, only one representative element can be kept for each region of the codon space in which codon usages cluster.

All the transcripts whose codon usage closely reflects their average genomic bias are clustered together. Only one representative kept, standing for occupancy of that region of the codon space, while atypical transcripts would need their own representatives. In this way the populated space (the region of the codon space that encompass the codon usages observed in sequenced genes) can be mapped at a defined level of granularity. In other words, all codon usages contained inside a hypersphere whose radius is the granularity radius, are represented by a single codon vector, located at the centre of the hypersphere. The number of elements needed to characterise the dataset is reduced, since each selected codon vector represents a class of codon usages that shares similar features.

The non-populated space can also be analysed. These are the remaining zones of the codon space for which no equivalent codon usage can be observed in the available sequence data. If the codon space mapping is imagined as the charting of land masses, the non-populated space could be thought of as the oceanic regions.

Comparisons of the populated space with the non-populated regions can help in understanding the size and structure of the former, with the scope of characterising universal constraints and correlations in codon usages. Why are these regions non-populated? In other words, why it is not possible to observe those codon usages in nature? Have they not been observed yet (because still so little has been sequenced) or are they always avoided in coding sequences due to some constraint? If they are avoided, why is it so? Are there "universally" optimal codons, or conversely, universally under-represented ones? Similarly, are there regions preferentially occupied or avoided by some taxonomical group? Is there any broad codon usage pattern preferentially found or never found within a group of similar organisms? The charting of the codon space is a way of providing answers to these and similar questions.

B.5 Fractal structures

In 1975 Mandelbrot introduced the term fractal (from Latin *fractus*: broken) to describe those phenomena that are continuous but not differentiable. Every attempt to divide a fractal into smaller parts results in the resolution of more structure: fractals are said to display *self-invariant* properties. Fractal structures will look the same regardless of the observation scale: the property of *scale independence*. Other fractal properties are self-similarity, self-affinity, complexity and infinite length or detail.

Recognition of the fractal geometry of nature has important implications to biology, as evidenced by the numerous applications: fractal properties have been studied – to name but a few examples – for chromosome architecture (Takahashi, 1989), protein surfaces (Lewis and Rees, 1985), cellular complexity (Smith *et al.*, 1989), DNA sequences (Xiao *et al.*, 1995), branching systems in the organs of animals or in the plant structures (Deering and West, 1992; Fitter and Strickland, 1992) and in the relations between size and populations of organisms (Jeffries, 1993).

Formally, a mathematical fractal is defined as any series for which the Hausdorff dimension D (a continuous function) exceeds the discrete topological dimension (Tsonis and Tsonis, 1987). The *fractal dimension* D is most commonly estimated from the regression slope of a log-log plot.

Unlike theoretical curves (such as the Koch curve or Sierpinski gasket), natural structures do not display exact self-similarity but many display some degree of statistical self-similarity (at least over a certain range of spatial or temporal scales) and are thus better referred to as *scale invariant* (Vicsek, 1989).

One of the first methods used to empirically estimate the fractal dimension is the *dividers method*, in which the length of a fractal curve is measured at various scale values. This procedure is analogous to moving a set of dividers of fixed size along the curve. By measuring the contour using different sizes of dividers one finds that, if the object is fractal, the length of the contour will increase as the size used to measure the contour is decreased.

In some cases the log-log plot does not have a constant slope (*i.e.* the fractal dimension is not constant). This may indicate different generative processes or it may simply reflect the limited spatial resolution of the analysed data.

Another method that can be applied to structures lacking strict self-similar properties is the *box-counting method*. For an image depicting a two dimensional curve, this technique subdivides the image into a number of equal sized boxes; the number of boxes which contain portions of the curve is then counted and the process is iterated with different sizes of the boxes. The fractal dimension of the contour is related to the slope of the plot between the logarithm of the number of boxes through which the contour passes and the logarithm of the size of the boxes (Longley and Batty, 1989).

B.6 Dimensionality of the codon space

B.6.1 Dimensionality of the synonymous sets

Although the synonymous codon vectors have 59 components (excluding Methionine and Tryptophan, coded by a single codon, and the STOP codons), the codon space is not a 59-dimensional space, because the components are not independent values but are instead relative frequencies. Taking for example the amino acid Cysteine, with its two synonymous codons TGT and TGC, it is probably intuitively clear that all the possible values corresponding to these two dimensions can be represented as a one dimensional segment, since the sum of the frequencies of TGT and TGC must be equal to one: if a two-dimensional square is considered, with the possible values for TGT and TGC as its x and y dimensions, all the possible distributions of frequencies lie on one of the two diagonals of the square.

Similarly, for the case of the three-fold degenerate amino acids the distribution of frequencies can be represented as a triangle (two-dimensional, "diagonal" of a cube), while for four-fold ones it has the three-dimensional shape of a tetrahedron (from the four-dimensional hypercube which would represent the four components if these were independent).

More difficult to visualise is the case of the triplets for six-fold degenerate amino acids, but by comparison with the other cases it is a five-dimensional space. In other words, in the theoretical case of six uncorrelated triplets the points lie on a five dimensional surface inside the six-dimensional hypercube. Having one less degree of freedom, the space of CPRO dimensions relative to those triplets is instead four-dimensional.

A way to calculate the dimensionality of the triplet distributions is to use the box-counting method which is usually employed in fractal analysis: subdividing the six-dimensional space in boxes of increasingly smaller size and counting the number of boxes which contain data points (see previous section).

When this method is applied on the biological data for 6-fold synonymous sets or on randomly generated frequencies, the result approximates but does not reach the expected value of five dimensions. In reality, this kind of procedure is heavily limited by the sparsity of data in high dimensions (see also C.2) and its limits are already perceived in this six-dimensional application. With approximately one million of biological frequencies for each synonymous set of six triplets, the slope of the linear regime of the log-log curve estimates a fractal dimension of 4.4. This result is with all probability an underestimation, due to the extremely high number of samples required by the boxcounting method for high-dimensional spaces. It is thus not possible to apply these methods to the analysis of the full-length codon vectors of the codon space.

B.6.2 Dimensionality of the entire codon usage

The synonymous codon usage space has a maximum of 41 uncorrelated dimensions, which is the product of the orthogonal *subspaces* with the different dimensionalities described above for each type of synonymous set. The total dimensionality derives from the sum of nine one-dimensional subspaces, one two-dimensional, five three-dimensional and three five-dimensional ones. If the distributions for the synonymous triplets were completely independent, with the relative usage of codons for one amino acid not correlated to the relative usage of codons for another amino acid, then all the possible distributions of relative usages could theoretically be found among the biological sequences. In other words, if the usage of the codons were random, or if there were no constraints, global trends and correlations, then the populated space should resemble the theoretical space.

In reality, the synonymous sets for the different amino acids are often observed as being not independent: when analysing biological data a correlation is almost always found, for example, between the triplets contributing to G+C content. The effective heterogeneity among codon usages is hence expected to be lower than what could be theoretically possible. Correlations between triplets and other constraints would limit the possible "exploration of the codon space" by biological sequences, *i.e.* the maximum divergence between codon usages.

C METHODS

C.1 Mapping the populated space

A filtered space was the object of the investigation, restricting the analyses only to complete codon usages, either in the sense of comprising at least a triplet for each amino acid (AA-filtered space), or encoding the full repertoire of triplets (CPRO or CSYN filtered space). The decision to filter the data was taken in order to investigate a coherently defined set, consistently removing any aberrant codon usage, like those deriving from very short transcripts and susceptible to large stochastic variation (the filtering procedure was presented in section II C.4). The more restrictive filters limit the amount of codon usages that can be analysed: considering all the analysed data sets, there are 948,938 transcripts encoding all the amino acids (AA-filtered), while 133,231 is the number of CPRO-filtered ones and 93,032 the number of CSYN-filtered ones. Although there is less data available for them, the codon usages encoding the full repertoire of codons are better suited at investigating the correlations among all the triplets in the genetic code.

Firstly, the transcripts obtained from the completed genomes were analysed. An annotated completely sequenced genome ensures a high level of quality for the sequences. Secondly, the space mapping was extended to the whole EMBL database (Stoesser *et al.*, 2003), thus significantly enlarging the size of the data set. Release 75 (June 2003) of this database was used. The sequences chosen were those employing the Standard genetic code and the equivalent Bacterial code (almost all prokaryotic sequences employ EMBL translation table 11, the *bacterial code*, which is the same as the Standard code, with the only difference of having additional potential initiation codons).

For too many data points (one for each transcript sequence) multivariate analysis requires an exponential amount of allocated memory and computation time (for example many MVA algorithms rely on the computation of matrices of distances between all the data vectors). Additionally, comparisons and visualisations are difficult to perform when there are too many data points.

A simple binning algorithm was hence adopted in order to keep only a minimal set of representatives, thus mapping the populated space at a specified level of granularity. The algorithm proceeds sequentially through the data set and accepts new data points only if their Euclidean distance to all the previously accepted data points is greater than the defined cut-off (the granularity level).

The result is that a reduced number of representative points are kept to characterise the dataset, all at a certain minimal distance from each other (see Figure VI-1). Each point represents the class of points that shares similar features, in this case similar codon usages. Taken together, the representative points reflect the topology of the region occupied by all the data points. In other words, they represent the portions of the codon space corresponding to the codon usages of the biological data.

Different distances can be used to sample the space, with shorter distances leading to more representatives but a tighter fit of the high-dimensional region. Granularity radii which are too short produce too many representatives. Conversely, granularity radii which are too long incur the risk of a loss of features in the representative space (Figure VI-1 *e f g*).

Apart from making the codon space manageable, the binning procedure has the benefit of producing a more uniform data set, where redundancies are eliminated and the sequencing sampling biases greatly reduced. For high enough granularity radii, the resulting space is even and unbiased, with all the representative points equally identifying the populated regions of the space (*i.e.* all points are treated equally and together they represent the total coverage of the populated space).

Alternatively, when it is desirable to maintain the density information (how many data points are in a given region) the algorithm can be set to compute the number of codon usages covered by each representative and that information can hence be analysed (as in Figure VI-4 and Figure VI-5 of section D.2). The density information reveals: 1) which are the main trends for specific subsets inside the codon space (for example, if there are preferential regions of the space where plant codon usages can be found); 2) the significance of the representative points (which points stand for many codon usages and which are instead marking isolated or rare occurrences); 3) which are the most populated areas of the codon space and what are the most commonly employed codon usages (in absolute terms or for selected taxonomical groups).

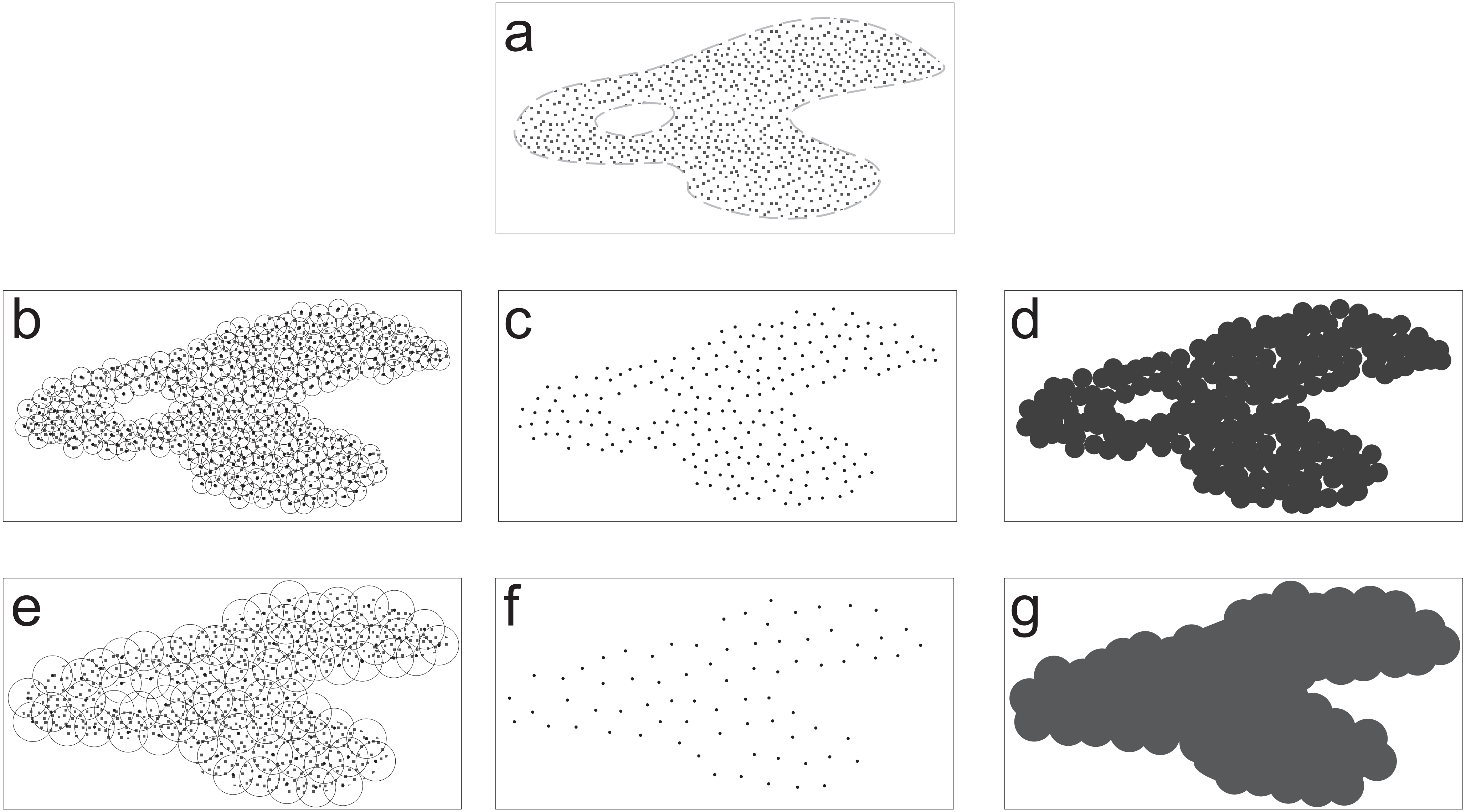


Figure VI-1: *Schematic representation of the binning algorithm for the case of a 2 dimensional space. (a) The space is populated in a region with a certain shape (approximated by the dashed line). The algorithm reduces the number of points while trying to preserve the topology. (b) As new representative points are added, a disc of the specified granularity radius prevents new representative points from being selected in that area. (c) The result is a reduced number of points that describe the populated region. (d) The topology obtained from the representative points at this granularity level. (e) A longer granularity radius can be used, resulting in (f) less points but (g) a coarser description of the populated space, with the risk of losing some features, like the empty inside region which will appear populated when sampling is performed on this space of representatives.*

C.1.1 Number of identified representatives and choice of granularity radius

The logarithm of the number of representatives is linearly proportional to the logarithm of the granularity radius (as shown in the following Figure VI-2). The linearity holds as long as the number of identified representatives does not become too close to the total number of points. Plotting this information can help decide on the choice of the granularity level at which analysing the codon space. Too short a radius would yield too many points representing only themselves (inefficient binning). Too large a radius loses any structure in the codon space (to the absurd limit case of a single representative for all the codon usages). Furthermore, it is possible to predict the number of representatives, limited to the linear regime, which can be found at a given radius, after a few iterations have been computed with different radii.

It is important to note that the number of representatives identified by the developed binning algorithm needs to be considered an approximate rather than an exact number. In fact the number of representatives can fluctuate, depending on the order in which the vectors appear in the data set when the algorithm is run.

It was experimentally determined that a near-optimal coverage can be obtained by randomising the order of the total data set analysed: if the order of the vectors of the original data set is randomised, the number of representatives found is about 16% lower than the number of representatives which are obtained from data sets to which a sorting procedure is applied (for example sorting by the values assumed by the vectors along a specified dimension). The representatives of sorted data sets would have a tighter packing (with more overlap between their hyperspheres).

Iterating the binning procedure over data sets in different randomised orders allows the estimation of the amount of fluctuation, measured as the deviation in the number of identified representatives, at a given granularity radius. The fluctuation is dependent on the size of the data set and on the granularity radius. It is low for short radii, increases with the radius and eventually decreases again when the granularity radius becomes too large and the number of found representatives becomes very low. For the size of the analysed data sets the standard deviation in the number of representatives is lower than 1% (percentage of the average representative number) for granularity radii below 1; it reaches 2% around 1.25, 4% for 1.4 and after reaching its maximum at 1.5 (6%) it decreases again, indicating loss of features in the codon space.

