DNA Composition, Codon Usage and Exon Prediction

Roderic Guigó
Informàtica Mèdica, Institut Municipal d'Investigació Mèdica *
and
Departament d'Estadística, Universitat de Barcelona

May 6, 1998

Contents

*corresponding address: Informàtica Mèdica, IMIM, Dr. Aiguader 80, 08003 Barcelona phone: +343 2211009, fax: +343 2213237, email: rguigo@imim.es

Overview

In this chapter, we review the sequence based measures indicative of protein-coding function in genomic DNA. As the genome projects are entering the large scale sequencing phase, computer programs are becoming essential to identify protein coding genes in large uncharacterized genomic sequences—typically of tens of thousands, or even hundreds of thousands of nucleotides— with efficiency and reliability. At the core all gene identification programs there exist one or more coding measures. Most such programs rely on additional information—mainly, potential sequence signals involved in gene specification, and sequence similarity database searches—, and use very complex frameworks for its integration. Still, a good knowledge of the core coding statistics is important to understand how gene identification programs work, and to interpret their predictions. Here we will review a few of the most important coding measures, and we will illustrate through examples the details involved in their computation.

1 Introduction

A coding statistic can be defined as a function that computes given a DNA sequence a real number related to the likelihood that the sequence is coding for a protein. Since the early eighties, a great number of coding statistics have been published in the literature. Most such coding statistics measure either codon usage bias, base compositional bias between codon positions, or periodicity in base occurrence (or a mixture of all them). Exhaustive reviews can be found elsewhere (see, for instance, Gelfand (1995) and the references therein). Here we follow loosely the critical review by Fickett and Tung (1992). Our classification of coding measures is, however, slightly different. The main distinction here is between measures dependent of a model of coding DNA, and measures independent of such a model. The model of coding DNA is always probabilistic, allowing to compute the probability of a DNA sequence, given that the sequence is coding. Although in the practice, the values (scores) of a given coding statistic in a query sequence can be computed in a number of different ways, here for the model-based coding statistics we will compute scores based on such a probability. Indeed, given a query sequence, we will compute the probability of the sequence under the model of coding DNA, and under an alternative model or non-coding DNA (which, here, for illustration purposes will be simply random DNA). We will take the logarithm of the ratio of these two probabilities—the log-likelihood ratio—as the score of the coding statistic in the query sequence.

Model dependent coding statistics are likely to capture more of the specific features of coding DNA—more of them as more complex the model is (i.e. dependent on more parameters)—. Therefore, model dependent coding statistics may be more powerful in discriminating coding from non-coding DNA. Model dependent coding statistics, however, require of a representative sample of coding DNA from the species under consideration where to estimate the model's parameters (probabilities). The more complex the model is, more sensible to sample bias and size. Model independent coding statistics, on the other hand, capture only the "universal" features of coding DNA; since they do not require of a sample of coding DNA, they can be used even in absence of previously known coding regions from the species under consideration.

To illustrate the coding statistics reviewed here, we will use a few test sequences. As a genomic test sequence, we use a 2000 bp DNA sequence from the human genome coding for the β-globin gene. The sequence has been extracted from the GenBank entry HUMHBB (EMBL entry HSHBB), from positions 62,001 to 64,000. The human β-globin gene has three coding exons in positions 187–278, 409–631, and 1482–1610, relative to the 2000 bp test sequence. In addition, we have extracted two subsequences from this test sequence. The complete sequence of the second coding exon of the β-globin gene, which is 223 bp long, and a 223 bp long sequence from the middle of the second intron (from positions 800 to 1022 of the test sequence). These two sequences will serve as exonic and intronic test sequences. We have also extracted from GenBank rel. 93 (February 1996), a set of 450 human genes following the protocol described in Burset and Guigó (1996) (see also Guigó (1997b)). From this set, exons longer than 100bp and introns longer than 100 bp and shorter than 2500 bp were kept. This resulted in 1753 introns and 1761 exons. The distribution of the scores of the different coding statistics analyzed here will be plotted in these two sets of sequences. Results obtained in the exonic, intronic and genomic test sequences are only for illustration purposes, and differences in the performance of different coding statistics can not be inferred from them. Readers interested in a rigorous comparative benchmarking should refer to Fickett and Tung (1992).

