Compositional features of eukaryotic genomes for checking predicted genes

Stéphane Cruveiller, Kamel Jabbari, Oliver Clay and Giorgio Bernardi

Date received (in revised form): 22nd January 2003

Stéphane Cruveiller
is a postdoctoral fellow at Genoscope (French National Sequencing Center, Evry) working on (re)annotation of prokaryotic genomes. His current research focus is on developing methods and criteria for systematically detecting and characterising horizontal transfer events.

Kamel Jabbari
is working in Jacques Monod Institut (Paris). His research experience is in structural and evolutionary genomics. His work focuses on genomic methylation, CpG islands and transposons, on functional diversification in gene families, and compositional properties of protein-coding sequences.

Oliver Clay
works in computational genome research. His interests include comparative analyses using chromosomal and coding sequences, ultracentrifuge experiments and gene expression data.

Giorgio Bernardi
is head of the Molecular Evolution Laboratory at the Stazione Zoologica Anton Dohrn (Naples, Italy). His research interests have been centred for many years on genome organisation and evolution.

Keywords: *base composition, gene prediction, correlations in DNA*

Giorgio Bernardi,
Laboratorio di Evoluzione Molecolare,
Stazione Zoologica Anton Dohrn,
Villa Comunale,
80121 Napoli, Italy

Fax: +39 081 764 1355
E-mail: bernardi@alpha.szn.it

Abstract

Gene prediction relies on the identification of characteristic features of coding sequences that distinguish them from non-coding DNA. The recent large-scale sequencing of entire genomes from higher eukaryotes, in conjunction with currently used gene prediction algorithms, has provided an abundance of putative genes that can now be analysed for their compositional properties. Strong, systematic differences still exist, in several species, between the compositional properties of sets of *ex novo* predicted genes and genes that have been experimentally detected and/or verified. This is particularly evident in the estimated gene set (>45,000 genes) of the recently sequenced rice genome, where roughly half the predicted genes are compositionally unusual and have no known orthologues in the dicot *Arabidopsis*. In a few cases such differences might suggest a bias in experimental gene-finding protocols, but the quasi-random nature of the compositionally aberrant predicted genes is a strong indication that many, if not most, of them are false positives. It therefore appears that some important features of coding regions have not yet been taken into account in existing gene prediction programs. Statistical base compositional properties of curated gene data sets from vertebrates, which we briefly review here, should therefore provide a useful benchmark for fine-tuning probabilistic gene models and model parameters that are currently in use.

INTRODUCTION

Vertebrate genomes are characterised by the mosaic organisation of their base composition, at different scales. Historically, the first indication of such large-scale mosaicism was the phenomenon of chromosomal (eg Giemsa/Reverse) banding. The bands that are obtained by standard staining techniques are now known to correlate with local GC level (ie with the molar ratio of guanine + cytosine in the DNA) and with replication timing: the dark-staining bands have lower average GC contents than the bands that flank them,[1,2] and replicate later than the light bands.[3]

The first rigorous evidence for large-scale compositional mosaicism came from analytical ultracentrifugation of mammalian and avian DNAs in density gradients. Such experimental analyses allowed an early demonstration of a strong large-scale heterogeneity in the GC of their genomes, quite apart from the contributions of highly repetitive satellite DNAs.[4] Subsequent ultracentrifugation analyses, performed at different molecular weights and for different species, provided conclusive evidence that the unique DNA of mammalian and avian genomes is organised into long, relatively homogeneous chromosomal regions that span more than 300 kb on average, but can often extend much further[5,6] (recent discussions of the methodology are given in Clay *et al.*[7] and Pavlícek *et al.*[8]). To emphasise the mosaic nature of these genomes, in which GC-poorer regions alternate with distinctly GC-richer regions along the chromosomes, the fairly homogeneous regions were called isochores.[9] Among the biological properties that are now known to correspond (in some cases sharply) with the DNA of GC-rich isochores in mammals, we mention here the earlier replication timing, the higher gene densities, the fewer and shorter introns, the higher concentration of the GC-rich CpG islands, and the different and larger chromosomal territories in interphase that are occupied by such GC-rich DNA (reviewed in Bernardi[10] and Saccone *et al*.[11]).

Codon positions

Correlations

The conservation, among mammals and birds, of many, if not most, of the broad compositional properties that will be discussed here, has been confirmed by experimental studies involving many taxa. The techniques used include ultracentrifugation experiments in density gradients and inter-taxon hybridisation studies (so-called zooblots).[12–14] Among the eutherians that have been well characterised at the genomic level, only some myomorph rodents, including murids, show an obvious departure from the general compositional patterns of other warm-blooded vertebrates. In particular, GC contrasts among different regions of a chromosome, and related contrasts such as those between CpG islands and inter-island DNA (including their respective methylation levels), are less pronounced in mouse, rat and other murids than in human or chicken (see Bernardi[15] and Douady *et al*.[16] and references therein). Birds differ from mammals in having even more pronounced large-scale heterogeneity in their GC content, and in having microchromosomes that contain much of their GC-richest, CpG-richest DNA (see Andreozzi *et al*.[17] and references therein).