From these two approaches, calculating the number of representatives and the deviation for different granularity radii, it is possible to choose the level at which to analyse the codon space, *i.e.* to determine the range of granularity radii that are neither too short nor too long.

The numbers that will be presented in **Results** for the total representatives are the averages among 50 iterated binning. A possible alternative would be to use the minimum number of selected representatives instead.

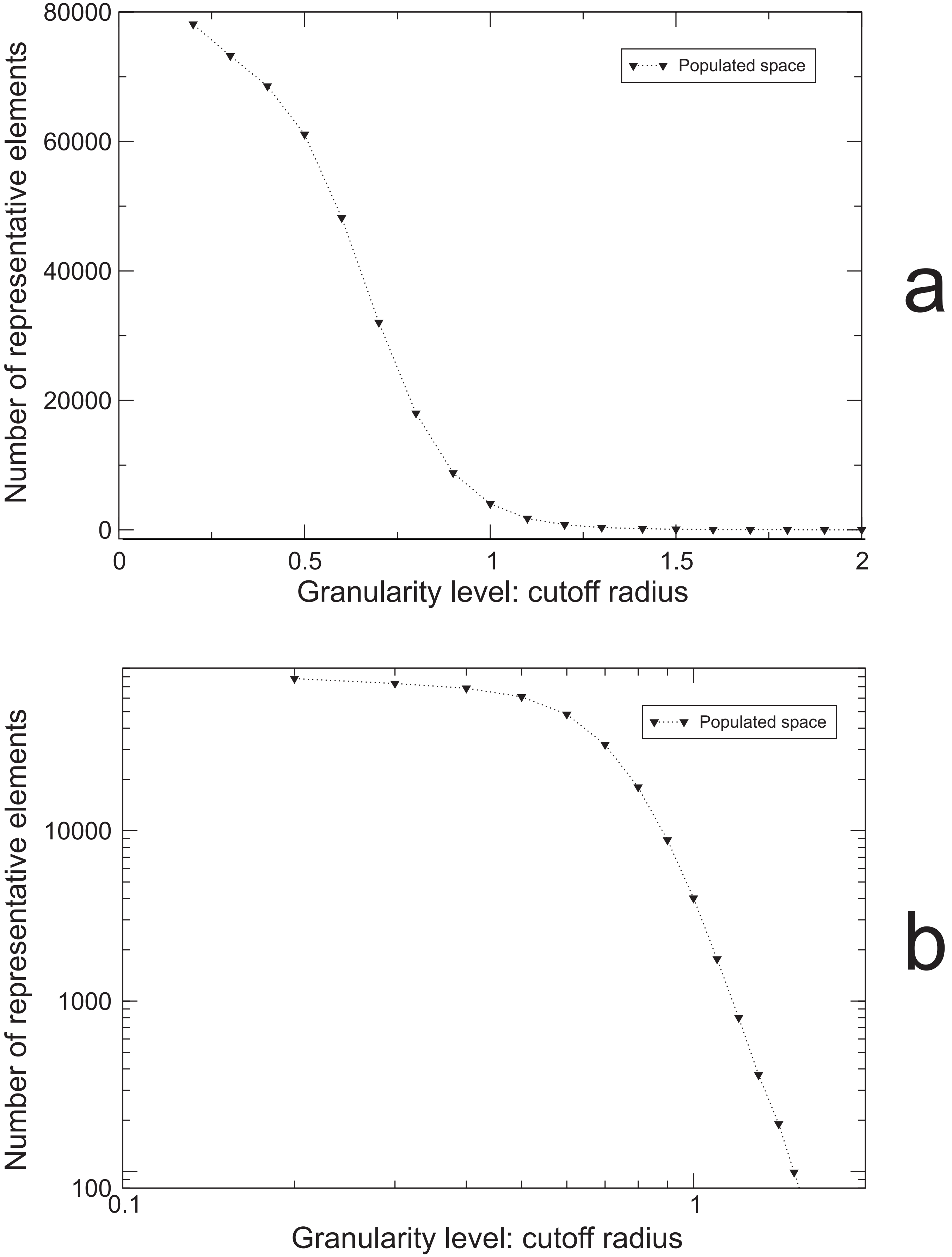


Figure VI-2: *Number of representative points kept for the populated space at each granularity level (for each choice of cut-off radius). (a) linear scale (b) logarithmic scale for both axes. The logarithm of the number of representative points is linearly proportional to the logarithm of the granularity radius. This figure additionally reveals the high number of similar codon usages: from a total of 133,232 non-identical codon vectors, 55,117 can be found lying at a distance shorter than 0.2 from another codon usage.*

C.1.2 Shortcomings and future improvements

The binning algorithm described above is not efficient for low granularity radii and large amounts of data. In fact, since every codon usage needs to be sequentially compared for proximity to all the previously accepted representatives, the algorithm slows down considerably as the number of representatives grows. This needs to be addressed with more efficient binning, otherwise the computational requirements for a high amount of representatives could be prohibitive. In fact the analysis of the AA-filtered space (roughly one million codon usages) becomes impractical at low granularity radii: in the worst case, the creation of bins containing single elements, the algorithm goes through $n^2/2$ distance evaluations.

The results presented in this chapter are hence those relative to the CSYN-filtered and CPRO-filtered data sets, which were compared against a CSYN-filtered theoretical space (*i.e.* all triplets are required to be present, although very low frequencies, approximating the zero-frequency case, are permitted). Apart from the lower computational requirements, these more restrictive spaces can be better compared with the theoretical space, because they can reveal correlations among the complete codon usage, *i.e.* among all the triplets.

The topology for the AA-filtered space at high granularity levels was found to be largely comparable to the ones investigated for the more restrictive codon spaces, so most of the results presented can be applied to it. Nevertheless, a thorough investigation is needed to appropriately compare these data sets.

Another possible improvement to the binning algorithm could address the fluctuations in the number of representatives (non-optimal coverage; previous section). One possible solution would be to pre-scan the data set with longer radii and use them to guide the binning at shorter-radii, and using the centroids of the bins (the points whose coordinates are the average of all the members of the bins) as representative points. This approach would increase the computational requirements (as the data set would need to be scanned more than once) but would produce better coverage and a more exact value for the number of representatives needed at the specified granularity level. Apart from the higher computational complexity, querying the model would become less direct, since the representatives would not be existing codon usages but codon usages averaged over existing ones. With the algorithm presented in this work, it is instead straightforward to obtain the accession number of the sequence-database entries whose codon usage is kept as representative, and retrieve their annotation and their sequence.

C.2 Random sampling and identification of the non-populated regions

In order to characterise the biological space, it is useful to compare it with the theoretical space, which could be approximated by randomly generated codon usages.

There are several possible studies that can be accomplished with a source of random codon vectors. For example, they can be used to sample the codon space in search for regions which are not populated (not visited by the biological sequences analysed). In this case the procedure consists of randomly generating codon usage vectors and sampling the mapped codon space (the representatives obtained after binning at the desired granularity level). Alternatively, the procedure could be used with an empty space to estimate the number of theoretical codon usages at a given granularity level. Each new random vector gets compared to all the stored data points, and if its distance from all points is found to be higher than a specified cut-off, the sample is accepted and added to the codon space. The cut-off can be the same as the one used for the mapping or a coarser one can be adopted, in order to reach saturation (indicated by a steep decrease in the number of new non-populated representatives being found) with a lower number of accepted random points.

In fact, these kinds of direct-sampling methods suffer from the sparsity of data in high dimensions (an effect usually called the *curse of dimensionality*, due to the exponential growth of hyper-volume as a function of dimensionality; Bellman, 1961). In order to maintain a given level of accuracy, the number of required samples increases exponentially with the number of variables. In practice this means that sampling at low granularity levels is both unfeasible and inaccurate, but it is possible to sample at coarser levels.

Additionally, both the number of representatives at a given radius and the change in the number of representatives found for different radii can be compared between the biological codon usages and the randomly generated ones. This allows the estimation of the total coverage and density of the biological space with respect to the theoretical space.

Four different algorithms have been subsequently developed to generate random codon vectors adopting different criteria. For each of these algorithms it is possible to specify the application of a filter (see II C.4), in order to generate only vectors which would satisfy that filter.

C.2.1 'Random triplets' codon usage generation algorithm

The first algorithm is based on randomly deciding how many triplets to assign to each amino acid and then randomly selecting between the triplets coding for that amino acid, until the decided amount has been assigned. The total size of the transcript is also randomly chosen, within a specified range.

This algorithm produces codon vectors which are not very different from the codon usage average vector: for a high number of triplets, the randomness will approximate a uniform distribution (*e.g.* 25% for each of four synonymous codons), while for low numbers the stochastic variation would be greater.

C.2.2 'Random frequencies' codon usage generation algorithm

The second algorithm proceeds by determining *a priori* position-specific relative base frequencies (TCAG123, relative use of the nucleotides at the three codon positions) and then generating random triplets whose total nucleotide content would be equivalent to those pre-determined frequencies. In this way the algorithm manages to generate more diverse codon usages. Additionally, the generated codon vectors for the 20 amino acids are not probabilistically independent as they reflect the total base frequencies. In this way they approximate the correlations observed in the codon usage of many genomes: for example the correlation between total G+C content and codon usage.

C.2.3 'Random usages' codon usage generation algorithm

The third algorithm removes this inter-dependence by determining *a priori* frequencies for the synonymous triplets for each amino acid species. For example it will randomly assign the four codons for Alanine to be in a proportion of 10:30:5:55. A minimum frequency can be set for each codon species (for example setting a minimum of 20% for triplets of 2-fold degenerates, while allowing a minimum of 10% for triplets of 4-fold degenerates). Additionally, it is possible to specify the granularity level for the resulting frequencies, for example producing only frequencies in multiples of 0.1.

This algorithm can produce very deviant codon usage vectors and hence explore zones of the codon space which are prevented to the previous algorithm due to its inter-dependency in nucleotide contents: the 'random frequencies' algorithm would rarely produce a codon usage which is GC-rich for the triplets coding for half of the amino acids and GC-poor for the other half. Also for this algorithm it is possible, specifying a total and minimum frequency for the triplets, to regulate the granularity in the generated random usages.

C.2.4 'Random distributions' codon usage generation algorithm

The possible arrangements of triplet frequencies of the random codon vectors generated by the 'random usages' algorithm are uniformly distributed. The 'random distributions' algorithm can be set to produce coarse frequencies which obey certain specified distribution schemes. For example this algorithm can be instructed to generate vectors missing the most extreme triplet distributions (like 0.9:0.1 for 2-fold degenerate amino acids or 0.7:0.1:0.1:0.1 for 4-fold degenerate ones). Conversely, it can generate vectors whose frequency distributions are only the most extreme ones (with one triplet greatly over represented over the synonymous ones for each amino acid type). There is no inter-dependence between the frequencies among different amino acids.

This is by far the most versatile and also the fastest of the presented algorithms, as it exploits a pre-generated library of all possible frequency distributions for the desired scheme (schemes like "only extremes", "no extremes", "coarser") from which to randomly select the distribution of triplets for each amino acid species in the vector.

C.2.5 General considerations on the developed random algorithms

Theoretically, the maximum Euclidean distance from the average codon vector is 3.629 and there are 340 million possible AA-filtered possible vectors with this distance from the average (the minimal codes, where only one codon is used per amino acid). For AA-filtered transcripts, all algorithms manage to generate these extremely deviant transcripts. CPRO/CSYN-filtered transcripts could theoretically get very close to that limit (with individual relative frequencies as low as 0.002, for example; enough to satisfy the filter) but in practice these extremes are never reached in random searches with the first two algorithms.

For the first algorithm ('random triplets'), a distance from the average codon vector of no more than 1.771 units could be achieved even after extremely long searches (of several millions of generated vectors), while the other ones can easily generate more deviant vectors.

Besides the maximally deviant codon usage that an algorithm can generate, another important property is the heterogeneity of the generated vectors, how diverse and independent from each other they are. This property is fundamental for a comprehensive exploration of all regions of the theoretical codon space.

The importance of a good source of randomness, to generate very diverse vectors in the exploration of the codon space, can be easily underestimated. The codon bias of the bacterium *Streptomyces coelicolor* is 2.627 distance units from the average vector (this genome – which represents one of the most deviant data points in the CPRO-filtered mapped space – has a total G+C content of 72.5%, with a GC3 averaging 93%), so vectors as diverse as this one need to be generated for sensible sampling.

Conversely, the algorithms can be set to avoid generating too extreme data points by measuring the distance to the average codon vector and discarding those vectors whose distance is larger than a specified threshold; for example discarding vectors whose distances from the average are greater than 2.3 units.

C.2.6 Characteristics of the generated codon usages

Figure VI-11 (presented in section D.4 of **Results**) compares the biological codon usages and the random codon usages generated by the different algorithms, binning them with a range of granularity radii.

The codon usages generated by 'random distributions' are the most diverse (they do not have correlations between non-synonymous triplets) and hence their number of representatives is always higher than that found for the biological codon usages or for the other randomly generated ones.

A similar reasoning can be applied to those generated by 'random usages', with the exception of the behaviour observed at high granularity radii. Since this algorithm was set to generate frequency distributions with a minimum of 0.2 (20%) for the 2-fold degenerate amino acids, the generated codon usages are not as deviant from the average as some of those which can be found in the biological space: there are biological codon usages in which some triplets are used almost exclusively over their synonymous alternatives. For this reason the number of representatives at very high granularity radii (which keep only very diverse codon usages) for the 'random usages' vectors is lower than the number of biological representatives.

The 'random usages' at these higher radii, beyond 1.5, are also less diverse than those generated by 'random frequencies'. In fact, these two algorithms are complementary: 'random frequencies' has correlation between the non-synonymous frequencies but it can generate very deviant vectors from a nucleotide compositional point of view (like a very extreme distribution for all GC-rich codons) while 'random usages' generates more diverse codon usages but (since it was limited to 0.2 minimum frequency for the synonymous couples) these do not reach the extreme perimeter of the space.

D RESULTS AND DISCUSSION

The scope of this study was to understand the topology and extension of the populated codon space. In other words, to compare the theoretical codon usages with those that can be found in the biological sequences, to examine the heterogeneity of codon usages employed by diverse organisms or groups of organisms, to quantify how diverse the biological codon usages can be or, conversely, how constrained.

Obviously the biggest impediments to the creation of a full picture of the biological codon space are the inevitable sampling bias and the low sampling size. The former relates to both the technological limits (only recently it has been possible to determine the sequence of large genomes) and to the differential relevance that guides the choice of species to analyse (model species and pathogens are the first to be studied). As for the sampling size, the sequences presently stored in the databases are a mere fraction of the global biological patrimony (even limiting ourselves to the existing species, not considering those which became extinct). Estimates of global species diversity vary greatly, ranging from as low as two to as high as one-hundred million species (Ozanne *et al.*, 2003). Even if the number of genes and genomes being sequenced grows dramatically day by day, when we consider that there are only 17,140 species with at least one coding sequence in the database (as to the February 2003 GenBank release) it is clear that the best possible picture of the biological codon space can only be a very blurred and minuscule one.

Nevertheless it is interesting to draw the picture, however tiny and inaccurate, with the awareness of its limits but also of the fact that the existing data can still be a representative of the total data.

D.1 Mapping the populated codon space

Following the aim of mapping the complete space, the analysed data sets contain codon usages employed by very diverse species from the main taxonomical subdivisions of life. Nevertheless it is also possible to chart the populated space of specific subsets, and two of them, namely the vertebrate and the prokaryotic spaces, will be briefly presented below.

D.1.1 Completed genomes

At a granularity level of 1.3 Euclidean distance units, 460 representatives are selected for the completed genomes; 49 of which come from archaea, 220 from bacteria, 26 from *Anopheles gambiae*, 15 from *Takifugu rubripes*, 14 from *Homo sapiens*, 4 from *Mus musculus*, 14 from the yeasts (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*) and 118 from the viruses.