We will try to avoid complex mathematical formulas to describe the algorithms, but we will indicate how to calculate them, so that interested readers can hopefully reproduce the computations. Although, we will thus keep mathematical formalisms to a minimum, it will be useful to maintain a consistent notation through the chapter. Thus,

$$S = S_1 S_2 \cdots S_l$$

will denote a DNA sequence of length l, while S_i $(i = 1 \cdots l)$ will denote the individual nucleotides. For instance if

$$S = \texttt{AGGACGGGATCA}$$

then

$$S_1 = \texttt{A}, \quad S_2 = \texttt{G}, \quad \cdots \quad S_l = \texttt{G}, \quad \text{and} \quad l = 12$$

A DNA sequence can be partitioned in a sequence of consecutive non-overlapping codons in three different ways depending on the nucleotide in the sequence on which the grouping of nucleotides into codons starts (that is, the sequence can be read in three different frames). If C is a sequence of codons

$$C = C_1 C_2 \cdots C_m$$

C_j will denote the codon occupying position j in the sequence. If S is a nucleotide sequence, we will use C_S^i (or simply C^i, $i = 1, 2, 3$) to denote the sequence of codons that results when the grouping of nucleotides from S into codons starts at nucleotide i. We willl use C_j^i to denote the codon occupying position i in the decomposition j of the sequence. For instance, if S is the sequence above, then

$$\begin{array}{llll} C_1^1 = \texttt{AGG} & C_2^1 = \texttt{ACG} & C_3^1 = \texttt{GGA} & C_4^1 = \texttt{TCA} \\ C_1^2 = \texttt{GGA} & C_2^2 = \texttt{CGG} & C_3^2 = \texttt{GAT} & \\ C_1^3 = \texttt{GAC} & C_2^3 = \texttt{GGG} & C_3^3 = \texttt{ATC} & \end{array}$$

On the other hand, if c is a codon, we will use $c[k]$ to denote the nucleotide occupying position k in the codon. For instance, in the example above

$$C_1^1[1] = \texttt{A} \quad C_2^2[3] = \texttt{G} \quad C_1^3[2] = \texttt{A}$$

2 Measures dependent on a Model of Coding DNA

All the measures dependent on a probabilistic model of coding DNA can be computed in a uniform way through the computation of the probability of the sequence given the model. That is, given a probabilistic model of what coding DNA is—the codon usage table, for instance—, we can compute the probability of a sequence of nucleotides S, assuming the sequence is coding in a given frame. We will use

$$P^i(S)$$

to denote the probability of the sequence of nucleotides S, given that S is coding in frame i ($i = 1, 2, 3$). On the other hand, we can compute the probability of S given a model of non-coding DNA. We will use

$$P_0(S)$$

to denote such a probability. $P^i(S)/P_0(S)$ is a likelihood ratio: the ratio of the probability of finding the sequence of nucleotides S, if S is coding in frame i over the probability of finding the sequence of nucleotides S, if S is non-coding. To measure the coding potential of sequence S in frame i given the model of coding DNA, we will compute the natural logarithm of this ratio—the log-likelihood ratio,

$$LP^i(S) = \log \frac{P^i(S)}{P_0(S)}$$

If $LP^i(S) > 0$, then the probability of the sequence of nucleotides S is higher assuming that S is coding in frame i, than assuming that S is non-coding in frame i, while if $LP^i(S) < 0$, then the probability of S is higher assuming that S does not code in frame i than assuming that S is coding in frame i. Given a sequence problem S, we compute the log-likelihood ratios for S in the three frames. If the sequence is coding, the log-likelihood ratio will larger for one of the frames than for the other two.

Through the chapter, we will assume non-coding DNA to be simply random DNA with nucleotide equiprobability and independence between positions. It could be argued that a model inferred from actual non-coding regions of the species under consideration should be used, instead. However, non-coding DNA is usually underrepresented in public databases, and it may exhibit a high degree of heterogeneity along the genome. Therefore, we believe that, at least for illustration purposes, the random assumption does not introduce a significant distortion.

2.1 Measures based on oligonucleotide counts

Unequal usage of codons in the coding regions appears to be a universal feature of the genomes across the phylogenetic spectra. This bias obeys mainly to (i) the uneven usage of the amino acids in the existing proteins and (ii) the uneven usage of synonymous codons (Grantham et al., 1980). The bias in the usage of the synonymous codons correlates with the abundance of the corresponding tRNAs (Ikemura, 1985). The correlation is particularly strong for highly expressed genes. Codon usage is specific of the taxonomic group, and there exist correlation between taxonomic divergence and similarity of codon usage (Ikemura, 1985).