COMPOSITIONAL CORRELATIONS INVOLVING GENES

Coding DNA

Within the isochores of vertebrates, there is again mosaicism of GC levels, although at a much smaller scale, namely at the scale of genes, exons and introns.[18–22] For example, in GC-rich isochores, the GC levels of coding exons rise above the background of intergenic DNA, as do the CpG dinucleotide frequencies and CpG observed/expected ratios.[23] Within the coding exons, the GC levels in third codon positions (GC_3) are, in turn, typically higher than in first codon positions (GC_1), and much higher than in second codon positions (GC_2). Conversely, genes in very GC-poor isochores ($GC < 40\%$) tend to contain more, and longer, introns than those in GC-rich isochores,[24] and are GC-poor. In such GC-poor genes, GC_3 levels are typically lower than GC levels in first and second codon positions. Thus GC_3 is, at least statistically, a sensitive monitor of the mean GC of 10–100 kb regions in which the genes are embedded, being substantially higher than the ambient intergenic GC in GC-rich isochores, yet similar to or lower than this intergenic GC in GC-poor isochores. Whereas second positions are strongly constrained by the encoded amino acids, third codon positions are largely free of such constraints and generally reflect the base composition of the isochore in which they are embedded. Thus, a steep, yet strong, correlation holds between the GC_3 levels of genes and the GC level of the DNA surrounding the genes.

Significant correlations exist also among the other characteristic GC levels within or surrounding genes (three codon positions, exons/coding sequence (CDS), introns, 5′ flanks, 3′ flanks). This fact corresponds to the observation that the corresponding bivariate distributions often have an approximately linear shape and are well characterised by their major axis (also called orthogonal regression line, or principal axis). In passing, it should be mentioned that traditional linear regression lines (such as are used to describe unilateral dependence relationships) do not provide a satisfactory characterisation, since they do not follow the points when scatterplots are characterised by steep slopes, but instead systematically slice the scatterplots at a lower angle.[21,25–27] In addition, such traditional regression lines are not invariant when 'dependent' and 'independent' variables are swapped.

The major axis equations, which are

equations that characterise a genome,[28] have been obtained and studied for human[20,21,29] and chicken,[22] and some examples are listed in Table 1. Relationships between codon positions require only cDNA (mRNA) sequences, and could therefore be obtained for a wide range of eukaryotic and prokaryotic species, either intra- and/or intergenomically.[19,29,35–38] The practical utility of these relations comes from their unusually wide conservation, especially for the relation between the third and second codon positions. Since the line that characterises this latter relation is far from the diagonal line that would be expected for random intergenic sequences ($GC_2 = GC_3$), it can be used to check the quality or plausibility of gene predictions in previously uncharacterised species.

GC_2 and GC_3

Table 1 lists the equation linking GC_2 and GC_3 for human, chicken and *Escherichia coli*, illustrating good conservation despite a huge taxonomic distance, and despite a relatively narrow distribution of GC_3 values in *E. coli*.

As a comparison, the equation is listed also for a sample of apparently *ex novo* or *ab initio* predicted 'human' genes. Such results for predicted human genes are sometimes quite different from those for true human, chicken or *E. coli* genes. Since human and *E. coli* have long been represented by large databases of experimentally determined or verified coding sequences (cDNA), and since the compositional properties of these growing databases have remained essentially the same for almost a decade (as will be illustrated below), the non-redundant sequences used in Table 1 should be a faithful representation of true protein-coding genes. For predicted genes, on the other hand, the steeper orthogonal regression line and higher correlation coefficient (*R*) tend toward the expectation for randomly chosen intergenic sequences. Indeed, for intergenic DNA one would expect the slope to be essentially 1.

The relation between GC_3 and the GC of the DNA surrounding the genes has

Table 1: Equations of the human and chicken genomes, describing linear relations (major axis, ie orthogonal regression) between base compositions of characteristic parts of genes and/or flanking (intergenic) DNA

x	*y*	Retrieval date	Species	Equation	*R*	Method	Reference
$GC_{flanking}$	GC_3	1995	Human	$y = 2.92x - 74.3$	n.a. (0.9995)*	Experimental/sequences*	Zoubak *et al.*[30]
		1999	Human	$y = 4.09x - 120.37$	0.62	CDS in large (>50 kb) GenBank contigs	Jabbari and Bernardi[31]
		2002	Human	$y = 3.06x - 79.4$	0.64	RefSeq (refGene) + draft genome sequence†	Pavlicek *et al.*, in preparation
		1998	Chicken	$y = 2.64x - 64$	0.78	Genes with sequenced flanking DNA	Musto *et al.*[22]
GC_{1+2}	GC_3	1995	Human	$y = 5.64x - 215.3$	0.42	4,270 non-redundant GenBank CDS sequences	Clay *et al.*[21]
GC_3	GC_2	2002	Human	$x = 5.846y - 187.7$	0.32	10,218 non-redundant CDS sequences (RefSeq)	Pruitt and Maglott[32]
		1998	Chicken	$x = 5.98y - 185$	0.36	1,037 non-redundant GenBank CDS sequences	Musto *et al.*[22]
		1997	*E. coli*	$x = 5.225y - 156.6$	0.23	4,286 CDS sequences	Lawrence and Ochman[33]
		2002	'Human'	$y = 1.95x - 33.2$	0.62	588 'not_experimental' GenBank sequences	GenBank (28 November 2002)
$GC_{flanking}$	GC_{CDS}	2002	Human	$y = 1.27x - 4.76$	0.65	RefSeq (refGene) + draft genome sequence†‡	Pavlícek *et al.*, in preparation
GC_{intron}	GC_{CDS}	1995	Human	$y = 0.83x + 14.2$	0.78	Genes with sequenced introns	Clay *et al.*[21]

*Indirect calculation, by matching four Gaussian components of *x* and *y* distributions (*x*: experimental CsCl profile; *y*: $N = 4,270$); confirmed by direct regression for smaller sets of genes with available flanking sequences or hybridisation/ultracentrifugation data, $R \sim 0.7–0.8$.