At that granularity level, 14 data points are identified by the binning algorithm as necessary to represent the extension of the codon space relative to *Homo sapiens*: the first covers the great majority of human genes, whose codon usage closely reflects the average genomic bias; the other points represent the codon usages of atypical transcripts. Together, they capture the variability that the human genome transcripts can reach inside the codon space. Very few human transcripts have a codon usage significantly different from the average codon vector (the vector with averagely distributed synonymous frequencies). Conversely, the *Pseudomonas aeruginosa* genome bias is more extreme than any of them, with an Euclidean distance from the average vector of 2.34 units. This great distance is due to the very high GC3 content of 88% in this genome and to the almost exclusive usage of some synonymous triplets, for example the TTC triplet for Phenylalanine used at 95% (Grocock and Sharp, 2002).

If representatives are selected among the total data (all the sequences from the completed genome analysed together rather than split in different groups), the mapping is reduced from 460 to 201 data points which lie no less than 1.3 Euclidean distance units apart, a measure of the extent of the redundancy and overlap that exists between the codon space extensions reached by these groups.

This data set was considered too limited to represent the complete biological space so the procedure was extended to a broader data set.

D.1.2 All sequenced transcripts

The whole EMBL database (see Table VI-2 for a list of the division files examined and release information) was then analysed and representatives for the codon space were selected at different levels of granularity.

If the representatives from EMBL are compared to the codon space binned in the previous analysis (when only sequences from completed genomes where analysed), approximately double the number of codon usage representatives are found: data points representing codon usages not covered by the previously mapped codon space. At the same granularity radius of 1.3 there are in fact 201 representative points describing the space of the completed genomes, while there are 397 for the EMBL data set.

Those are the values when the two sets are independently binned. If instead the completed genomes codon space is used as the *starting space* for binning the EMBL codon usages (to prevent overlap: only those from previously empty regions are added), the novel EMBL codon usages are 179 (an increase of 89%).

The majority of the new data points found (in the EMBL data set compared to the completed genomes data set) are codon usages from plants and invertebrates (mainly from the genomic projects of model species: like *A.thaliana*, *O.sativa*, *C.elegans*, *D.melanogaster*), as was expected due to the bias towards prokaryotes in the data from completely sequenced genomes.

To integrate the two sources of data, a new comprehensive data set was constructed by joining the codon usages computed from the EMBL database sequences and those relative to the completely sequenced genomes. These codon usages were arranged in nine groups, as detailed in Table VI-2. The obvious redundancy between the two sets was eliminated in the process of selection of representative elements.

groups	Data origin	average number of representatives				
		1.1	**1.2**	**1.3**	**1.4**	**1.5**
prokaryotes	archaea†, bacteria†, PRO	880	426	213	104	63
fungi	FUN, yeastsΔ	193	101	56	29	19
plants	PLN	398	189	90	44	25
invertebrates	INV, *A.gambiae*‡, *D.melanogaster*◊	487	239	121	64	37
vertebrates	VRT, *T.rubripes*‡	118	63	31	16	11
mammals	MAM, ROD, *H.sapiens*‡, *M.musculus*‡	112	58	33	18	12
viruses	VRL	373	192	101	52	32
phages	PHG (bacteriophages)	67	40	24	14	11
other	ORG (organelles), UNC (unclassified)	72	42	27	16	11
	total	2700	1351	695	358	220
	without overlap	1736	792	376	179	106
	overlapping points	36%	41%	46%	50%	52%

Table VI-2*: The groups in which the data for the populated space was sorted and the representative elements for each group at different granularity radii. Uppercase codes in the second column refer to EMBL divisions (Release 75 June 2003). †: Completely sequenced archaea and prokaryotes as from appendix III F.1; Δ: genomes of S.pombe, S.cerevisiae; ‡: from Ensembl; ◊: from FlyBase. The values in the "without overlap" row refer to the representatives computed from the total data set, without creating overlapping representatives between the nine groups. These are hence the real estimation of the size of the codon space at the set granularity level. Note that all the numbers presented are the rounded up average values on fifty randomised-order runs of the binning algorithm: see Methods C.1.1*

As mentioned above, considerable overlap exists between the groups: those representative points of different groups lying at a distance shorter than the granularity radius (having approximately the same codon usage). The amount of overlap increases with the length of the granularity radius, since the total number of all the possible codon usages decreases for higher granularity radii (B.3).

Prokaryotes (and in particular the bacteria) constitute the most diverse group and this is reflected both in the number of representatives as well as in the extension of the codon space that they cover, with the maximum distance between two points of this group being 4.227 Euclidean distance units (basically spanning two opposite "corners" of the codon space, completely opposite codon usages). Independently from the granularity level at which the codon space is mapped, the representative data points for the prokaryotes are the most numerous. But apart from the diversity between the prokaryotic species, this could also reflect the sampling bias in sequencing: up to now only few eukaryotes have been sequenced completely (or even partially), while prokaryotes account for half of the sequences in the analysed EMBL divisions.

A large part of such sampling bias is removed by the binning procedure: selecting representatives using relatively high granularity radii produces a more uniform data set, lowering redundancy and reducing the bias towards the most populated groups. Without binning, or using short granularity radii, the similar codon usages between certain highly populated groups (those for which more sequences are available) influence analyses of the codon space: for example when plotting the distribution of the values for each dimension in the codon space, the distributions for the total space are skewed towards the values found for the prokaryotic space, because of the higher amount of data. Increasing the granularity radius, this effect diminishes considerably and the distributions become more flattened and less skewed, although remaining almost in the same ranges (*i.e.* covering the same extension of codon space), as shown in Figure VI-3.

It is then important to use relatively long granularity radii in order to reduce the sampling bias (so that any over or under-representation observed can be attributed to a universal feature rather than to the bias given by a single group). Very long radii are also to be avoided, because the number of representatives becomes too low and too many features could be lost at the coarser levels (as explained in **Methods**, section C.1).

Although not all sequencing sampling bias can be removed in this way (since the number of species sequenced for the different groups varies greatly), the prokaryotes do appear the most biologically diverse group in codon usage, both in the number of representatives and in the extension over the codon space. Invertebrates, plants and then viruses follow them in terms of codon diversity. The vertebrates are a quite compact group, with a maximum distance between two vertebrate representative codon usages amounting to 2.549 units (thus not a large span over the codon space, if compared to the prokaryotes, spanning 4.227 at their maximum).

It is now too premature to extrapolate from the existing sequenced data, but it will be interesting to follow the number of representative codon usages identified as more sequences get determined. At some point a plateau effect should become visible (like it happened for the number of protein folds being discovered), with less and less new codon usages being found (at a certain specified granularity radius). It is probable that the number of representative points describing the total biological codon space will raise greatly, but without an extreme change in the order of magnitude (an expected estimate of the maximal upper bound being in the order of thousands, at 1.41 granularity; possibly under one thousand). This guess is based on the observation of the distributions of genomic nucleotide frequencies, compactness of the populated space and comparison with the randomly generated codon usage vectors (presented below).

D.2 Analyses of the populated space

The space of codon usage representatives can be analysed in several ways: looking for universal biases between synonymous triplets, computing the ranges of total nucleotide contents, investigating particular regions or particular subsets of it and measuring its extension and density. The main advantage of this model is the reduction of the number of points that can be kept while maintaining the topology of the total space and removing redundancy. This makes very large scale analyses possible and allows the easy extension of the model with new data. Every newly determined sequence can be tested against the representative space to see if it represents a novel codon usage. Similarly the representative space could be used for studies on biodiversity, species evolution or sequence identification. One possible application is that the model can instantly provide the codon usages most similar to that of a query sequence, with the desired amount of similarity.

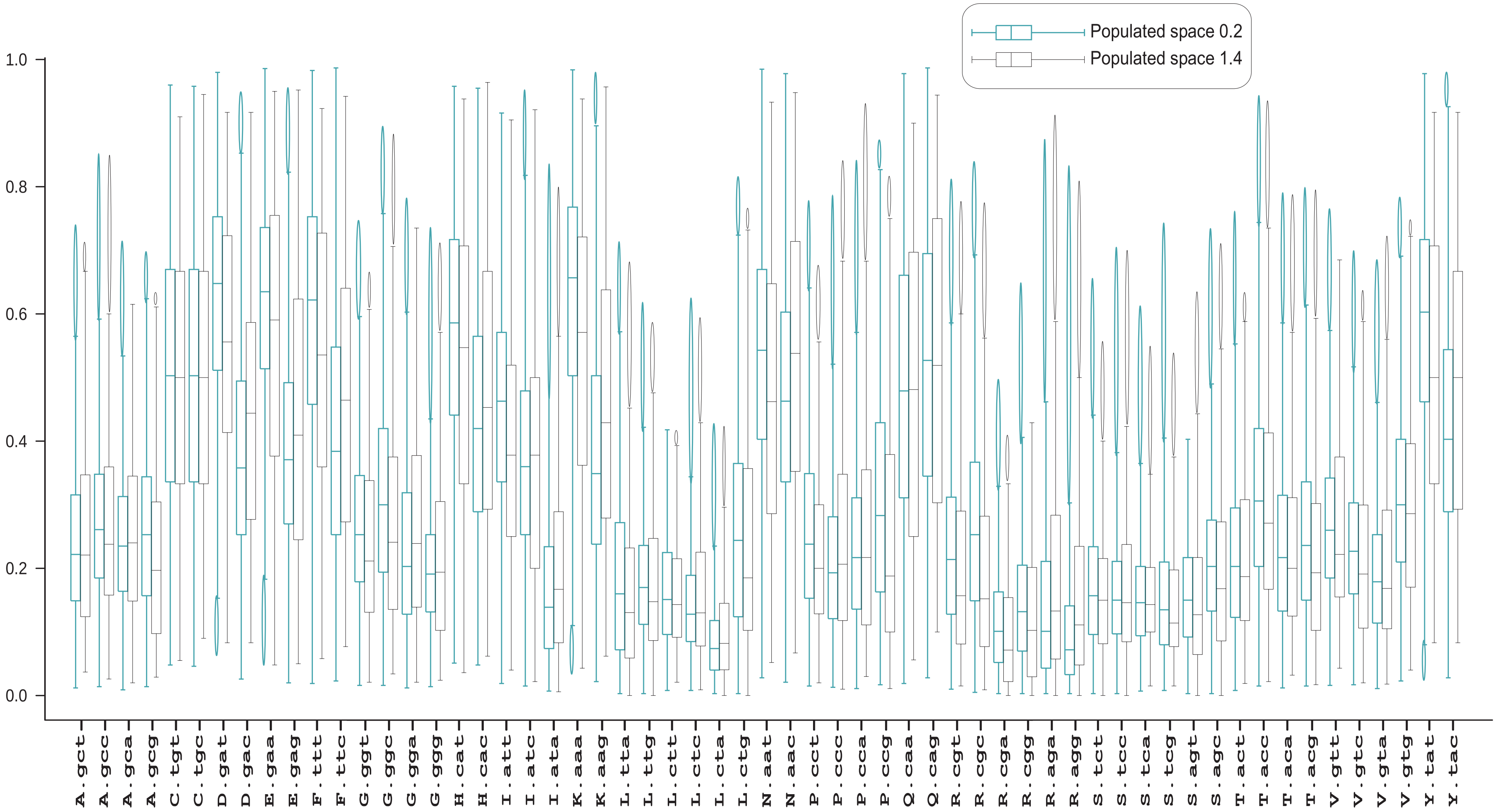


***Figure VI-3**: The distributions, in boxplot representation, of the values for the dimensions in the synonymous codon usage space, with comparison between a low (radius of 0.2) and a high granularity level (1.4). For clarity reasons, the outliers are represented as a single ellipse, which covers the range in which they appear, rather than as individual points. Several of the greatest differences in synonymous usage which appear at low granularity level are due to a larger amount of data from certain groups (in particular from prokaryotes). They are reduced in the binning with longer radius. The remaining biases are hence a better representation of possible universal constraints and less the effects of the sequencing sampling bias. Please refer to the text and to the next section (D.2) for a better presentation of the populated space and a discussion of the commonly preferential usages.*

D.2.1 Low-dimensional representations and principal coordinates of separation

Two multivariate analysis procedures (III C.4), correspondence analysis (CA) and multidimensional scaling (MDS), were used for reducing the dimensionality of the space and understanding the major trends in codon usage. Adopting multiple techniques allows the comparison between the separation axes identified by each, to see if they diverge in the determination of the major trends.

A multidimensional scaling plot of the populated space appears in Figure VI-4. Density information (the number of effective codon usages being represented by the points) is included in the form of gray discs. If this information is ignored, the plot represents the populated space in an effectively unbiased way, showing the spread of the groups over the regions of the codon space, even if occupied only by one codon usage; if conversely this information is considered, it reveals the most populated regions of the space and where the majority of the codon usages for the different groups are to be found.

At the edges of the map there are extreme (very biased) codon usages while the centre contains more average ones. For example, on the rightmost side there are two very GC-rich representatives: AK094712, a human transcript coding for a probable zinc finger protein (87% GC3 content) and ANG14849, a mosquito transcript coding for the sizeable 4095 amino acid long glycoprotein gp330/megalin (Saito *et al.*, 1994) with its 93% GC3 content. On the left of the map lie the representatives of the most GC-poor codon usages, like that of the bacterium *B.aphidicola* with its 13% GC3 content.

The codon space can also be represented using three dimensions, obtaining a cloud of points which could be rotated and observed from different angles. Figure VI-5 presents three of these plots at different granularity radii and with density information. Several points that appear superimposed in the two-dimensional plot (Figure VI-4) can be discriminated along the third principal coordinate.