2.1.1 Codon Usage.

By comparing the frequency of codons in a region of an species genome read in a given frame with the typical frequency of codons in the species genes, it is possible to estimate a likelihood of the region coding for a protein in such a frame. Regions in which codons are used with frequencies similar to the typical species codon frequencies are likely to code for genes. This idea was first introduced by Staden and McLahlan (1982). In the practice, the likelihood can be computed in a number of different ways. Here we compute it as a log-likelihood ratio. Let $F(c)$ be the frequency (probability) of codon c in the genes of the species under consideration (in other words, F is the codon usage table, see Table 1). Then, given a sequence of codons $C = C_1 C_2 \cdots C_m$, and assuming independence between adjacent codons

$$P(C) = F(C_1)F(C_2)\cdots F(C_m)$$

is the probability of finding the sequence of codons C knowing that C codes for a protein. For instance, if S is the sequence S=`AGGACG`, when read in frame 1, it results in the sequence $C_1^1 = \texttt{AGG}$, $C_2^1 = \texttt{ACG}$. Then

$$P^1(S) = P(C^1) = F(\texttt{AGG})F(\texttt{ACG})$$

Substituting the appropriate values from Table 1, we obtain

$$P^1(S) = P(C^1) = 0.022 \times 0.038 = 0.000836$$

On the other hand, let $F_0(c)$ be the frequency of codon c in a non-coding sequence.

$$P_0(S) = P_0(C) = F_0(C_1)F_0(C_2)\cdots F_0(C_m)$$

is the probability of finding the sequence S if C is non-coding. Assuming the random model of coding DNA, $F_0(c) = 1/64 = 0.0156$ for all codons, and P_0 for the above sequence of codons C would be

$$P_0(C) = 0.0156 \times 0.0156 = 0.000244$$

The log-likelihood ratio for S coding in frame 1, LP^1, is

$$LP^1(S) = \log(0.000836/0.000244) = \log(3.43) = 0.53$$

The log-likelihood ratios for S coding in frames 2, and 3 (LP^2 and LP^3) are computed in a similar way. Table 4 shows the values of the log-likelihood ratios computed on our test exon and intron sequences, using the values of F from Table 1. As it can be seen, in this case the log-likelihood ratio LP is indeed greater than zero in the coding frame of the exon sequence, while is smaller than zero in the non-coding frames of the exon sequence and in all frames of the intron sequence.

The distribution of the scores of the Codon Usage log-likelihood ratios in the larger sets of intron and exon sequences are shown in Figure 2. Because the sequences in these sets are of very different lengths, the scores are divided by the sequence length in order to derive these distributions. As it is possible to see, although the distributions are clearly distinct, there is substantial overlap between the Codon Usage scores in the sets of intron and exon sequences. As we will see, this is a general situation for all coding statistics.

In the practice, the problem is not usually to determine the likelihood that a given sequence is coding or not, but to locate the (usually small) coding regions within large genomic sequences. The typical procedure is to compute the value of a coding statistic in successive (usually overlapping) windows (an sliding window), and record the value of the statistic for each of the windows. This generates a profile along the sequence in which peaks may point to the coding regions and valleys to the non-coding ones. In Figure 1, we plot the result of sliding a window of length 120 bp, the distance between consecutive windows being 10 bp, computing LP in the three different frames, and plotting the highest value obtained. In this case, the resulting profile reproduces fairly well the exonic structure of the human β-globin gene.

As we have pointed out, codon usage bias is a mixture of both: bias in the usage of amino acids, and bias in the usage of synonymous codons. Methods can be used to measure these effects separately