†Flanking regions of 100 kb were used after removing repetitive DNA ($N = 14,652$ coding sequences).[2, 34]

‡Similar results were found earlier for fewer sequences/fragments.[21]

also proved very useful. Already a decade ago, this linear relation was employed to estimate the gene density distribution in the human genome.[27,30] The equation, given in Table 1, was applied to the distribution of genes' GC_3 levels, in order to obtain the distribution of GC levels of the DNA that is expected to surround the genes. The distribution of the GC levels of long DNA fragments was obtained by ultracentrifugation in caesium chloride density gradients, and a simple division of the two distributions yielded the gene density curve. This curve rises steeply: in the GC-poor regions, DNA is abundant and genes are scarce, whereas in the GC-rich regions there is little DNA, so that genes are crowded. In the GC-richest regions, genes are found at densities that are 15–20 times higher than in the GC-poorest regions. This ratio was recently confirmed at the sequence level, using the draft genome sequence[2] (see Bernardi[39] for a discussion).

Gene distribution

Other linear relationships that can be used to recognise or verify genes and their exon–intron structures include those that exist between the GC levels of coding sequences (GC_{CDS}) or introns (GC_{intron}) and the surrounding DNA ($GC_{flanking}$). Interestingly, CpG island genes (in which CpGs, and therefore GC levels, are elevated at the 5′ or 3′ end of the gene) do not appear to strongly influence the lines describing these relationships, at least in human.[21,30]

The observed linearities often extend over wide ranges, which span most of the relevant GC values that are found in protein-coding sequences. This fact is noteworthy, because such linear relationships will obviously no longer hold at 0 per cent GC and 100 per cent GC (excepting the trivial identity $y = x$).

Figure 1 shows the scatterplots of GC_2 and GC_3 for coding sequences from human (left; 10,218 sequences) and *E. coli* (right; 4,286 sequences), corresponding to data sets listed in Table 1. The orthogonal regression lines that characterise them are shown, together with the main diagonal of slope 1 ($GC_2 = GC_3$) as a comparison.

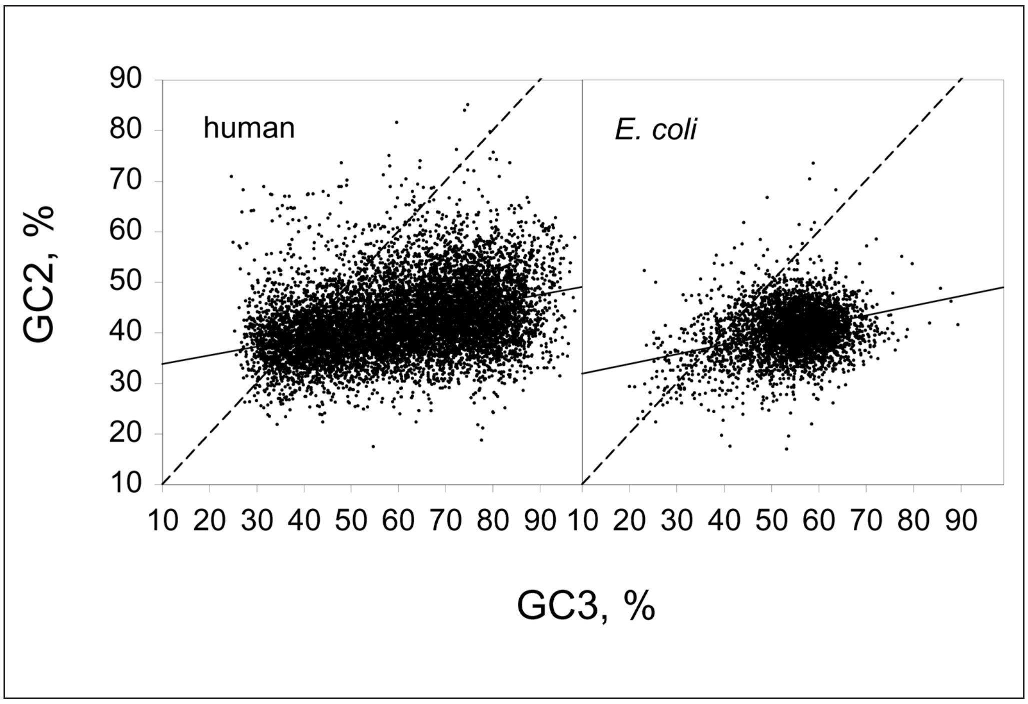

Figure 1: Scatterplots of GC_2 versus GC_3 for non-redundant, representative collections of coding sequences for human (left, 10,218 non-redundant RefSeq[32] sequences) and *E. coli* (right, 4,286 sequences[33]). In each scatterplot, the main diagonal and orthogonal regression line (major axis; equations listed in Table 1) are shown

In GC-rich coding DNA, a striking difference is seen between the GC levels in different codon positions.