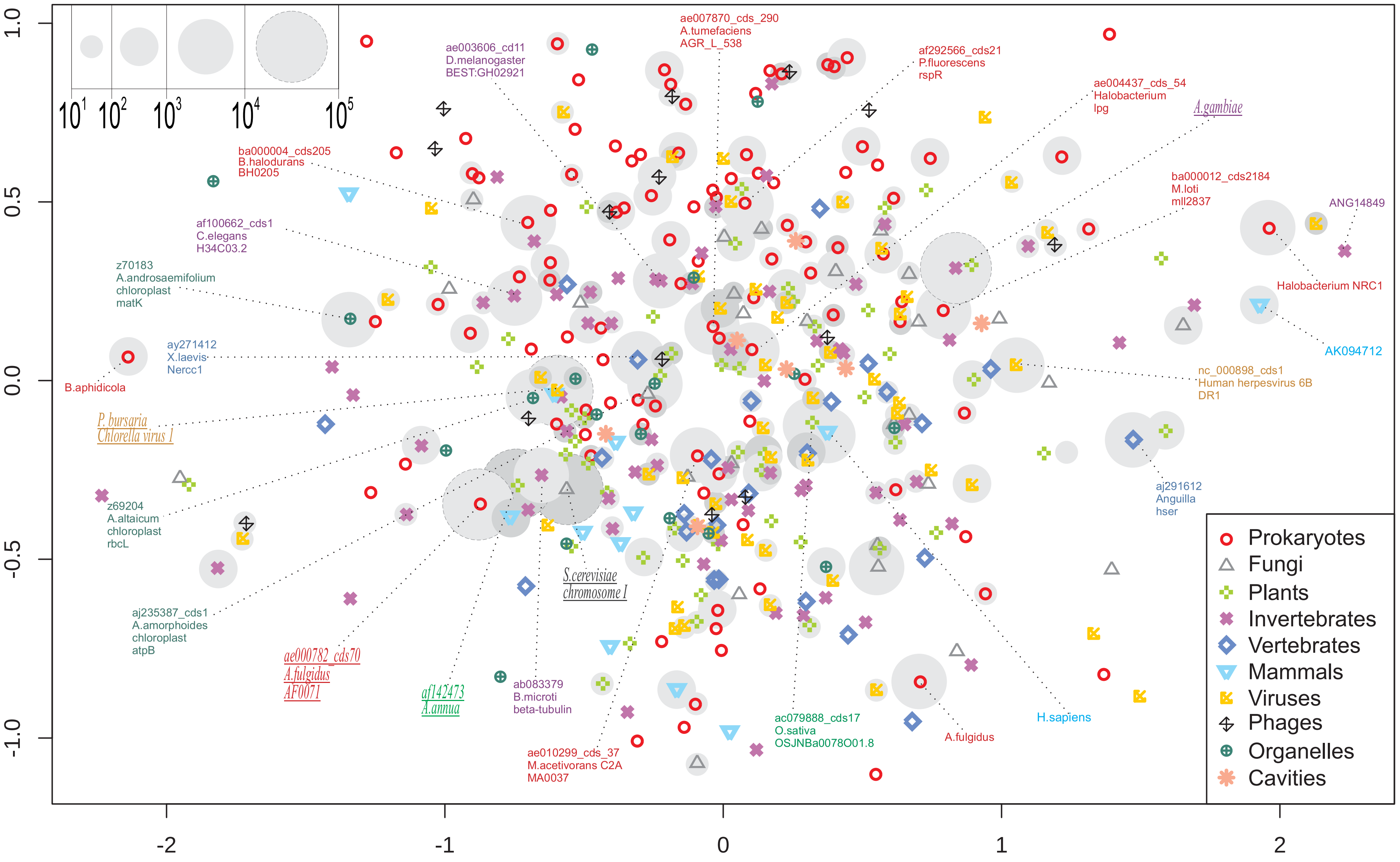
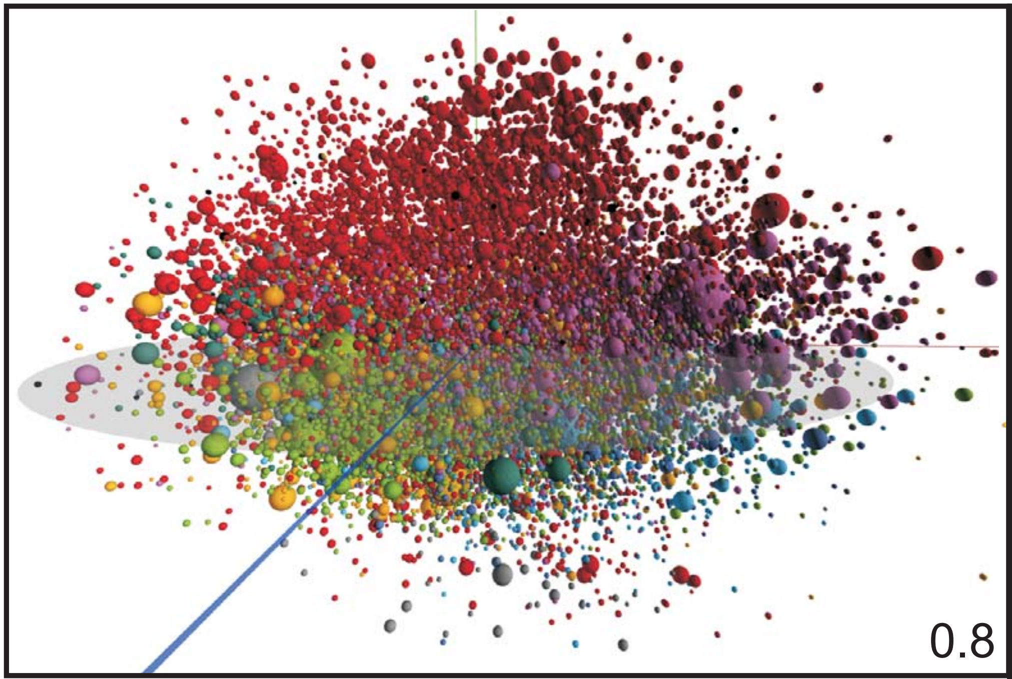
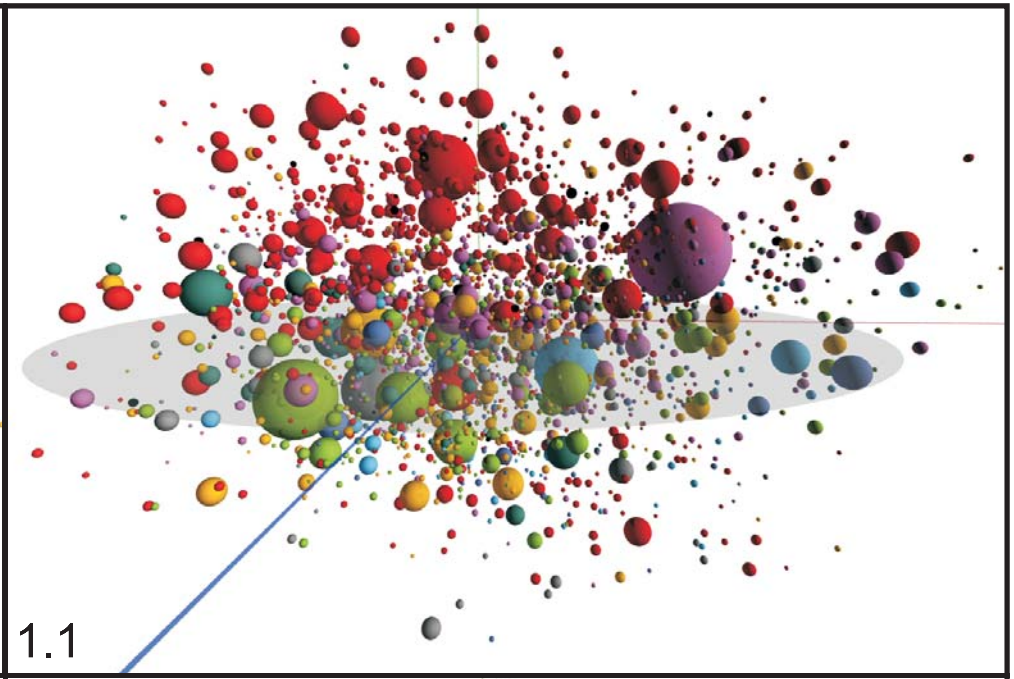


***Figure VI-4**: Codon space multidimensional scaling plot of the available populated space (main EMBL divisions plus completed genomes) at 1.4 granularity radius. Representatives are chosen for each group and hence overlap exists between representative points. The groups are outlined in Table VI-2. Density information (number of codon usages covered by the representative points) is indicated with gray discs of different areas behind the symbols for the representative points; if a representative covers less than ten codon usages, the disc is not drawn. The representatives with highest density are labelled with their database accession number, the scientific name of the species they belong to and the gene name where applicable. This map also shows the location of the cavities found (see section D.2.3).*





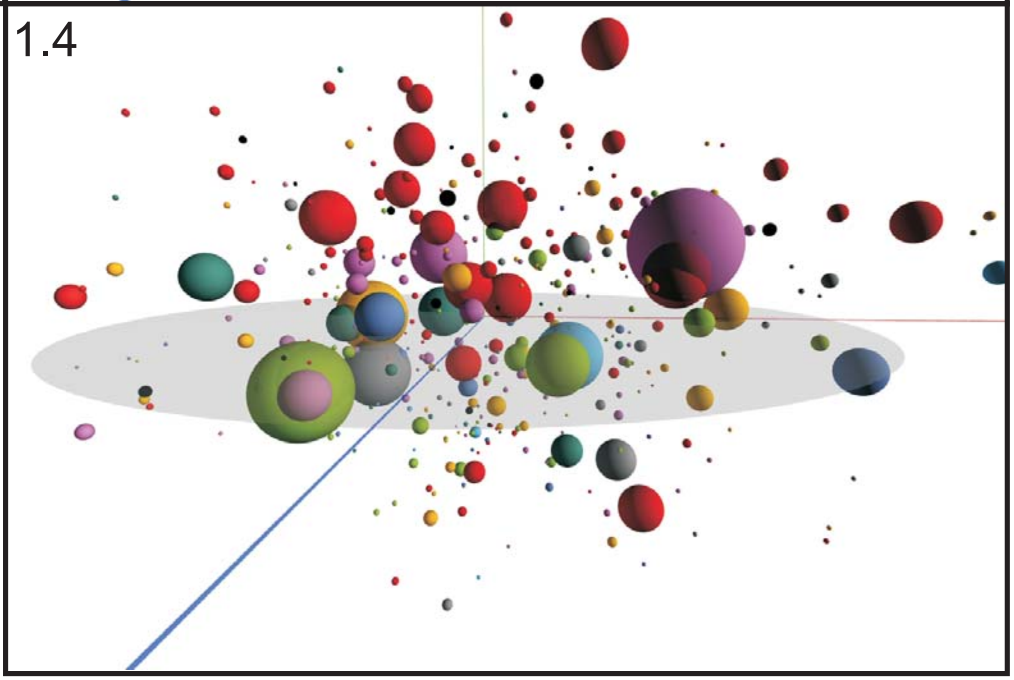


Figure VI-5: *Codon space correspondence analysis plots in three dimensions at different granularity levels. The representative points are here visualised as spheres, with their volume proportional to the density of the codon space in that region (the number of codon usages covered by the representative point). The first three principal coordinates are represented by the red, green and blue axes, respectively. The colouring for the nine groups are the same as in the previous figure (Figure VI-4). A horizontal disc indicates the position of the x-z (red-blue) plane. Although the prokaryotes (in red) have representatives all over the populated space, the majority of the prokaryotic codon usages is positioned in the upper part of the plot above the disc (see also Figure VI-9 below), together with the majority of those from bacteriophages (in black). Plants (in green) and vertebrates (in blue) are mostly limited to the lower part while invertebrates (violet) are prevalently distributed on the upper part, almost parallel to the x-z plane and not too distant from it (with the exception of a sizeable subset of invertebrate codon usages in the lower left quadrant, at negative x, positive y and z coordinates). The contributions to these axes are shown in the next Figure VI-6.*

Figure VI-6 reports the contributions for the orthogonal axes identified by multivariate ordination procedures as accounting for the largest fraction of variation among the codon space. The contributions are shown as codon difference matrices. These were computed for each orthogonal axis as the difference between the centroid of the 10% points with higher positive coordinate on that axis and the centroid of the 10% points with higher negative coordinate (a centroid is the point whose coordinates are the averages of the coordinates of all the points belonging to a cluster).

The principal axis, accounting for the largest variation among the codon usages (22.5% of the total), is the one related to G+C content (in particular to GC3), with the presence of extremely GC-rich and GC-poor codon usages.

The main contribution to the second coordinate is the usage of AGR codons (Arg-A1 in codon profile labelling), which effectively separates the majority of bacterial codon usages (and those coming from bacteriophages) from the other taxonomical groups.

The third axis is mainly related to the usages for the aminoacids Lysine, Glutamate and Glutamine, the three synonymous couples with A/G alternative in third position (*i.e.* the NAA/NAG triplets with the exclusion of TAA and TAG which are terminators), with high A3 content in the usage for these triplets at positive coordinates of the z/blue axis of Figure VI-5 and conversely higher G3 content at negative coordinates. Additionally there is a common TG3 correlation (T+G over A+C in third coding position).

The fourth axis separates according to T+C content (over A+G content) in all the 4-fold and 6-fold degenerate amino acids, while the fifth axis is mainly accounted by the NCG codons (Alanine, Proline, Serine and Threonine) over their NCC synonyms and by Phenylalanine triplets. The separation along the sixth axis is due to relative usage for Cysteine triplets. The major contributors to the seventh axis are the synonymous couples for Histidine, Phenylalanine and Cysteine.

The axes identified at a granularity level of 1.1 are almost identical, although there is a stronger contribution from Phenylalanine codons in the fifth axis and from Asparagine codons in the sixth axis.

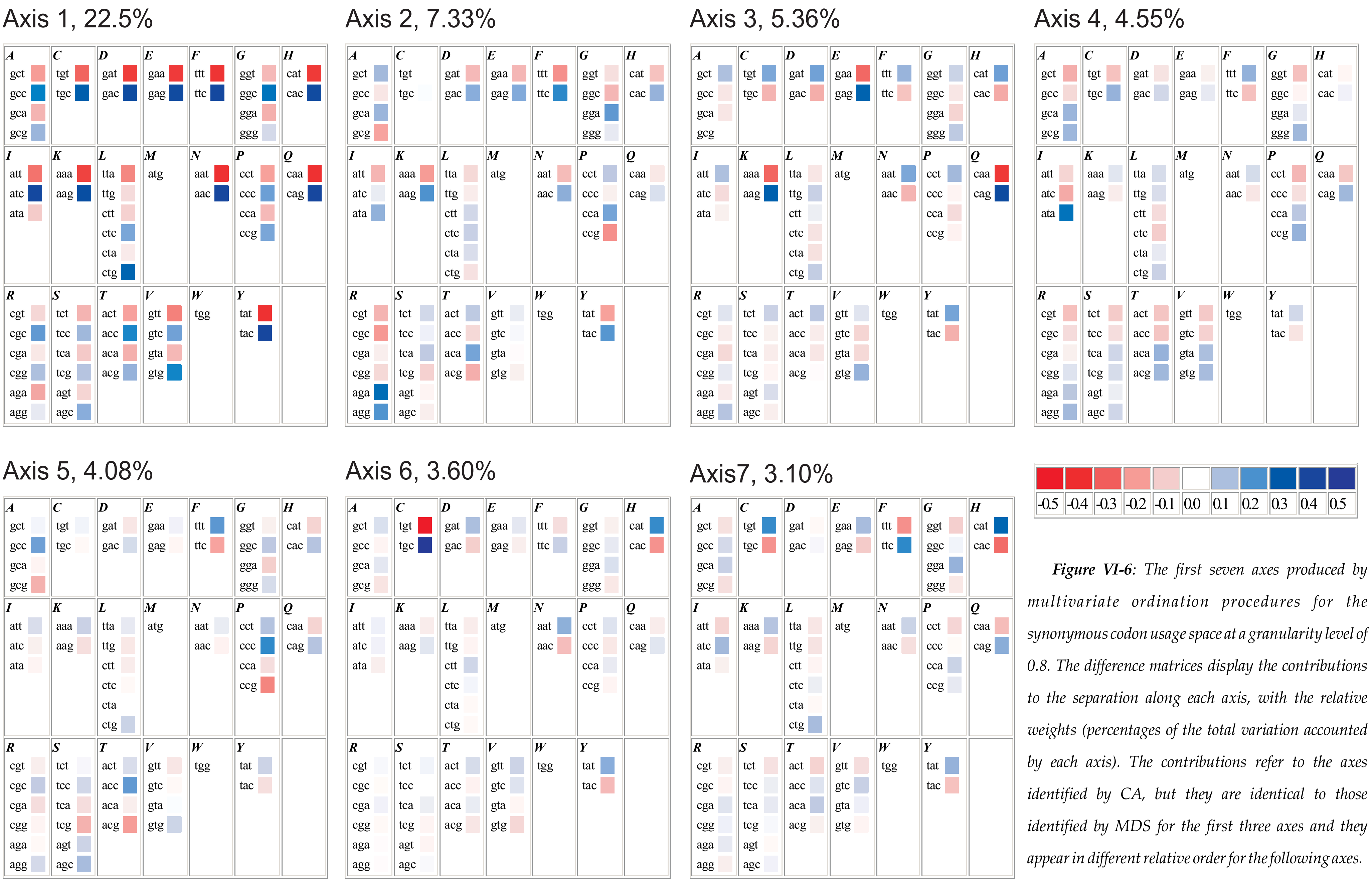


Figure VI-6: *The first seven axes produced by multivariate ordination procedures for the synonymous codon usage space at a granularity level of 0.8. The difference matrices display the contributions to the separation along each axis, with the relative weights (percentages of the total variation accounted by each axis). The contributions refer to the axes identified by CA, but they are identical to those identified by MDS for the first three axes and they appear in different relative order for the following axes.*

D.2.2 Analysis of the individual dimensions and identification of common biases

A detailed representation of the relative usage of the synonymous triplets is given in Figure VI-7, where each boxplot stands for a dimension of the codon space.

There are only few extremely deviant distributions for three-fold, four-fold and six-fold degenerate amino acids: for these synonymous sets there are few cases in which a single triplet is extremely over-represented over all its synonymous alternatives. In the two-fold case, however, the complete range of the frequency values (almost from 0 to 1) is present. Of course this is largely due to the use of filtered data, codon usages containing the full repertoire of triplets. In the AA-filtered space, when only one synonymous triplet is used to code for an amino acid, its relative frequency is 1 and 0 are the relative frequencies of its synonyms.