2.1.2 Amino Acid Usage.

McCaldon and Argos (1988) compute the probabilities of occurrence of different oligopeptides in existing proteins. By translating a sequence of codons to a sequence of amino acids, the probability of the resulting sequence of oligopeptides assuming the region to be coding can be computed. Here, following Fickett and Tung (1992), we compute a measure of amino acid bias based on the observed frequencies of single amino acids in the existing proteins. The measure is identical to the log-likelihood ratio introduced to measure codon usage bias, but the probability of each codon is now the observed probability of the amino acid encoded by the codon. That is, $F_A(c)$ is the observed probability of the amino acid encoded by codon c in the existing proteins; This value can be directly derived from a codon usage table by summing up the probabilities of synonymous codons; that is, given a codon c

$$F_A(c) = \sum_{c' \equiv c} F(c')$$

The Human Codon Usage Table

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	38.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Table 1: The human codon usage and codon preference table as published in `http://bioinformatics.weizmann.ac.il/databases/codon`. For each codon, the table displays the frequency of usage of each codon (per thousand) in human coding regions (first column) and the relative frequency of each codon among synonymous codons (second column).

where, $c' \equiv c$ means c' synonymous to c. Then,

$$P_A^i(S) = P_A(C^i) = F_A(C_1^i)F_A(C_2^i)\cdots F_A(C_m^i)$$

is the probability of finding the sequence of amino acids resulting of translating the sequence S in frame i given that S is coding in frame i. As a model of non-coding DNA, we assume the probability of each amino acid to be proportional to the number of synonymous codons coding for the amino acid, that is $F_{A0}(c) = n_c/64$, where n_c is the number of codons synonymous to c, and we can compute $P_{A0}(S)$. Then, from $P_A^i(S)$ and $P_{A0}(S)$, we compute the Amino Acid Usage log-likelihood ratio as usual. Results obtained using this measure in our test sequences are shown in Table 4, and Figures 1 and 2

2.1.3 Codon Preference.

Gribskov et al. (1984) introduce a coding statistic to measure uneven usage of synonymous codons solely. Indeed, from a codon usage table, we can compute the relative probability of each synonymous codon to code for a given amino acid. For instance, from Table 1, we can see that codons `GAG` and `GAA`—the two codons coding for Glutamic Acid—are used in coding regions with probabilities 0.03882 and 0.02751 respectively, which results in a relative probability of 0.59 and 0.41, respectively. Let $F_R(c)$ be the relative probability in coding regions of codon c among codons synonymous to c,

$$F_R(C) = \frac{F(C)}{\sum_{c' \equiv c} F(c')}$$

Then

$$P_R^i(S) = P_R(C^i) = F_R(C_1^i) F_R(C_2^i) \cdots F_R(C_m^i)$$

is the probability of the sequence S given the particular sequence of amino acids coded by S in frame i (that is, in P_R the effects of unequal usage of amino acids have been eliminated.) We will assume that in non-coding DNA, there is no preference between synonymous codons to code for a given amino acid. Therefore the probability of codon c in non-coding DNA is $F_{R0} = 1/n_c$. From $P_R^i(S)$ and $P_{R0}(S)$ we compute the Codon Preference log-likelihood ratio as usual. Results obtained using this measure in our test sequences are shown in Table 4, and Figures 1 and 2.

As it can be seen from Table 4 and Figure 1, although amino acid usage and codon preference carry a lot of information about coding function, neither of these measures appears to be as discriminant as codon usage. In fact, it is easy to see that, as we have introduced them, codon usage is *the* composition of amino acid usage and codon preference. Indeed, from the definitions above, it follows directly that for a given codon:

$$F(c) = F_A(c) F_R(c)$$
$$F_0(c) = F_{A0}(c) F_{R0}(c)$$

which results in

$$P^i(S) = P_A^i(S) P_R^i(C)$$
$$P_0(S) = P_{A0}(C) P_{R0}(C)$$

for a sequence S in frame i, which in turn leads to

$$LP^i(S) = LP_A^i(S) + LP_R^i(S)$$

which states that Codon Usage bias is the sum of Amino Acid usage bias and Codon Preference.

2.1.4 Hexamer Usage.

Bias in the distribution of oligonucleotides other than codons (tri-nucleotides) can also be used to discriminate between coding and non-coding regions. Bias in the usage of hexamers may be the most discriminant one (probably because of dependence between adjacent amino acids in the proteins). Claverie et al. (1990) were the first to use hexamer frequencies to locate coding regions. Bias in hexamer usage can be computed exactly as bias in codon usage. An hexamer usage table, $F(h_i)$ ($i = 1, \cdots, 4096$) from the species under consideration is computed "a priori". Then, the probability of a sequence of hexanucleotides, $H = H_1, H_2, ..., H_m$, in the coding frame of a coding sequence

nucleotide	codon position		
	1	2	3
A	0.27	0.31	0.18
C	0.24	0.24	0.31
G	0.32	0.20	0.29
T	0.17	0.26	0.22