COMPOSITIONAL DISTRIBUTIONS

Bivariate frequency distributions (ie probability density functions), such as the joint distribution of GC_2 and GC_3, can be represented in several ways. One way is to show a scatterplot, and the major axis that best characterises it, as in Figure 1. When there are many sequences, important information is however obscured, especially in the dense region around the major axis, and even the shape of the modal crest cannot be discerned. Contour plots or three-dimensional plots can reveal such information (Clay *et al.*[21] show an example). Similarly, lateral views or, alternatively, thin transects of the 3D landscape can be easily obtained and plotted: they are simply one-dimensional histograms. A histogram can, for example, be plotted for a thin transect (slice) along the major axis or modal crest of such a landscape, and the views from (or projections onto) the two axes correspond to the histograms of the variables (as is illustrated by Figure 5 in Rijsdijk and Sham[40]).

The distribution of genic GC_2 and GC_3 levels subtends only a small angle with the GC_3 axis. This fact has two simple consequences. The first consequence is that departures from the expected clustering along the major axis are well captured by GC_2 histograms. The second consequence is that the extent to which the genes spread out along the major axis is well captured by GC_3 histograms. Both of these consequences can be used to recognise sets of anomalous genes, which may represent incorrect predictions.

Anomalous genes

Rice genome

Figure 2 shows a scatterplot of GC_2 and GC_3 levels for the predicted gene set of chromosome 1 in rice, according to a very recent analysis by Sasaki *et al.*[41] As a guide, two lines are shown: the main diagonal $x = y$, and the line $x = 6y - 2$ (ie $GC_3 = 6GC_2 - 200$ per cent), which is close to the major axis in other species (see Table 1). It can be seen that, although many of the genes follow the widely conserved major axis for the relation between GC_2 and GC_3, a large number of genes depart from it and follow closely the main diagonal, as would be expected for incorrectly predicted genes that are in fact intergenic sequences.

Figure 2 also shows, in projection along the GC_2 axis, the GC_3 distribution for a different predicted gene set of 53,398 sequences from the entire genome of the *indica* subspecies of rice obtained by Yu *et al.*[42] It is shown as a sum of two components: the distribution for genes that have homologues in *Arabidopsis* (a distribution which one would expect if the major axis is conserved in rice), and the distribution for predicted genes that have no homologue in *Arabidopsis*. Many of these latter sequences have, furthermore, no homologous sequences, from any species, in currently accessible databases.[42] Such sequences are seen to have a much wider GC_2 distribution, extending effortlessly up to high GC_2 levels that are found only very rarely in other organisms.

Although two different sets of rice sequences (histogram and scatterplot) are shown in Figure 2, both of them apparently used a similar approach (hidden Markov models) and similar training sets, and both of them show similar compositional anomalies for a large proportion of the predicted genes. In fact, as both Yu *et al.*[42] and Sasaki *et al.*[41] have pointed out, over 50 per cent of the sequences in their predicted gene sets (50.6 per cent in the first set, 53.2 per cent in the second set) have apparently no obvious homologue in *Arabidopsis*. Possible reasons for this are discussed below.

It has been suggested that the compositionally anomalous genes observed in rice may be a characteristic of some cereal plants that is not shared by other well-characterised taxa, including other angiosperms such as the dicot *Arabidopsis thaliana*.[42] Such a departure