The distributions of the values (which come from almost uniformly spaced codon vectors) identify some triplets which are universally under-represented, often very slightly, while in other cases more substantially (like the ATA triplet for Isoleucine or CTA for Leucine). The most noticeable under-representations are those of NTA codons (where N stands for any nucleotide), which in the case of Leucine triplets is made more obvious by the codon profile combined-contribution dimensions *La3* and *Lg3* (corresponding to YTA and YTG, where Y = C or T), markedly more different than the equivalent third position combined-contribution dimensions of Arginine and Serine. This observation is consistent with previous studies on dinucleotide abundances (Karlin and Burge, 1995; Karlin *et al.*, 1998; see also section I C.4.1) which identified under-representation of the TA dinucleotide in both eukaryotes and prokaryotes. Here it becomes apparent that this is an almost universal state, which indicates the probable presence of a constraint (either due to a mutational bias or to negative selection pressure).

The reasons for the relative scarcity of TA in nucleotide sequences are not clearly understood. It may be due to selection related to the susceptibility of UA in the messenger RNA, which appears to be a preferential target for ribonucleases (Beutler *et al.*, 1989), although bias against TA has been reported also for noncoding regions (in humans; Karlin and Mrázek, 1996). However, according to a study by Duret and Galtier (2000), a substantial part of the observed departures from expected frequencies of the dinucleotide TA (in humans) are a mathematical artefact. A more general reason could be the low thermodynamic stacking energy of this dinucleotide (Delcourt and Blake, 1991). Furthermore, because of the presence of TA in many regulatory signals (like TATA box or the polyadenylation signal) it has been suggested that TA suppression could reduce inappropriate binding of regulatory factors (Karlin and Mrázek, 1997).

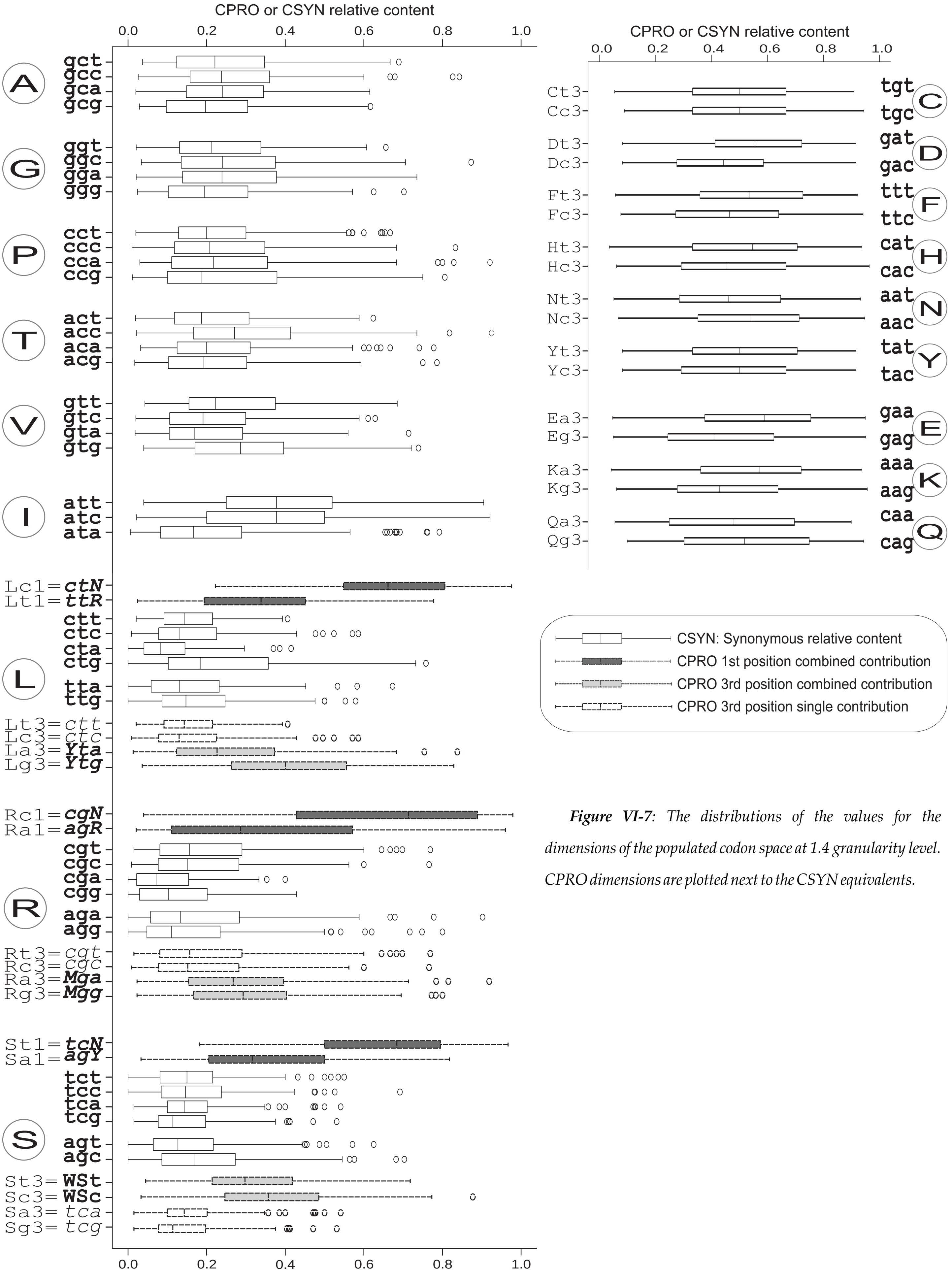


Figure VI-7: *The distributions of the values for the dimensions of the populated codon space at 1.4 granularity level. CPRO dimensions are plotted next to the CSYN equivalents.*

D.2.3 Convexity of the space

One of the possible tests which can be applied to the populated codon space model is a test of convexity. This question naturally arises both from a desire to get a better understanding of the shape of the populated space, and from the need to verify the amount of features which could be lost in the mapping by using long binning radii. As exemplified by Figure VI-1 above, a longer radius can conceal certain features in the shape of the data set, such as the presence of “holes”, empty inner regions.

A very basic search for convexity was hence performed, computing the middle points (*midpoints*) between each couple of vectors belonging to the populated space, and testing whether they lie in a non-populated region. If the shape of the space were similar to the letter “C” (as an example in two-dimensions), the number of midpoints found in non-populated space would be very high. If instead the space had the shape of a solid disc without holes, all the midpoints would lie inside populated space. Even very long granularity radii would not hide a situation like the “C-shape”, while small internal cavities (“holes”) might not become apparent. Small holes (identified by no more than one midpoint found in a non-populated area) are probably not significant, whereas when many midpoints are found in non-populated space (a large hole), this would indicate the possibility of a relevant constraint preventing codon usages with those codon frequencies.

As a matter of fact, there are no midpoints found to lie in non-populated space for granularity radii equal to or longer than 1.3, while only seven of them are found at a granularity of 1.2. Testing the codon space model for proximity of these points to the representatives reveals that they are well inside the core, not far from the average (they are marked in the map of Figure VI-4).

These results exclude the possibility of a very convex shape or the presence of significant cavities in the codon space, which appears instead to have a solid hyper-spherical shape.

D.3 Subsets of the codon space

Beyond representing the extension of the biological world in the codon space – according to the available sequences – it is also possible and desirable to analyse subsets of it, like the space of vertebrate codon usages or the one occupied by prokaryotic ones.

D.3.1 The vertebrate space

The under-representation of synonymous triplets that can be observed in the vertebrate space (Figure VI-8) is consistent with previous reports of dinucleotide frequencies and optimal codons (for example Karlin and Mrázek, 1996; Kanaya *et al.*, 2001a). There are strong biases in the usage of NCG triplets (Alanine GCG, Proline CCG, Serine TCG, Threonine ACG), which are under-represented. This is due to the well known CG deficiency observed in vertebrates: the frequency of the dinucleotide CG is up to five times lower than the product of C and G frequencies (Bird, 1980). This deficiency is the consequence of a mutational bias: the methylation dependent CG→TG mutation. The TA deficiency observed for the whole codon space (see previous section for possible causes) appears stronger in the vertebrate space, with the third quartile of the distributions of NTA triplets being lower than the first quartile of the synonymous alternatives.

Less pronounced, but still noticeable, is the under-representation of NAT, NTT and NAA triplets. Besides a certain tendency towards GC-rich codons, this could be explained by a preference for WWC codons over WWA (where W = T or A). WWC codons were found to be preferred over their WWA synonyms in several unicellular organisms (Sharp and Devine, 1989; Andersson and Sharp, 1996; Kanaya *et al.*, 1999) for reasons of optimal codon-anticodon interaction energy: translational efficiency is greater for triplets favouring a codon-anticodon interaction with higher binding strength.

Mutation biases or translational selection? The question is still unresolved and the answer is most probably a combination of these and other factors (*q.v.* I C.3), as found in several prokaryotic species. Nevertheless, differences in isoaccepting tRNA gene copy numbers (found correlated to tRNA abundances in species where translational efficiency is a codon usage shaping force) are low in vertebrates, which also have a larger set of tRNA species, a fact that, together with small effective population sizes (see I C.1.5) would tend to suggest an absence of selection acting towards translational efficiency (Urrutia and Hurst, 2001; but see Haas *et al.*, 1996).

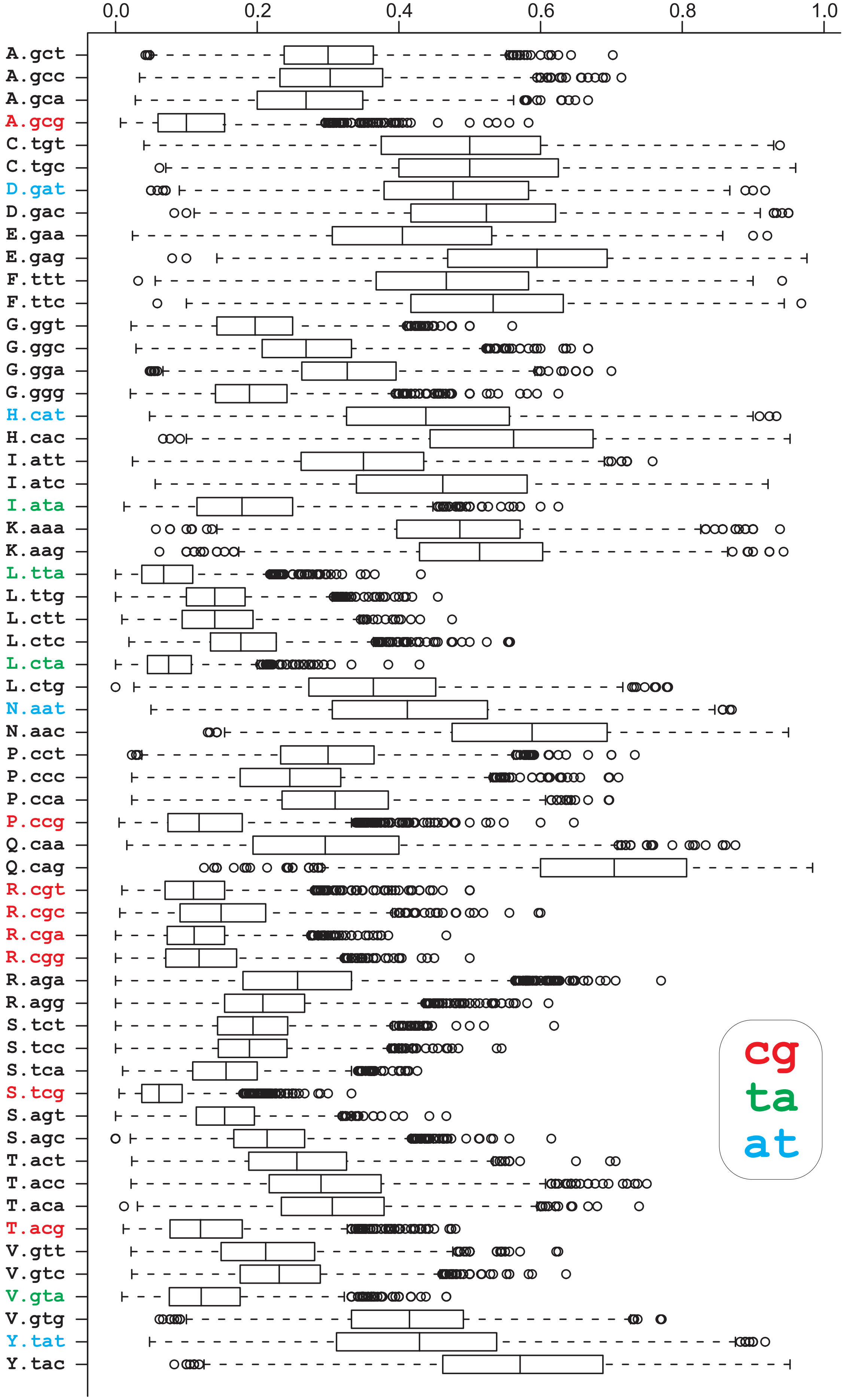


***Figure VI-8**: The dimensions of the vertebrate space, in boxplot representation. The usage of synonymous codons is clearly biased, even when all vertebrate sequences are considered together, where biases are observed between distributions of codon usages rather than between single codon usages.*

D.3.2 The prokaryotic space

The prokaryotic space is one of the subsets of the codon space for which many completely sequenced genomes are available. The low-dimensional representation of the genomic (average) codon usage vectors for all the completely sequenced prokaryotic genomes, Figure VI-9, reflects the one relative to the whole codon space. Since the prokaryotes are more widely spread over the codon space, this is not surprising. But analysing this subset reveals how archaea and bacteria occupy two different regions identified by the second axis of separation (particularly associated with the AGR triplets for Arginine; see D.2.1). The first coordinate of separation for these genomes is related to G+C content (in particular GC3).

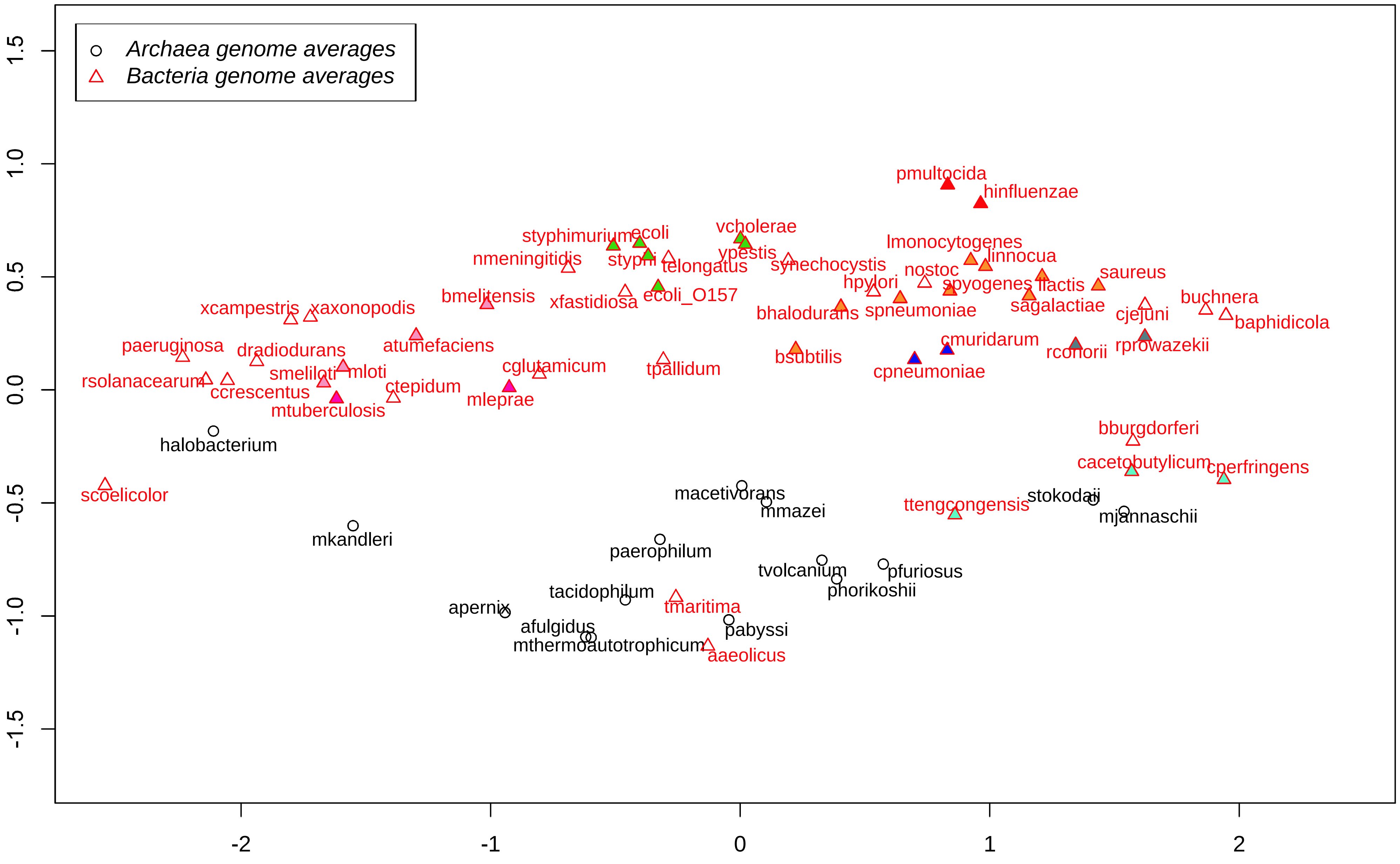


Figure VI-9*:Multidimensional scaling of the completely sequenced prokaryotic genomes. Archaea and three bacteria occupy the lower side of themap.*

Archaea occupy a well defined region in which only three of the sequenced bacteria can be found, namely *T.maritima*, *A.aeolicus* and *T.tengcongensis*, hyper-thermophilic bacteria (their optimal growth temperature being beyond 80° Celsius) whose placement indicates codon usage patterns similar to those of archaea, in line with the observation that they contain a large number of genes similar to those of thermophilic archaea (Nelson *et al.*, 1999; Ochman *et al.*, 2000).