Table 2: Frequency of the four different nucleotides at the three different codon positions in human coding regions. Derived from Table 1

is $P(H) = F(H_1)F(H_2)\cdots F(H_m)$. If P_0 is the background probability distribution, the log-likelihood ratio LP can be computed as before. Now, a test sequence can be decomposed in six different sequences of hexamers, instead of three, and, thus, six log-likelihood rations can be computed (LP^i, $i = 1 \cdots 6$). Table 4 shows the values of these ratios in our test intron and exon sequences, Figure 2 shows the distributions of the standardized scores (by the sequence length) in the larger sets of intron and exon sequences, and Figure 1 shows the results of sliding a window, and plotting the maximum of the six values at each position.

2.2 Measures based on base compositional bias between codon positions

From the codon usage table (Table 1), we can derive the probability of each base at each codon position in coding regions (Table 2). As it is possible to see, there are clear differences in the frequency with which the different bases appear at the different codon positions; for instance, `G` is almost twice as frequent as `T` in the first codon position, while `T` is more frequent than `G` in the second codon position. Similarly, `C` is almost twice as frequent as `A` in the third codon position, but `A` is more frequent than `C` in the second codon position. Sheperd (1981) already noted that the most frequent codons were of the form `RNY` (`R` = `A` or `G`, `Y` = `C` or `T`, `N` any nucleotide). He suggested a method to test for the existence and frame of a coding region by measuring the number of differences between the sequence and the pattern `RNYRNRY`$\cdots$`RNY`. A number of other measures have been latter proposed to exploit the asymmetry in the base composition between codon positions in order to locate potential coding regions in genomic DNA.

Before discussing some of these measures, we would like to point out that asymmetry in the base composition between codon positions arises, not only because of uneven usage of amino acids and synonymous codons, but also because of the particular structure of the genetic code. Indeed, uneven usage of amino acids and uneven usage of synonymous codons are not enough to produce asymmetry in base composition, as the following example illustrates:

EXAMPLE. CODON PREFERENCE AND AMINO ACID USAGE BIAS DO NOT NECESSARILY RESULT IN CODON ASYMMETRY IN BASE COMPOSITION

Let's assume a three letter "DNA" and "amino acid" codes:

$$DNA = \{A, B, C\} \text{ and } AA = \{P, Q, R\}$$

Let's assume "codons" to be di-nucleotides, and let's assume strong bias in "amino acid" usage, and in "codon" preference, as expressed in the following "genetic code" table:

P	0.7
Q	0.2
R	0.1

amino acid usage bias

P	AA	0.6	Q	BB	0.6	R	AB	0.5
P	BC	0.2	Q	AC	0.2	Q	BA	0.5
P	CB	0.2	Q	CA	0.2	stop	CC	

codon preference bias

This results in a very biased "codon" usage table, but not in "codon" asymmetry in base composition

AA	0.42	AB	0.05	AC	0.04
BA	0.05	BB	0.12	BC	0.14
CA	0.04	CB	0.14	CC	

"codon" usage

	codon position	
	1	2
A	0.51	0.51
B	0.31	0.31
C	0.18	0.18

base composition at "codon" positions

Because of the structure of the genetic code, synonymous codons almost always share the first two nucleotides (the exception being obviously the amino acids coded by six codons, Arginine, Serine, and Leucine). This implies that the first two positions in the codons will be more abundant in those nucleotides common to the synonymous codons corresponding to the most abundant amino acids. On the other hand, differences between synonymous codons are mostly confined to the third codon position. At this position, C and G are usually preferred, as genes (at least, in higher eukariotes) tend to occur in G+C rich regions (references...)

2.2.1 Codon Prototype.

The distribution of base frequencies at codon positions (Table 2) can be assumed to describe statistically a prototypical codon. Then, given a sequence problem S, we can measure how similar to the prototypical distribution is the observed distribution of base frequencies at the three codon positions in S. Closer the distributions, more likely for S to be coding. As usual, there are a number of ways in which such a "proximity" can be measured (Fickett and Tung, 1992; Mural et al., 1991). Here we will compute the usual log-likelihood ratio.