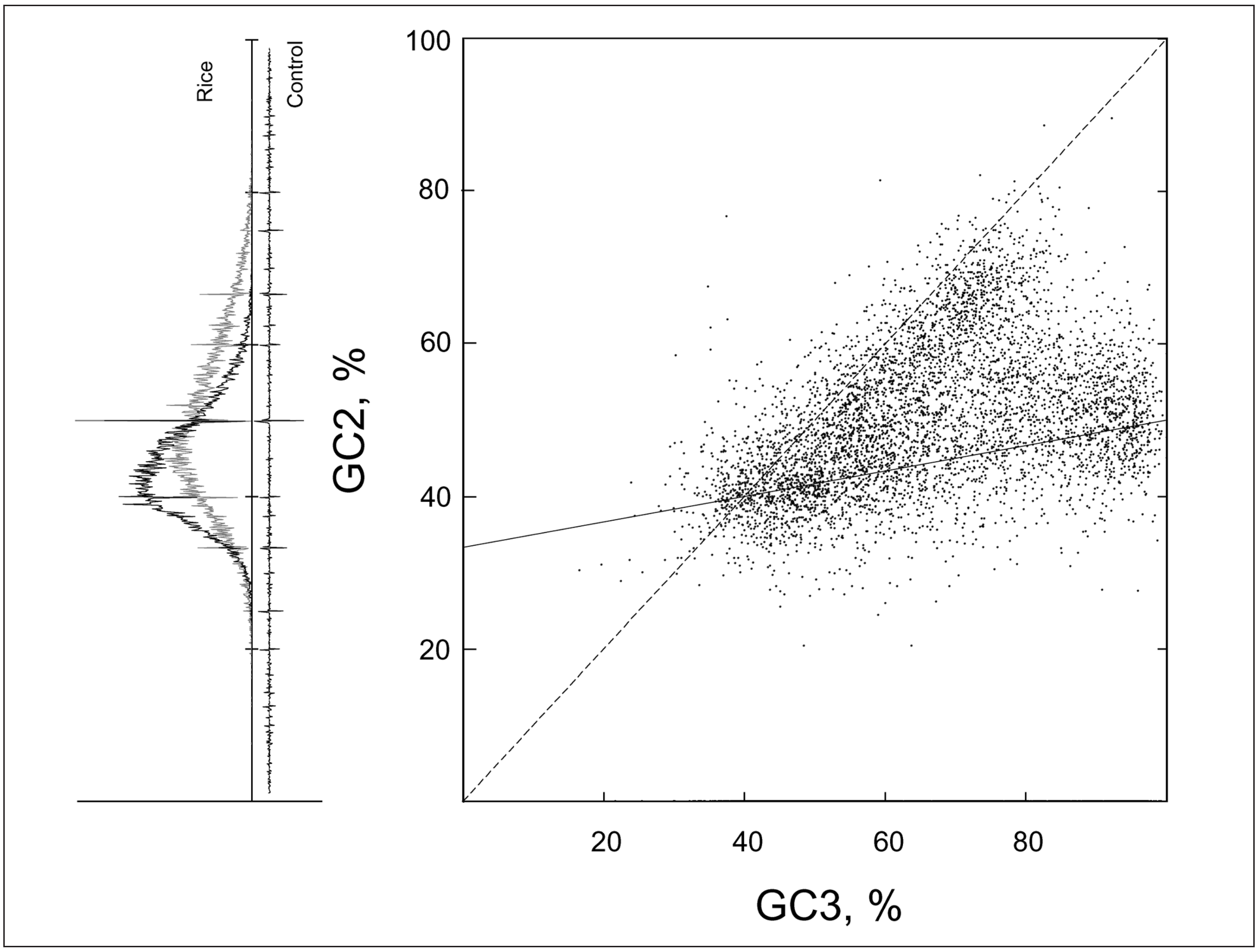

Figure 2: Scatterplot of GC_2 versus GC_3 for a predicted set of protein-coding sequences of rice chromosome 1 obtained by Sasaki *et al.*[41] The main diagonal (which also characterises randomly chosen DNA) and the expected major axis (approximate; see text) are shown. Along the GC_2 axis on the right, two corresponding histograms of GC_2 are shown, for the predicted set of 53,398 rice coding sequences obtained by Yu *et al.*:[42] those that have homologous genes in *Arabidopsis thaliana* (black histogram), and those that do not (grey histogram). Bin size is 0.1 per cent GC_2. A control histogram is also shown for a hypothetical set of sequences having the same lengths but a uniform GC_2 distribution, in order to highlight the inevitable peaks. For example, if the 1,001 possible GC_2 values (<100 per cent) of a 1001 bp sequence are distributed in 1,000 bins, one of the bins will always contain two values, ie an elevated frequency. More generally, the expected frequency in the *i*th of *b* equal bins dividing the range $0 \leqslant x < 1$ (ie $0\% \leqslant x < 100\%$) will be the sum, over all lengths 1, of the terms $[\mathrm{f}(l)/l]\infty\{\mathrm{floor}(il/b) - \mathrm{ceiling}[(i-1)l/b] + r_{il,b}\}$ where f(*l*) denotes the number of sequences that have length *l*. Here, floor (ceiling) denotes the highest (lowest) integer not higher (lower) than the expression enclosed in the parentheses, and the 'rest indicator' $r_{il,b}$ is 1 unless *il* divides *b* exactly, in which case it is 0

Gene predictions

from widely conserved compositional features (such as the one illustrated in Figure 1) would, in all likelihood, require cereal plants to have undergone a massive change in elementary genomic processes such as transcription, transcriptional regulation, or translation since the monocot–dicot divergence. There are no known independent findings that support such a hypothesis. It is possible, therefore, that many, if not most, of the obviously aberrant sequences in Figure 2 are simply incorrectly predicted as coding DNA, and are in fact non-coding sequences.[43]

Figure 3 shows GC_3 distributions that can be used for highlighting potentially erroneous gene predictions. Whereas the GC_2 distributions of 'false positive' sequences will tend to be wider and/or GC_2-richer than confirmed or correctly

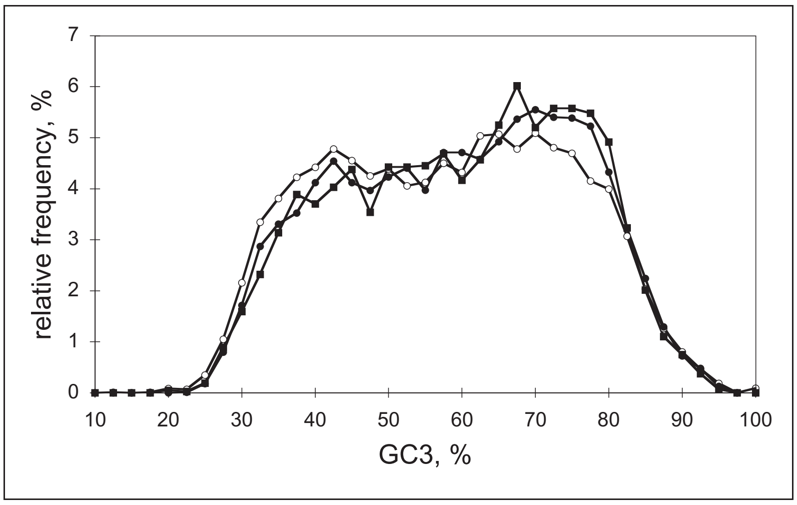

Figure 3: Histograms of GC levels of third codon positions (GC_3) in confirmed human coding sequences, retrieved from curated sequence databases over the past eight years (symbols connected by black lines; $N = 4,270$, 10,218 and 14,652 for the filled squares, filled circles and open circles, respectively), and, narrower histograms were found for *ab initio* predicted sequence sets. The confirmed GC_3 histograms are similar to each other and to earlier histograms.[27] Sources are given in Table 1. Bin size is 2.5 per cent GC_3, and histograms are normalised to 100 per cent.

recognised genes, their GC_3 distributions will tend to be narrower, at least in warm-blooded vertebrates.