The bacterial kingdom is phylogenetically extremely diverse (Olsen *et al.*, 1994), there was even a proposal to create twelve bacterial kingdoms to reflect the great differences inside this group. The biggest separation is the one between high and low G+C content. The biodiversity of bacteria is here reflected by their codon usage diversity.

The codon usages of the atypical transcripts (those differing significantly from their genomic bias) are plotted over the map of prokaryotic genomes in Figure VI-10. This reveals how even the most atypical transcripts are mainly restricted to their own region, either bacterial or archaeal. In fact the average distance computed between all archaeal versus all bacterial atypical transcripts is 2.957 units (with a standard deviation of 0.525). There are nevertheless several contact points, many archaeal transcripts in 'bacterial codon space' and vice versa.

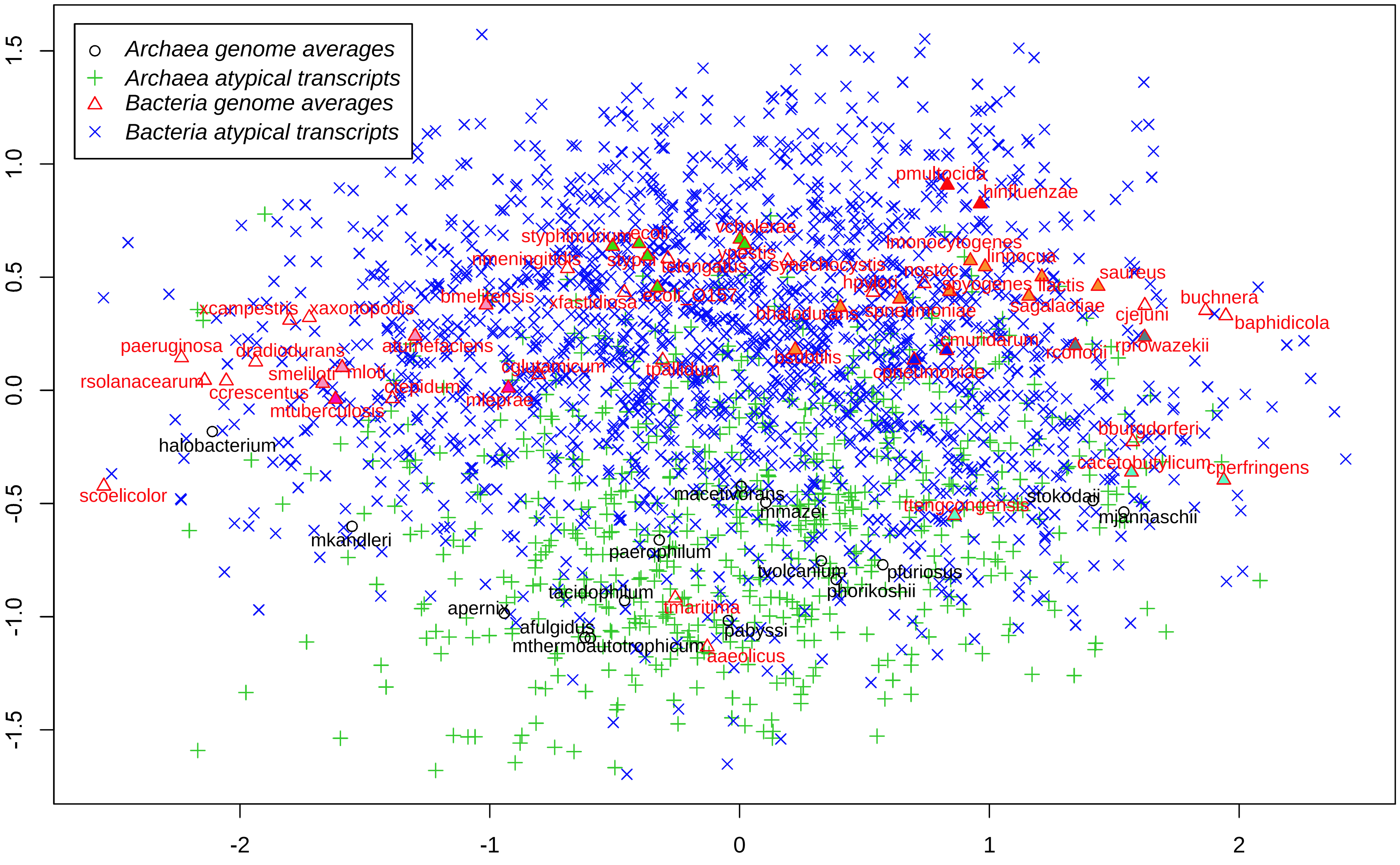


Figure VI-10:*Multidimensional scaling of the sequenced prokaryotic genomes alongside the transcripts atypical to each genome. OnlytheAA-filtered transcriptswhose Euclidean distance from their genome is higher than 1.8 units are included, as in the analysis of chapter IV. Themajority of the atypical transcripts are confined to their respective bacterial or archaeal region of the codon space.*

D.4 Comparisons with the theoretical space

A very important feature which emerges from the sampling of the populated space is the observation that it is tightly confined, occupying but a fraction of the theoretical space. A quantification of its limited coverage is possible by comparison with randomly generated codon usages.

The change in the number of representative points identified at different granularity radii for the biological space (previously shown as Figure VI-2) was compared to those identified on random codon spaces (codon spaces comprising only randomly generated codon usages): Figure VI-11 (see also section C.2.6).

For lower granularity radii the theoretical space is exceedingly large and the randomly generated codon usages are extremely sparse, they are not grouped together by the binning procedure under a radius of 0.7-0.8, with the exception of the codon usages generated by ‘random triplets’, which are not very deviant from the average codon vector (*i.e.* not far from average frequency distributions). The biological codon space instead contains many very similar codon usages (higher proximity, less scattering between the codon vectors), which are grouped together in representatives even when very short radii are used. The biological codon space hence appears more compact, not very sparse and not very diverse (as was shown also by the multivariate ordination techniques, with the high correlation between triplets and the first separation axes accounting for most of the variation).

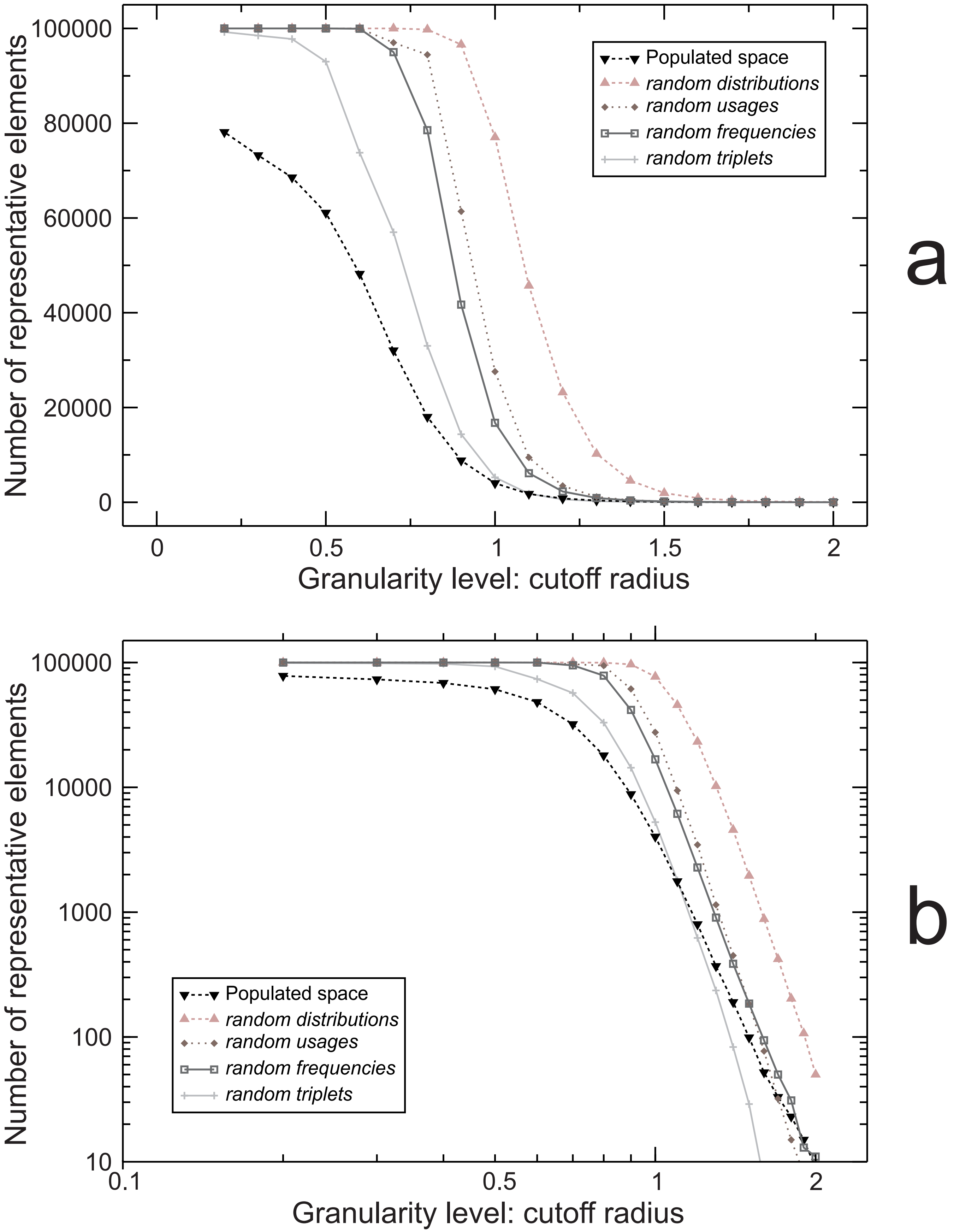


Figure VI-11: *Comparison between the populated (biological) space and randomly generated codon spaces in linear scale (a) and logarithmic scale (b). The codon usage generation algorithms are described in section C.2. The 'random usages' algorithm was set to generate synonymous frequencies of at least 0.1 for each triplet and at least 0.2 in the case of 2-fold degenerate amino acids. The number of representatives found is plotted against the granularity radius. The absence of a constant regime (slope=0) at low granularity radii for the curve corresponding to the biological space is due of the high number of very similar codon usages: a total of 133,232 codon vectors are grouped in 78,115 representatives at a 0.2 granularity radius. All the randomly generated codon usages (100,000 for each algorithm) are instead separated in individual bins at that radius, indicating the high sparsity (heterogeneity) of the random usages.*

D.4.1 Non-populated sampling

Another comparison between biological and theoretical spaces can be made by sampling the populated codon space using randomly generated codon usages (C.2). In this procedure the random usages are tested for proximity to the representative points of the populated space: if the generated vector is found at a distance greater than a cut-off distance from all representative data points in the space (the representatives for the populated space and the previously accepted random vectors), then it is included in the space (hence preventing further sampling in that zone).

The *random usages/distribution* algorithms have uncorrelated synonymous sets and find many more vectors representing the non-populated space than those that are found by the *random frequencies* algorithm (whose codon usages are constrained by total nucleotide relative frequencies, similar to how the biological codon usages are constrained). A brief resume of the sampling is reported in Table VI-3.

Algorithm	**samples**	**estimation of non-populated space**
random frequencies	1013	85.0%
random usages minimum 0.2	7794	97.8%
Random distributions (coarse)	38349	99.5%

***Table VI-3**: Estimation of the ratio between populated and non-populated space from a comparison at 1.4 granularity radius between the representatives of the biological codon usages (179 representatives see Table VI-2) and random codon usages generated by different algorithms. The randomly generated vectors are all at a distance of at least 1.4 units from any point of the populated space. The percentage indicated is the amount of non-populated space over total space (total space as populated plus non-populated). If the estimation is done on the minimum number found for populated space representatives (near-optimal coverage: 163 representatives) instead of on the average, the percentages are slightly different: 86.1%, 98.0% and 99.6%.*

The proportion between non-populated space and populated space increases with shorter radii, so the percentages reported in the table (relative to a granularity radius of 1.4) are biased towards the populated space and should be considered an under-estimation.

D.5 General considerations regarding the populated space

Using several different and complementary analyses, it was possible to show how constrained and confined the region of the populated space is.

The distribution of values for each dimension (D.2.2) revealed the ranges for the maximum variation of the synonymous frequencies and identified a common trend of under-represented triplets. The absolute ranges indicate that not all of the most deviant distributions of frequencies are present (those with a triplet used almost exclusively over its synonymous alternatives, which would show values approximating 1 for that dimension).

The observations on the convexity of the codon space (D.2.3) precluded the possibility that the codon space is "C-shaped" or "L-shaped" (using a two-dimensional analogy) or that it contains significant empty regions (cavities) and instead suggested a solid hyper-spherical shape.

The low-dimensional representations and the identification of the axes accounting for the largest fractions of the variation between codon usages (D.2.1) showed how the very general correlations between triplets limit the total 'exploration' of the theoretical space by the biological sequences. In other words, they indicate the presence of constraints limiting the divergence of codon usages.

This aspect was verified using randomly generated codon usages of different inherent degrees of correlation (D.4) which identified a very large number of non-populated regions and hence a very limited extension of the populated space, whose coverage over the theoretical one was estimated.

The whole populated space is hence mainly composed of codon vectors with interdependent relationships, apparently preventing a more heterogeneous spread over the theoretical space. This result will need to be verified as more sequence data becomes available.

Using a linguistic analogy, there is a maximum number of consecutive consonants (in consonantal clusters) that can be pronounced and recognized. Furthermore, there are correlations in the set of phonemes present in a natural language. For example, if there is one phoneme for a place of articulation, such as palatal or labial, then the language will also include other phonemes produced at the same place.

The mechanisms of speech production and perception limit the effective possible divergence of natural languages. In a conceptually similar way, constraints such as

DNA structure, information content, translational efficiency and message superimposition, limit the divergence of codon usages.

E Conclusions

The theoretical set of all possible codon usages was defined as *codon space* and several tools to represent it and to analyse it were devised. The region encompassing all the codon usages found in the currently determined sequences (the *populated space*) was mapped at different levels of detail into models, in which a limited number of codon vectors represent the populated space without redundancy. The aim was to obtain an even and unbiased representation without losing information on the occupancy of even lowly populated regions.

The codon space model was developed to find features common to all codon usages and allowed to quantify the limitations to the divergence of codon usage among the species.

This model can also find practical applications like proximity searching, namely the comparison of a codon usage to the model, allowing rapid retrieval of the codon usages which are most similar to the query. Furthermore, the model could be used in theoretical studies of evolution and in gene-prediction algorithms.