Let $f(b,r)$ be the probability of nucleotide b at codon position r, as estimated from known coding regions (Table 2). Then, if c is a codon

$$F(c) = f(c[1],1)f(c[2],2)f(c[3],3)$$

is the probability of codon c in coding regions, assuming independence between adjacent nucleotides. On the other hand, we will assume $F_0(c) = 1/64$ the probability of for all triplets c in non-coding DNA. From F and F_0, P^i and P_0 can be computed exactly as done before in deriving the codon usage log-likelihood ratio. For instance, if the sequence S is `AGGACG`, the probability of S, if S is coding in frame 1, can be computed as

$$P^1(S) = F(\texttt{AGG})F(\texttt{ACG}) = f(\texttt{A},1)f(\texttt{G},2)f(\texttt{G},3)f(\texttt{A},1)f(\texttt{C},2)f(\texttt{G},3)$$

From Table 2, we obtain

$$P^1(S) = \underbrace{0.27 \times 0.20 \times 0.29}_{F(\texttt{AGG}) = 0.01566} \times \underbrace{0.27 \times 0.24 \times 0.29}_{F(\texttt{ACG}) = 0.01879} = 0.0002943$$

$P^2(S)$ and $P^3(S)$ are computed in a similar way. From P^i, and P_0, the Codon Preference log-likelihood ratio is derived as usual. Results obtained using this measure are shown in Table 4, and Figures 1 and 2.

2.3 Measures based on dependence between nucleotide positions

Both Codon Prototype and Codon Usage are based on a model of coding DNA described by the probabilities of the codons. The models, however, are very different. In Codon Usage, the model is described by the explicit probability of each codon. In Codon Prototype, the model is simply described by the probability of occurrence of each base at each position in a codon. Codon Prototype and Codon Usage would be equivalent if codon positions were independent. This is not clearly the case: frequencies of codons derived from Table 2 assuming independence between codon positions are substantially different than the observed codon frequencies (Table 1). Measures based on the frequency of usage of oligonucleotides, such as Codon Usage, implicitly capture such dependences between nucleotide positions within codons in coding regions. Dependencies between nucleotide positions in coding regions, however, can be explicitly described by means of Markov Models.

2.3.1 Markov Models

Borodovsky and McIninch (1993) first introduced the usage of Markov Models to locate potential coding regions in DNA sequences. For illustration purposes, it may be helpful to introduce Markov Models from Codon Prototype. As we have seen, in Codon Prototype, the probability of a nucleotide to appear in a given codon position is constant, independent of the nucleotides in nearby positions. For instance, if S is the sequence above ($S =$ `AGGACG`), the probability of `G` at codon position 3 (S_3 and S_6) is constant, 0.29, whether the nucleotide preceding `G` is `G` (as in S_2) or `C` (as in S_5). In the Markov Models, however, the probability of a nucleotide at a particular codon position depends (is conditionated) on the nucleotide(s) preceding it.

In the simplest of the Markov Models, the Markov Models of order 1, the probability of a nucleotide depends only on the preceding nucleotide. In this case, the model of coding DNA is based on the probabilities of the four nucleotides at each codon position, depending on the nucleotide occurring at the preceding codon position (technically called the *transition probabilities*). Thus, instead of one single matrix, as in Codon Prototype, three 4×4 matrices (the *transition matrices*) are required, F^1, F^2, and F^3, each one corresponding to a different codon position. Coefficient i, j from matrix F^r, $F^r(i, j)$, corresponds to the probability of nucleotide i in codon position $r + 1$ (position 1, if $r = 3$), given that nucleotide j is at codon position r. We have estimated these matrices from the sample of 1761 human exons. The conditional probability of nucleotides i in codon position $r + 1$, given nucleotide j in codon position r, is estimated by the number of times that di-nucleotide j, i appears at codon position r over the total number of times that nucleotide j appears at codon position r. These matrices are shown in Table 3. Indeed, we can see from them that the probability of `G` at codon position 3, given that `C` is at codon position 2, is 0.27, but the probability of `G` at codon position 3 given that `G` is at codon position 2 is 0.37.