Genome annotation

In Figure 3, it can be seen that the GC_3 distribution of available curated sequences of human genes has remained essentially unchanged during the past decade (symbols connected by black lines), in spite of the increasing number of sequences. Indeed, a very similar histogram of non-redundant sequenced genes was found 12 years ago.[27] The shape and the width of the human GC_3 distribution of confirmed genes can, therefore, be relied upon as robust.

Training sets

Ex novo (*ab initio*) predicted gene sets still often depart from this established GC_3 distribution of human genes. The shapes of such gene sets' GC_3 histograms approach that of the GC distribution of bulk DNA, which is narrower and has a mode around 40 per cent GC.

Although it cannot be ruled out that all experimental gene-finding procedures so far have been highly biased, ie that we are on the verge of discovering a new class of abundant genes that systematically eluded detection procedures for over a decade, we consider that high levels of contamination by non-coding DNA is a simpler, and quite reasonable, explanation for such discrepancies.

METHODOLOGICAL BLIND SPOTS IN MACHINE LEARNING AND THE 'SNOWBALL EFFECT'

Machine learning has been used in a variety of contexts during the past decade. An early example was the training of neural networks that learned to distinguish undersea mines from benign rocks, on the basis of their sonar echo patterns (discussed in Churchland[44]). Machine learning is also used, to differing extents, in predicting genes in DNA contigs that are produced by bulk sequencing projects. Probabilistic hidden Markov models, in which the parameters need to be fitted from information in a 'training' or 'learning' set, play a role in several gene prediction programs now employed in whole-genome sequencing and annotation projects, such as GENSCAN,[45] FGenesh[46] or RiceHMM.[47] Currently used gene prediction programs involve models in which only parameters with an obvious biological meaning need to be fitted and their role is clearly defined from the outset. This structure gives the models a distinct advantage over more abstract models: the results obtained after a fitting or 'learning' stage can often be formulated in traditional terms as easily interpreted if-then-else rules (or 'sentences or propositions, expressible in the first-order predicate calculus'[44]). A characteristic of the machine learning framework persists, however: the need for a learning set or 'training set' consisting of *bona fide* coding sequences, such as can be obtained by checking the laboratory evidence for each gene in the set (as was done, for example, in Salzberg *et al.*[48]).

The sensitivity of machine learning protocols to contaminated training sets is well recognised. Yet present gene prediction programs are often trained on data sets that partly contain, in turn, sequences that were previously classified as coding by other prediction programs (cf. the discussion in Yu *et al.*[42]). Alternatively, they may give higher scores when a candidate gene is detected by another prediction program.[47] Such iterative training or lateral reinforcements could lead to a fatal 'snowball effect', in which incorrectly predicted genes could occupy an increasingly large proportion of the total putative gene set at successive steps of the iteration. One might expect such an effect if the recognised features are frequent in the genome's non-coding DNA.

'Snowball effect'

Learning protocols, used together with unbiased sets of *bona fide* coding sequences, can be a powerful aid when one is predicting genes. Surprises can occur, however, when the original training set of examples is small (as in the case of rice, where it was one or two orders of magnitude smaller than for human), or when two or more taxa with different compositional organisation are used during the tuning of parameters or models. In some cases, simply raising the stringency can eliminate a fair number of compositionally aberrant sequences that are likely to be false positives (see eg Jabbari and Bernardi[31]). In other cases, it may be necessary to try to understand in biological terms how the models' parameters are being learned, where the algorithm may have become trapped by artefacts or intergenomic differences, and how one might subsequently improve the models. As we have illustrated here, compositional approaches could be useful in such analyses.

Sequence assembly

Finally, it should be mentioned that compositional approaches can be useful also for visualising, comparing and checking draft genome assemblies (contig order and orientation). Colour-coded moving window plots,[8,49] with a standard coding scheme and appropriate scales (eg colour changes every 2.5 or 5 per cent GC for a window size of 100 kb), can give a concise overview of a chromosomal sequence, including its gaps. The coloured images of the compositional landmarks (GC-rich and GC-poor isochores) are easy to remember, so that disparities between two or more assemblies of a vertebrate chromosome can be quickly detected. In addition, such GC or isochore maps permit comparisons with studies from fluorescent *in situ* hybridisation (FISH)[50,51] (see also BACRC[52]). FISH experiments can provide independent compositional classification of the bands, and sequence assemblies could then be checked for consistency with the expected mosaic of GC-rich and GC-poor chromosomal bands. Long duplicated regions can confound some sequence assembly protocols (reviewed in Eichler[53]), and some candidates for such duplications could be revealed by the compositional studies. GC plots, at an appropriate scale, could therefore complement the usual criss-cross plots that connect orthologous regions of two species' chromosomes, or (as in Hattori and Taylor[54]) of two draft assemblies of a chromosome.

Acknowledgments

We thank Adam Pavlícek for helpful discussions and references and for communicating results prior to publication. We also thank an anonymous referee for valuable suggestions.