Comparing the biological data to several types of randomly generated codon usages made the estimation of the extension and heterogeneity of the populated space possible, showing the very small portion of the theoretical space which is populated and the high correlation between non-synonymous triplets. The populated space appears as a convex and non-hollow region inside the multivariate space, centred on the average codon vector and with a high degree of correlation between its dimensions.

Using multivariate ordination procedures the codon space was represented in low-dimensional plots and the principal coordinates of separation revealed the general trends of variation between large taxonomical groups, which were also characterised in their maximum spread over the codon space. The triplet contributions to the most significant axes of separation were presented as codon difference matrices.

Common biases to synonymous usage, in the form of universally under-represented synonymous triplets (codons which are never found in high relative proportion to their alternative synonyms), were observed and commented on in the light of previous studies.

A comprehensive picture of the biological codon space is very difficult to obtain, for two main reasons. Firstly, the available data is but a fraction of the extant species and a very biased one. Secondly, humans are poor at seeing structure in a large number of dimensions, such as those of the codon space. A blurred shadow (blurred because of limited data, shadow because of the projection onto few dimensions) is the best to be expected, but this first approximation can still prove useful in the investigation of universal constraints or of the major trends among taxonomic groups. More points of view need to be investigated, in a similar way as to how several projections of an object onto a wall can better reveal its three dimensional shape. More data needs to be obtained, to understand what percentage of the observed trends is a result of sampling bias or of missing data.

As the genomic sequences of more and new species are determined, it will be possible to verify whether the total coverage of the theoretical codon space will remain as low as it was observed in this study. Although it might not be possible to reach a final conclusion on this issue (as we might never have the complete sequence of every species), a periodic re-evaluation of the codon space will show the trend in novel data: when the discovery of novel codon usages will decrease (and how steeply) and how the coverage of the theoretical codon space will change.

General conclusions

With the exponentially increasing volume of DNA sequences becoming available, it is important to develop general and automated procedures to analyse, to compare and to present this data. In fact, as in many other fields, the gap between the amount of data that is generated and stored, and the amount of data which is actually studied, is rapidly growing.

Codon usage is probably the most informative aspect that can be analysed in the coding sequences, which are known to contain several superimposed messages. The knowledge of codon usage patterns can be used, among others, to optimise the levels of translation, to estimate the degree of sequence conservation and the rate of mutation, to back-translate protein sequences to their probable nucleotide counterparts, to identify imported genes and to assist in the prediction of protein-coding sequences.

In this work the codon usage information was studied in several domains, including the recently determined sequences from the eukaryotic genome projects. Different approaches were combined and new tools and methodologies were developed for comprehensive systematic analyses. It was shown that an alternative point of view on synonymous codon usage is possible, and that it performs as well as the classical method. Both schemes were used to study codon usage patterns, to predict horizontally transferred genes and to identify significantly atypical codon usages.

A measure of genomic heterogeneity was devised as a function of codon usage distances from the average bias. The histogram of distances reveals which species have the most diversified codon usage patterns. Intra-genomic heterogeneity was also compared across the species, with the simple but effective boxplot representations, or with the less straightforward but more accurate visualisation of the principal components of variation. Viruses were shown to have very different degrees of intra-genomic heterogeneity, while prokaryotic species have similar proportions of normal and atypical codon patterns.

The acquired knowledge on the patterns of intra-genomic heterogeneity lead to the development of a methodology for the identification of Horizontal Gene Transfers. The measure of codon usage dissimilarity was used in the comparison of all the transcripts which are atypical in their own genomic context to the codon biases of all the other genomes. The identified matches were linked in regions according to their location in the genome and their sequence homologues were retrieved from protein databases. It was shown that it is possible to characterise donor/acceptor relationships combining the compositional detection method with a phylogenetic verification.

All the developed procedures are general, efficient, automated and scalable, all of which are fundamental requirements in the genomic era.

Finally, the non-randomness of the codon usages was explored at the largest possible scale. This was made possible by the construction of models, which represent the spread of the available sequences over the theoretical space of possible codon usages. The characterisation of the codon space provided both qualitative and quantitative evaluation of the limitations which influence codon usage divergence among the species.

The knowledge about the absolute ranges of the variation among synonymous usage and of the major trends of correlation (computed either globally or for a specific subset of the codon space), can find applications in gene prediction, in the analyses of information content, in the estimation of sequence conservation and in the studies on the overlap of biological messages.

Since the currently available data, which may or may not be representative, is only a tiny fraction of the actual biodiversity, the picture drawn using these methods might change significantly with the determination of a wider range of genomic sequences. Hence, a periodic update of the codon space model can also provide a measure of the sequence diversity which has been so far observed and stored.

Bibliography

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997). *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucl. Acid Res. **25**, 3389-3402

Andersen E.S., Jeeninga R.E., Damgaard C.K., Berkhout B., Kjems J. (2003). *Dimerization and template switching in the 5′ untranslated region between various subtypes of Human Immunodeficiency Virus type 1.* J.Virol. **77**, 3020-3030

Anderson M.J., Willis T.J. (2003). *Canonical analysis of principle coordinates: a useful method of constrained ordination for ecology*. Ecology **84**, 511-525

Andersson G.E., Sharp P.M. (1996). *Codon usage in the Mycobacterium tuberculosis complex.* Microbiol. **142**, 915-925

Aparicio S., Chapman J., Stupka E., Putnam N., Chia J.M., Dehal P., Christoffels A., Rash S., Hoon S., Smit A., Gelpke M.D., Roach J., Oh T., Ho I.Y., Wong M., Detter C., Verhoef F., Predki P., Tay A., Lucas S., Richardson P., Smith S.F., Clark M.S., Edwards Y.J., Doggett N., Zharkikh A., Tavtigian S.V., Pruss D, Barnstead M., Evans C., Baden H., Powell J., Glusman G., Rowen L., Hood L., Tan Y.H., Elgar G., Hawkins T., Venkatesh B., Rokhsar D., Brenner S. (2002). *Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.* Science **297**, 1301-1310

——

Baisnée P.F., Baldi P., Brunak S., Pedersen A.G. (2001) *Flexibility of the genetic code with respect to DNA structure.* Bioinformatics **17**, 237-248

Baldi P., Brunak S., Chauvin Y., Krogh A. (1996). *Naturally occurring nucleosome positioning signals in human exons and introns.* J. Mol. Biol. **263**, 503-510

Bellman R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Rapp B.A., Wheeler D.L. (2000). *GenBank*. Nucleic Acids Research **28**, 15-18

Benzécri J.P. (1973). *L'analyse des données*. L'Analyse des Correspondances. 2, Dunod, Paris

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

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

Beutler E., Gelbart T., Han J.H., Koziol J.A., Beutler B. (1989). *Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage*. Proc. Natl. Acad. Sci. USA. **86**, 192-196

Bhattacharyya A. Stilwagen S. Reznik G. Feil H. Feil W.S., Anderson I., Bernal A., D'Souza M., Ivanova N., Kapatral V., Larsen N., Los T., Lykidis A., Selkov E. Jr, Walunas T.L., Purcell A., Edwards R.A., Hawkins T., Haselkorn R., Overbeek R., Kyrpides N.C., Predki P.F. (2002). *Draft sequencing and comparative genomics of Xylella fastidiosa strains reveal novel biological insights*. Genome Res. **12**, 1556-63

Bird A.P. (1980). *DNA methylation and the frequency of CpG in animal DNA*. Nucl. Acids Res. **8**, 1499-1504

Bolshoy A., Shapiro K., Trifonov E.N., Ioshikhes I. (1997). *Enhancement of the nucleosomal pattern in sequences of lower complexity.* Nucl. Acids Res. **25**, 3248-3254

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

Brukner I., Sánchez R., Suck D., Pongor S. (1995). *Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides*. EMBO J. **14**, 1812-1818

Burge C., Campbell A.M., Karlin S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA **89**, 1358-1362

——

Cain D., Erlwein O., Grigg A., Russell R.A., McClure M.O. (2001). *Palindromic sequence plays a critical role in human foamy virus dimerization*. J. Virol. **75**, 3731-3739

Chai N.N., Zhou H., Hernandez J., Najmabadi H., Bhasin S., Yen P.H. (1998). *Structure and organization of the RBMY genes on the human Y chromosome: transposition and amplification of an ancestral autosomal hnRNPG gene*. Genomics **49**, 283-289

Chargaff E. (1951). *Structure and function of nucleic acids as cell constituents*. Fed. Proc. **10**, 654-659

Cheeseman P. and Stutz J. (1996). *Bayesian classification (AutoClass): Theory and results. Advances in Knowledge Discovery and Data Mining*. AAAI Press

Chen C., Yang T.P. (2001). *Nucleosomes are translationally positioned on the active allele and rotationally positioned on the inactive allele of the HPRT promoter*. Mol. Cell Biol. **21**, 7682-7695

Chen G.T., Inoue M. (1994). *Role of the AGA/AGG codons, the rarest codons in global gene expression in Escherichia coli*. Genes Dev. **8**, 2641-2652

Cox T.F., Cox M.A.A. (1994). *Multidimensional scaling.* Chapman and Hall, London

——

DeBry R.W., Marzluff W.F. (1994). *Selection on silent sites in the rodent H3 histone gene family*. Genetics **138**, 191-202

Deering W., West B.J. (1992). *Fractal physiology*. IEEE Engin. Med. Biol. **11**, 40-46

Delcourt S.G., Blake R.D. (1991). *Stacking energies in DNA*. J. Biol. Chem. **266**, 15160-15169

Deml L., Bojak A., Steck S., Graf M., Wild J., Schirmbeck R., Wolf H., Wagner R. (2001). *Multiple effects of codon usage optimization on expression and immunogenicity of DNA candidate vaccines encoding the human immunodeficiency virus type 1 Gag protein*. J. Virol. **75**, 10991-11001

Deschavanne P.J., Giron A., Vilain J., Fagot G., Fertil B. (1999). *Genomic signature: characterization and classification of species assessed by chaos game representation of sequences*. Mol. Biol. Evol. **16**, 1391-1399

Dong H., Nilsson L., Kurland C.G. (1996). *Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates.* J. Mol. Biol. **260**, 649-663

Duret L., Hurst L.D. (2001). *The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution*. Mol. Biol. Evol. **18**, 757-762

Dutta C., Pan A. (2002). *Horizontal gene transfer and bacterial diversity*. J. Biosci. **27**, 27-33

Echarri A., Gonzalez M.E., Carrasco L. (1996). *Human immunodeficiency virus (HIV) Nef is an RNA binding protein in cell-free systems*. J. Mol. Biol. **262**, 640-651

Enright A.J., Van Dongen S., Ouzounis C.A. (2002). *An efficient algorithm for large-scale detection of protein families.* Nucl. Acid Res. **30**, 1575-1584

Eyre-Walker A., Bulmer M. (1993). *Reduced synonymous substitution rate at the start of enterobacterial genes*. Nucl. Acids Res. **21**, 4599-4603

Felsenstein J. (1989). *PHYLIP -- Phylogeny Inference Package (Version 3.2)*. Cladistics **5**, 164-166

Fitter A.H., Strickland T.R. (1992). *Fractal characterization of root system architecture*. Funct. Ecol. **6**, 632-635

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A., Merrick J.M., *et al.* (1995). *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd*. Science **269**, 496-512

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

Garcia-Vallvé, Romeu A., Palau J. (2000). *Horizontal gene transfer in bacterial and archaeal complete genomes.* Genome Research **10**, 1719-1725

Gardiner E.J., Hunter C.A., Packer M.J., Palmer D.S., Willett P. (2003). *Sequence-dependent DNA Structure: A Database of Octamer Structural Parameters*. J. Mol. Biol. **332**, 1025-1035

Gautier C. (2000). *Compositional bias in DNA*. Curr. Opin. Genet. Dev. **10**, 656-661

Gelfand M.S., Koonin E.V. (1997). *Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes*. Nucl. Acids Res. **25**, 2430-2439

Geyer M., Fackler O.T., Peterlin B.M. (2001). *Structure--function relationships in HIV-1 Nef*. EMBO Rep. **2**, 580-585

Gough J., Karplus K., Hughey R., Chothia C. (2001). *Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure*. J. Mol. Biol. **313**, 903-919

Gower J.C. (1966). *Some distance properties of latent root and vector methods used in multivariate analysis*. Biometrika **53**, 325-328

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

Grantham R., Gautier C., Gouy M., Jacobzone M., Mercier R. (1981). *Codon catalog usage is a genome strategy modulated for gene expressivity.* Nucleic Acids Res **9**, r43-r74

Greenacre M.J. (1984). *Theory and applications of correspondence analysis.* 1st Ed. Academic Press, London

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

Haas J., Park E.C., Seed B. (1996). *Codon usage limitation in the expression of HIV-1 envelope glycoprotein*. Curr. Biol. **6**, 315-324

Hayashi K., Munakata N. (1984). *Basically musical*. Nature. **310**, 96

Hertz G.Z., Young M.R., Mertz J.E. (1987). *The A+T-rich sequence of the simian virus 40 origin is essential for replication and is involved in bending of the viral DNA.* J. Virol. **61**, 2322-2325

Holm L. (1986). *Codon usage and gene expression*. Nucleic Acids Res. **14**, 3075-3087

Holm L., Sander C. (1996). *Mapping the protein universe.* Science **273**, 595-603

Holm L., Sander C. (1998). *Touring protein fold space with Dali/FSSP*. Nucl. Acids Res. **26**, 316-319

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., Durbin R., Eyras E., Gilbert J., Hammond M., Huminiecki L., Kasprzyk A., Lehväslaiho H., Lijnzaad P., Melsopp C., Mongin E., Pettett R., Pocock M., Potter S., Rust A., Schmidt E., Searle S., Slater G., Smith J., Spooner W., Stabenau A., Stalker J., Stupka E., Ureta-Vidal A., Vastrik I., Clamp M. (2002). *The Ensembl genome database project*. Nucleic Acids Res. **30**, 38-41

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

——

Ihaka R., Gentleman R. (1996). *R: a language for data analysis and graphics*. J. of Computational and Graphical Statistics **5**, 299-314

Ikemura T. (1981). *Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes.* J. Mol. Biol. **151**, 389-409

Ikemura T. (1982). *Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs.* J. Mol. Biol. **158**, 573-597

Ikemura T. (1992) in: Hatfield D.L., Lee B.J., Pirtle R.M. (eds) *Transfer RNA in protein synthesis.* CRC Press, Boca Raton, 87-111

Ioshikhes I., Bolshoy A., Derenshteyn K., Borodovsky M., Trifonov E.N. (1996). *Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences*. J. Mol. Biol. **262**, 129-139

——

Jain R., Rivera M.C., Moore J.E., Lake J.A. (2002). *Horizontal gene transfer in microbial genome evolution.* Theor. Pop. Biol. **61**, 489-495

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

Jeffries M. (1993). *Invertebrate colonization of artificial pondweeds of differing fractal dimension*. Oikos **67**, 142-148

——

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 tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis*. Gene **238**, 143-155

Kanaya S., Yamada Y., Kinouchi M., Kudo Y., Ikemura T. (2001a). *Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis.* J. Mol. Evol **53**, 290-298

Kanaya S., Kinouchi M., Abe T., Kudo Y., Yamada Y., Nishi T., Mori H., Ikemura T. (2001b). *Analysis of codon usage diversity of bacterial genes with a self-organising map (SOM): characterization of horizontally transferred genes with emphasis on the E.coli O157 genome*. Gene **276**, 89-99

Karlin S., Ladunga I. (1994). *Comparisons of eukaryotic genomic sequences*. Proc. Natl. Acad. Sci. USA **91**, 12832-12836

Karlin S., Burge C. (1995). *Dinucleotide relative abundance extremes: a genomic signature*. Trends Genet. **11**, 283-290

Karlin S., Mrázek J. (1996). *What drives codon choices in human genes?* J. Mol. Biol. **262**, 459-472

Karlin S., Mrázek J. (1997). *Compositional differences within and between eukaryotic genomes*. Proc. Natl. Acad. Sci. USA **94**, 10227-10232

Karlin S., Mrázek J., Campbell A.M. (1997). *Compositional biases of bacterial genomes and evolutionary implications*. J. Bacteriol. **179**, 3899-3913

Karlin, S. (1998). *Global dinucleotide signatures and analysis of genomic heterogeneity.* Curr. Opin. Microbiol. **1**, 598-610