References

1. Saccone, S., Pavlícek, A., Federico, C. *et al.* (2001), 'Genes, isochores and bands in human chromosomes 21 and 22', *Chromosome Res.*, Vol. 9, pp. 533–539.

2. IHGSC, International Human Genome Sequencing Consortium (2001), 'Initial sequencing and analysis of the human genome', *Nature*, Vol. 409, pp. 860–921 (data from URL: http://genome.ucsc.edu).

3. Federico, C., Saccone, S. and Bernardi, G. (1998), 'The gene-richest bands of human chromosomes replicate at the onset of the S-phase', *Cytogenet. Cell Genet.*, Vol. 80, pp. 83–88.

4. Filipski, J., Thiery, J. P. and Bernardi, G. (1973), 'An analysis of the bovine genome by Cs_2SO_4–Ag^+ density gradient centrifugation', *J. Mol. Biol.*, Vol. 80, pp. 177–197.

5. Macaya, G., Thiery, J. P. and Bernardi, G. (1976), 'An approach to the organization of

eukaryotic genomes at a macromolecular level', *J. Mol. Biol.*, Vol. 108, pp. 237–254.

6. Thiery, J. P., Macaya, G. and Bernardi, G. (1976), 'An analysis of eukaryotic genomes by density gradient centrifugation', *J. Mol. Biol.*, Vol. 108, pp. 219–235.

7. Clay, O., Carels, N., Douady, C. *et al.* (2001), 'Compositional heterogeneity within and among isochores in mammalian genomes. I. CsCl and sequence analyses', *Gene*, Vol. 276, pp. 15–24.

8. Pavlícek, A., Paces, J., Clay, O. and Bernardi, G. (2002), 'A compact view of isochores in the draft human genome sequence', *FEBS Lett.*, Vol. 511, pp. 165–169.

9. Cuny, G., Soriano, P., Macaya, G. and Bernardi, G. (1981), 'The major components of the mouse and human genomes. I. Preparation, basic properties and compositional heterogeneity', *Eur. J. Biochem.*, Vol. 115, pp. 227–233.

10. Bernardi, G. (2000), 'Isochores and the evolutionary genomics of vertebrates', *Gene*, Vol. 241, pp. 3–17.

11. Saccone, S., Federico, C. and Bernardi, G. (2002), 'Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds', *Gene*, Vol. 300, pp. 169–178.

12. Sabeur, G., Macaya, G., Kadi, F. and Bernardi, G. (1993), 'The isochore patterns of mammalian genomes and their phylogenetic implications', *J. Mol. Evol.*, Vol. 37, pp. 93–108.

13. Kadi, F., Mouchiroud, D., Sabeur, G. and Bernardi, G. (1993), 'The compositional patterns of the avian genomes and their evolutionary implications', *J. Mol. Evol.*, Vol. 37, pp. 544–551.

14. Cacciò, S., Perani, P., Saccone, S. *et al.* (1994), 'Single-copy sequence homology among the GC-richest isochores of the genomes from warm-blooded vertebrates', *J. Mol. Evol.*, Vol. 39, pp. 331–339.

15. Bernardi, G. (2000), 'The compositional evolution of vertebrate genomes', *Gene*, Vol. 259, pp. 31–43.

16. Douady, C., Carels, N., Clay, O. *et al.* (2000), 'Diversity and phylogenetic implications of CsCl profiles from rodent DNAs', *Mol. Phylogenet. Evol.*, Vol. 17, pp. 219–230.

17. Andreozzi, L., Federico, C., Motta, S. *et al.* (2001), 'Compositional mapping of chicken chromosomes and identification of the gene-richest regions', *Chromosome Res.*, Vol. 9, pp. 521–532.

18. Bernardi, G., Olofsson, B., Filipski, J. *et al.* (1985), 'The mosaic genome of warm-blooded vertebrates', *Science*, Vol. 228, pp. 953–958.

19. Bernardi, G. and Bernardi, G. (1986), 'Compositional constraints and genome evolution', *J. Mol. Evol.*, Vol. 24, pp. 1–11.

20. Aïssani, B., D'Onofrio, G., Mouchiroud, D. *et al.* (1991), 'The compositional properties of human genes', *J. Mol. Evol.*, Vol. 32, pp. 497–503.

21. Clay, O., Cacciò, S., Zoubak, S. *et al.* (1996), 'Human coding and noncoding DNA: compositional correlations', *Mol. Phylogenet. Evol.*, Vol. 5, pp. 2–12.

22. Musto, H., Romero, H., Zavala, A. and Bernardi, G. (1999), 'Compositional correlations in the chicken genome', *J. Mol. Evol.*, Vol. 49, pp. 325–329.

23. Jabbari, K. and Bernardi, G. (1998), 'CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families', *Gene*, Vol. 224, pp. 123–127.

24. Duret, L., Mouchiroud, D. and Gautier, C. (1995), 'Statistical analysis of vertebrate sequences reveal that long genes are scarce in GC-rich isochores', *J. Mol. Evol.*, Vol. 40, pp. 308–317.

25. Jolicoeur, P. (1990), 'Bivariate allometry: Interval estimation of the slopes of the ordinary and standardized normal axes and structural relationship', *J. Theor. Biol.*, Vol. 144, pp. 275–285.

26. Harvey, P. H. and Pagel, M. D. (1991), 'The Comparative Method in Evolutionary Biology', Oxford University Press, Oxford.

27. Mouchiroud, D., D'Onofrio, G., Aïssani, B. *et al.* (1991), 'The distribution of genes in the human genome', *Gene*, Vol. 100, pp. 181–187.

28. Bernardi, G. (1995), 'The human genome: organization and evolutionary history', *Annu. Rev. Genet.*, Vol. 29, pp. 445–476.

29. D'Onofrio, G., Jabbari, K., Musto, H. and Bernardi, G. (1999), 'The correlation of protein hydropathy with the base composition of coding sequences', *Gene*, Vol. 238, pp. 3–14.

30. Zoubak, S., Clay, O. and Bernardi, G. (1996), 'The gene distribution of the human genome', *Gene*, Vol. 174, pp. 95–102.

31. Jabbari, K. and Bernardi, G. (2000), 'The distribution of genes in the *Drosophila* genome', *Gene*, Vol. 247, pp. 287–292.

32. Pruitt, K. D. and Maglott, D. R. (2001), 'RefSeq and LocusLink: NCBI gene-centered resources', *Nucleic Acids Res.*, Vol. 29, pp. 137–140.

33. Lawrence, J. G. and Ochman, H. (1998), 'Molecular archaeology of the *Escherichia coli* genome', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 9413–9417 (data from URL: ftp://ftp.pitt.edu/dept/biology/lawrence/eco.txt).

34. URL: http://genome.ucsc.edu/goldenPath/gbdDescriptions.html

35. Sueoka, N. (1988), 'Directional mutation pressure and neutral molecular evolution', *Proc. Natl Acad. Sci. USA*, Vol. 85, pp. 2653–2657.

36. D'Onofrio, G., Mouchiroud, D., Aïssani, B. *et al.* (1991), 'Correlations between the compositional properties of human genes, codon usage and amino acid composition of proteins', *J. Mol. Evol.*, Vol. 32, pp. 504–510.

37. D'Onofrio, G. and Bernardi, G. (1992), 'A universal compositional correlation among codon positions', *Gene*, Vol. 110, pp. 81–88.

38. Sueoka, N. (1992), 'Directional mutation pressure, selective constraints, and genetic equilibria', *J. Mol. Evol.*, Vol. 34, pp. 95–114.

39. Bernardi, G. (2001), 'Misunderstandings about isochores. Part I', *Gene*, Vol. 276, pp. 3–13.

40. Rijsdijk, F. V. and Sham, P.C. (2002), 'Analytic approaches to twin data using structural equation models', *Brief. Bioinformatics*, Vol. 3, pp. 119–133.

41. Sasaki, T., Matsumoto, T., Yamamoto, K. *et al.* (2002), 'The genome sequence and structure of rice chromosome 1', *Nature*, Vol. 420, pp. 312–316 (data from URL: http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/status.pl/

42. Yu, J., Hu, S., Wang, J. *et al.* (2002), 'A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)', *Science*, Vol. 296, pp. 79–92 (data from URL: http://www.sciencemag.org/cgi/content/full/296/5565/79/DC1, Web Supplement 2).

43. Cruveiller, S., Jabbari, K., Clay, O. and Bernardi, G. (2003), 'Incorrectly predicted genes in rice?', *Science*, in press.

44. Churchland, P. M. (1990), 'On the nature of theories – a neurocomputational perspective', *Minnesota Studies in the Philosophy of Science*, Vol. 14, pp. 59–101.

45. Burge, C. and Karlin, S. (1997), 'Prediction of complete gene structures in human genomic DNA', *Genome Res.*, Vol. 268, pp. 78–94.

46. Salamov, A. A. and Solovyev, V. V. (2000), '*Ab initio* gene finding in *Drosophila* genomic DNA', *Genome Res.*, Vol. 10, pp. 516–522.

47. Sakata, K., Nagamura, Y., Numa, H. *et al.* (2002), 'RiceGAAS: An automated annotation system and database for rice genome sequence', *Nucleic Acids Res.*, Vol. 30. pp. 98–102.

48. Salzberg, S. L., Pertea, M., Delcher, A. L. *et al.* (1999), 'Interpolated Markov models for eukaryotic gene finding', *Genomics*, Vol. 59, pp. 24–31.

49. Pavlícek, A., Clay, O., Jabbari, K. *et al.* (2002), 'Isochore conservation between MHC regions on human chromosome 6 and mouse chromosome 17', *FEBS Lett.*, Vol. 511, pp. 175–177.

50. Saccone S., Federico, C., Solovei, I. *et al.* (1999), 'Identification of the gene-richest bands in human prometaphase chromosomes', *Chromosome Res.*, Vol. 7, pp. 379–386.

51. Federico, C., Andreozzi, L., Saccone, S. and Bernardi, G. (2000), 'Gene density in the Giemsa bands of human chromosomes', *Chromosome Res.*, Vol. 8, pp. 737–746.

52. BACRC, The BAC Resource Consortium (2001), 'Integration of cytogenetic landmarks into the draft sequence of the human genome', *Nature*, Vol. 409, pp. 953–958.

53. Eichler, E. E. (2001), 'Segmental duplications: What's missing, misassigned, and misassembled – and should we care?', *Genome Res.*, Vol. 11, pp. 653–656.

54. Hattori, M. and Taylor, T. D. (2001), 'Part three in the book of genes', *Nature*, Vol. 414, pp. 854–855.