Karlin S., Mrázek J., Campbell A.M. (1998). *Codon usages in different gene classes of the Escherichia coli genome*. Mol. Microbiol. **29**, 1341-1355

Kohonen T., Oja E., Simula O., Visa A., Kangas J. (1996). *Engineering applications of the self-organizing map*. Proc. IEEE **84**, 1358-1384

Koonin E.V., Mushegian A.R., Galperin M.Y., Walker D.R. (1997). *Comparison of archaeal and bacterial genomes: computer analysis of protein sequence predicts novel functions and suggests a chimeric origin for the archaea.* Mol. Microbiol. **25**, 619-637

Koonin E.V., Makarova K.S., Aravind L. (2001). *Horizontal gene transfer in prokaryotes: quantification and classification.* Annu. Rev. Microbiol. **55**, 709-742

Koski L.B., Morton R.A., Golding G.B. (2001). *Codon bias and base composition are poor indicators of horizontally transferred genes.* Mol. Biol. Evol. **18**, 404-412

Krecic A.M., Swanson M.S. (1999). *hnRNP complexes: composition, structure, and function*. Curr. Opin. Cell Biol. **11**, 363-371

Kunisawa T., Kanaya S., Kutter E. (1998). *Comparison of synonymous codon distribution patterns of bacteriophage and host genomes.* DNA Res. **5**, 319-326

——

Lafay B., Atherton J.C., Sharp P.M. (2000). *Absence of translationally selected synonymous codon usage bias in Helicobacter pylori*. Microbiology **146**, 851-860

Lawrence J.G., Ochman H. (1997). *Amelioration of bacterial genomes: rates of change and exchange.* J. Mol. Evol. **44**, 383-397

Lawrence J.G., Ochman H. (1998). *Molecular archaeology of the Escherichia coli genome.* Proc. Natl. Acad. Sci. U.S.A. **95**, 9413-9417

Lawrence J.G., Ochman H. (2001). *Reconciling the many faces of lateral gene transfer.* Trends in Microbiol. **10**, 1-4

Lee K.Y., Wahl R., Barbu E. (1956). *Contenu en bases puriques et pyrimidiques des acides désoxyribonucléiques des bactéries*. Ann. Inst. Pasteur **91**, 212-224

Levitsky V.G., Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Kolchanov N.A. (1999). *Nucleosomal DNA property database*. Bioinformatics **15**, 582-592

Lewis M., Rees D.C. (1985). *Fractal surfaces of proteins*. Science **230**, 1163-1165

Li H., Luo L. (1996). *The relation between codon usage, base correlation and gene expression level in Escherichia coli and yeast.* J. Theor. Biol. **181**, 111-124

Longley P.A., Batty M. (1989). *On the fractal measurement of geographical boundaries*. Geogr. Anal. **21**, 47-67

——

Malim M.H., Hauber J., Le S.Y., Maizel J.V., Cullen B.R. (1989). *The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA*. Nature **338**, 254-257

Mandelbrot B.B. (1975). *Stochastic models for the Earth's relief, the shape and the fractal dimension of the coastlines, and the number-area rule for islands*. Proc. Nat. Acad. Sci. USA **72**, 3825-3828

Mardia K.V., Kent J.T., Bibby J.M. (1979). *Chapter 14 of Multivariate Analysis*. London: Academic Press

Mathé C., Peresetsky A., Déhais P., Van Montagu M., Rouzé P. (1999). *Classification of Arabidopsis thaliana gene sequences: clustering of coding sequences into two groups according to codon usage improves gene prediction.* J. Mol. Biol. **285**, 1977-1991

McInerney J.O. (1997). *Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns.* Microbial Comp. Genomics **2**, 1-10

Médigue C., Rouxel T., Vigier P., Henaut A., Danchin A. (1991). *Evidence of horizontal gene transfer in Escherichia coli speciation.* J. Mol. Biol. **222**, 851-856

Mooers A.O., Holmes E.C. (2000). *The evolution of base composition and phylogenetic inference*. Trends Ecol. Evol. **15**, 365-369

Mrázek J., Karlin S. (1999). *Detecting alien genes in bacterial genomes.* Ann. New York Acad. Sci. **870**, 314-329

Murzin A.G., Brenner S.E., Hubbard T., Chothia C. (1995). *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J. Mol. Biol. **247**, 536-540

——

Nakamura Y., Gojobori T., Ikemura T. (2000). *Codon usage tabulated from international DNA sequence databases: status for the year 2000*. Nucl. Acids Res. **28**, 292

Nelson K.E., Clayton R.A., Gill S.R., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Nelson W.C., Ketchum K.A., McDonald L., Utterback T.R., Malek J.A., Linher K.D., Garrett M.M., Stewart A.M., Cotton M.D., Pratt M.S., Phillips C.A., Richardson D., Heidelberg J., Sutton G.G., Fleischmann R.D., Eisen J.A., White O., Salzberg S.L., Smith H.O., Venter J.C., Fraser C.M. (1999). *Evidence for lateral gene transfer between archaea and bacteria from genome sequence of Thermotoga maritima.* Nature **399**, 323-329

Nesbo C.L., L'Haridon S., Stetter K.O., Doolittle W.F. (2001). *Phylogenetic analyses of two "archaeal" genes in thermotoga maritima reveal multiple transfers between archaea and bacteria.* Mol. Biol. Evol. **18**, 362-375

Nunes L.R., Rosato Y.B., Muto N.H., Yanai G.M., da Silva V.S., Leite D.B., Gonçalves E.R., de Souza A.A., Coletta-Filho H.D., Machado M.A., Lopes S.A., de Oliveira R.C. (2002). *Microarray analyses of Xylella fastidiosa provide evidence of coordinated transcription control of laterally transferred elements*. Genome Res. **13**, 570-578

——

Ochman H., Lawrence J.G., Groisman E.A. (2000). *Lateral gene transfer and the nature of bacterial innovation*. Nature **405**, 299-304

Ohno S., Ohno M. (1986). *The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition*. Immunogenetics **24**, 71-78

Ohno H., Sakai H., Washio T., Tomita M. (2001). *Preferential usage of some minor codons in bacteria*. Gene **276**, 107-115

Olendzenski L., Liu L., Zhaxybayeva O., Murphey R., Shin D.G., Gogarten J.P. (2000). *Horizontal transfer of archaeal genes into the deinococcaceae: detection by molecular and computer-based approaches.* J. Mol. Evol. **51**, 587-599

Olsen G.J., Woese C.R., Overbeek R. (1994). *The winds of (evolutionary) change: breathing new life into microbiology*. J. Bacteriol. **176**, 1-6

Olson W.K., Gorin A.A., Lu X.J., Hock L.M., Zhurkin V.B. (1998). *DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.* Proc. Natl. Acad. Sci. USA. **95**, 11163-11168

Ozanne C.M., Anhuf D., Boulter S.L., Keller M., Kitching R.L., Korner C., Meinzer F.C., Mitchell A.W., Nakashizuka T., Dias P.L., Stork N.E., Wright S.J., Yoshimura M. (2003). *Biodiversity meets the atmosphere: a global view of forest canopies*. Science **301**, 183-186

——

Pazin M.J., Kadonaga J.T. (1997). *SWI2/SNF2 and related proteins: ATP-driven motors that disrupt protein-DNA interactions?* Cell **88**, 663-673

Pedersen A.G., Baldi P., Chauvin Y., Brunak S. (1998). *DNA structure in human RNA polymerase II promoters*. J. Mol. Biol. **281**, 663-673

Peleg O., Brunak S., Trifonov E.N., Nevo E., Bolshoy A. (2002). *RNA secondary structure and sequence conservation in C1 region of human immunodeficiency virus type 1 env gene*. AIDS Res. Hum. Retroviruses **18**, 867-878

Peleg O., Trifonov E.N., Bolshoy A. (2003). *Hidden messages in the nef gene of human immunodeficiency virus type 1 suggest a novel RNA secondary structure*. Nucl. Acids Res. **31**, 4192-4200

Percudani R., Pavesi A., Ottonello S. (1997). *Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae*. J. Mol. Biol. **268**, 322-330

Perrière G. and Thioulouse J. (2002). *Use and misuse of correspondence analysis in codon usage studies.* Nucl. Acids Res. **30**, 4548-4555

Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Podkolodny N.L., Savinkova L.K., Kolchanov N.A., Overton G.C. (1999). *Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins*. Bioinformatics **15**, 687-703

——

Ragan M.A. (2001). *On surrogate methods for detecting lateral gene transfer.* FEMS Microbiol. Letters **201**, 187-191

Ratner L., Haseltine W., Patarca R., Livak K.J., Starcich B., Josephs S.F., Doran E.R., Rafalski J.A., Whitehorn E.A., Baumeister K., Ivanoff L., Petteway S.R. Jr., Pearson M.L., Lautenberger J.A., Papas T.S., Ghrayeb J., Chang N.T., Gallo R.C., Wong-Staal F. (1985). *Complete nucleotide sequence of the AIDS virus, HTLV-III.* Nature **313**, 277-284

Ravatn R., Studer S., Zehnder A.J., van der Meer J.R. (1998). *Int-B13, an unusual site-specific recombinase of the bacteriophage P4 integrase family, is responsible for chromosomal insertion of the 105-kilobase clc element of Pseudomonas sp. Strain B13.* J. Bacteriol. **180**, 5505-5514

Rice P., Longden I., Bleasby A. (2000). *EMBOSS: The European Molecular Biology Open Software Suite*. Trends in Genetics **16**, 276-277

Rowe G.W., Szabo V.L., Trainor L.E.H. (1984). *Cluster analysis of genes in codon space*. J. Mol. Evol. **20**, 167-174

Rowe G.W. (1985). *A three-dimensional representation for base composition of protein-coding DNA sequences*. J. Theor. Biol. **112**, 433-444

——

Saito A., Pietromonaco S., Loo A.K., Farquhar M.G. (1994). *Complete cloning and sequencing of rat gp330/"megalin," a distinctive member of the low density lipoprotein receptor gene family*. Proc. Natl. Acad. Sci. USA **91**, 9725-9729

Saitou N., Nei M. (1987). *The neighbor-joining method: a new method for reconstructing phylogenetic trees*. Mol. Biol. Evol. **4**, 406-425

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

Sanger F., Coulson A.R., Friedmann T., Air G.M., Barrell B.G., Brown N.L., Fiddes J.C., Hutchison C.A. 3rd, Slocombe P.M., Smith M. (1978). *The nucleotide sequence of bacteriophage phiX174*. J. Mol. Biol. **125**, 225-246

Schaap T. (1971). *Dual information in DNA and the evolution of the genetic code*. J. Theor. Biol. **32**, 293-298

Scipioni A., Anselmi C., Zuccheri G., Samori B., De Santis P. (2002). *Sequence-Dependent DNA Curvature and Flexibility from Scanning Force Microscopy Images*. Biophys. J. **83**, 2408-2418

Shan J., Moran-Jones K., Munro T.P., Kidd G.J., Winzor D.J., Hoek K.S., Smith R. (2000). *Binding of an RNA trafficking response element to heterogeneous nuclear ribonucleoproteins A1 and A2*. J. Biol. Chem. **275**, 38286-38295

Sharp P.M., Rogers M.S., McConnell D.J. (1985). *Selection pressures on codon usage in the complete genome of bacteriophage T7*. J. Mol. Evol. **21**, 150-160

Sharp P.M. (1986). *Molecular evolution of bacteriophages: evidence of selection against the recognition sites of host restriction enzymes*. Mol. Biol. Evol. **3**, 75-83

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 Res. **15**, 1281-1295

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, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster, and Homo sapiens; a review of the considerable within-species diversity*. Nucleic Acids Res. **16**, 8207-8211

Sharp P.M., Devine K.M. (1989). *Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons*. Nucl. Acids Res. **17**, 5029-5039

Sharp P.M., Stenico M., Peden J.F., Lloyd A.T. (1993) *Codon usage: mutational bias, translational selection, or both?* Biochem. Soc. Trans. **21**, 835-841

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

Skiena S.S. (2001). *Designing better phages*. Bioinformatics **17** Suppl. 1, S253-S261

Smit A.F. (1996). *The origin of interspersed repeats in the human genome*. Curr. Opin. Genet. Dev. **6**, 743-748

Smith T.G., Marks W.B., Lange G.D., Sheriff W.H., Neale E.A. (1989). *A fractal analysis of cell images*. J. Neurosci. Meth. **27**, 173-180

Smith M.W., Feng D.F., Doolittle R.F. (1992). *Evolution by acquisition: the case for horizontal gene transfers.* Trends Biochem. Sci. **17**, 489-493

Smith N.G.C., Eyre-Walker A. (2001). *Synonymous codon bias is not caused by mutation bias in G+C rich genes in humans.* Mol. Biol. Evol. **18**, 982-986

Stoesser G., Baker W., van den Broek A., Garcia-Pastor M., Kanz C., Kulikova T., Leinonen R., Lin Q., Lombard V., Lopez R., Mancuso R., Nardone F., Stoehr P., Tuli M., Tzouvara K. and Vaughan R. (2003). *The EMBL Nucleotide Sequence Database: major new developments.* Nucl. Acids Res. **31**, 17-22

——

Takahashi M. (1989). *A fractal model of chromosomes and chromosomal DNA replication*. J. Theor. Biol. **141**, 117-136

The art of DNA. Economist 26th April 2000, 81-83 (No author. Published weekly)

The FlyBase Consortium (2002). *The FlyBase database of the Drosophila genome projects and community literature.* Nucl. Acids Res. **30**, 106-108

Thompson J.D., Higgins D.G., and Gibson T.J. (1994). *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice.* Nucl. Acids Res. **22**, 4673-4680

Trifonov E.N. (1989). *The multiple codes of nucleotide sequences*. Bull. Math. Biol. **51**, 417-432

Tsonis A.A., Tsonis P.A. (1987). *Fractals: a new look at biological shape and patterning*. Persp. Biol. Med. **30**, 355-361

Tsukiyama T., Wu C. (1997). *Chromatin remodeling and transcription.* Curr. Opin. Genet Dev. 7, 182-191

——

Urrutia A.O., Hurst L.D. (2001). *Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection*. Genetics **159**, 1191-1199

——

Vemula M., Kandasamy P., Oh C.S., Chellappa R., Martin C.E. (2003). *Maintenance and regulation of mRNA stability of the saccharomyces cerevisiae OLE1 gene requires multiple elements within the transcript that act through translation-independent mechanisms*. J. Biol. Chem. **278**, 45269-45279

Vicsek T. (1989). *Fractal growth phenomena*. World Scientific, Singapore

——

Wells D., Bains W., Kedes L. (1986). *Codon usage in histone gene families of higher eukaryotes reflects functional rather than phylogenetic relationships*. J. Mol. Evol. **23**, 224-241

Wolf Y.I., Rogozin I.B., Kondrashov A.S, Koonin E.V. (2001). *Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context.* Genome Res. **11**, 356-372

Wolffe A.P. (1994). *Nucleosome positioning and modification: chromatin structures that potentiate transcription.* Trends Biochem. Sci. **19**, 240-244

Wong P., Bergeron R. (1997). *30 Years of Multidimensional Multivariate Visualization* in *Scientific Visualization - Overviews, Methodologies and Techniques*. IEEE Computer Society Press, pp. 3-33. Los Alamitos, CA

——

Xiao Y., Chen R., Shen R., Sun J., Xu J. (1995). *Fractal dimension of exon and intron sequences*. J. Theor. Biol. **175**, 23-26

Xie G., Bonner C.A., Brettin T., Gottardo R., Keyhani N.O., Jensen R.A. (2003). *Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in Xylella species and in heterocystous cyanobacteria.* Genome biology **4**, R14

——

Zhang C., DeLisi C. (1998). *Estimating the number of protein folds*. J. Mol. Biol. **284**, 1301-1305

Zuccheri G., Scipioni A., Cavaliere V., Gargiulo G., De Santis P., Samori B. (2001) *Mapping the intrinsic curvature and flexibility along the DNA chain.* Proc. Natl. Acad. Sci. USA. **98**, 3074-3079

zur Megede J., Chen M.C., Doe B., Schaefer M., Greer C.E., Selby M., Otten G.R., Barnett S.W. (2000). *Increased expression and immunogenicity of sequence-modified human immunodeficiency virus type 1 gag gene*. J. Virol. 2000 **74**, 2628-2635