The probability of S above, given that S is coding in frame 1, can be computed now as

$$P^1(S) = f(\mathtt{A}, 1)F^1(\mathtt{G}, \mathtt{A})F^2(\mathtt{G}, \mathtt{G})F^3(\mathtt{A}, \mathtt{G})F^1(\mathtt{C}, \mathtt{A})F^2(\mathtt{G}, \mathtt{C})$$

Obviously, the probability of the first nucleotide in the sequence is not described by the transition matrices, because it is not preceded by any nucleotide. As the probability of the first nucleotide in the sequence (the *initial probability*) we can assume simply the probability of such nucleotide depending on its codon position, that is, the value in the Codon Prototype Table (Table 2). Then, substituting the appropriate values from Tables 3 and 2 in the above equation, we obtain the probability of S coding in frame 1:

$$P^1(S) = 0.27 \times 0.19 \times 0.27 \times 0.24 \times 0.21 \times 0.41 = 0.0002862$$

Similarly, the probability of the sequence of nucleotides S, given that S codes in frame 2 is

$$\begin{array}{ccccccccc} P^2(S) & = & f(\mathtt{A}, 2) & F^2(\mathtt{G}, \mathtt{A}) & F^3(\mathtt{G}, \mathtt{G}) & F^1(\mathtt{A}, \mathtt{G}) & F^2(\mathtt{C}, \mathtt{A}) & F^3(\mathtt{G}, \mathtt{C}) \\ \| & & \| & \| & \| & \| & \| & \| \\ 0.0006744 & = & 0.31 & 0.40 & 0.37 & 0.35 & 0.28 & 0.15 \end{array}$$

$P^3(S)$ can be computed in exactly the same way. If F_0 is the matrix of conditional probabilities of the different nucleotides given the preceding ones, we can compute the usual log-likelihood ratio. Assuming a random model of non-coding DNA, the probability

codon position 2 \ codon position 1	A	C	G	T
A	.36	.27	.35	.18
C	.21	.23	.24	.27
G	.19	.14	.23	.23
T	.24	.35	.19	.31

codon position 3 \ codon position 2	A	C	G	T
A	.16	.19	.15	.07
C	.28	.44	.41	.33
G	.40	.12	.27	.45
T	.16	.25	.17	.16

codon position 1 \ codon position 3	A	C	G	T
A	.22	.33	.24	.13
C	.21	.29	.27	.21
G	.44	.15	.37	.53
T	.13	.22	.12	.13

Table 3: Probabilities of the four nucleotides at the different codon positions conditioned to the nucleotide in the preceding codon position. Estimated from our set of human exon and intron sequences.

of a nucleotide i does not depend on the preceding nucleotide j, then $F_0(i,j) = 0.25$ for each pairs of nucleotides i, j, Results obtained in the test sequences with the log-likelihood ratios corresponding to a Markov Model of order 1 are shown in Table 4, and Figures 1 and 2.

In Markov Models of order 2, the probability of a given nucleotide at a given codon position depends on the di-nucleotide preceding it. The transition matrices have now 4 raws and 16 columns (one for each possible di-nucleotide). $F^2(\texttt{A}, \texttt{GC})$, for instance, would be the probability of `A` following the di-nucleotide `GC` at codon position 2 (that is, `A` would be at codon position 1). In general, the order of the Markov Model indicates the number of preceding nucleotides on which the probability of a given nucleotide depends. In a Markov Model of order k, thus, the coefficients $F_k^r(i,j)$ correspond to the probability of oligonucleotide j, i of length $k+1$ at codon position r, given oligonucleotide i of length k at codon position r. These probabilities are estimated by the frequency of oligonucleotide j, i of length $k+1$ at codon position r over the frequency of oligonucleotide j of length k at position r. Borodovsky and McIninch (1993) investigate Markov Models of order up to $k = 5$. Results obtained in the test sequences with Markov Models of order 2 and 5 are shown in Table 4, and Figures 1 and 2.

Markov Models of higher order may capture more of the intrinsic features of coding DNA, but on the other hand they also depend on more parameters. Since Markov Models of order 2 are based on counts of tri-nucleotides they are, somehow, similar to Codon Usage. However, while Codon Usage reflects only dependences between contiguous nucleotides within codons, a Markov Model of order 2 also reflects nucleotide dependences between nucleotides in contiguous codons. Consequently, it also depends on more parameters: while Codon Usage depends on 64 probabilities, a Markov Model of order 2 depends on four 4×16 matrices (corresponding to the three transition matrices and the initial probabilities matrix), that is, it depends on 256 estimated probabilities. A similar reasoning can be applied to the relationships between Hexamer Usage and Markov Models of order 5.

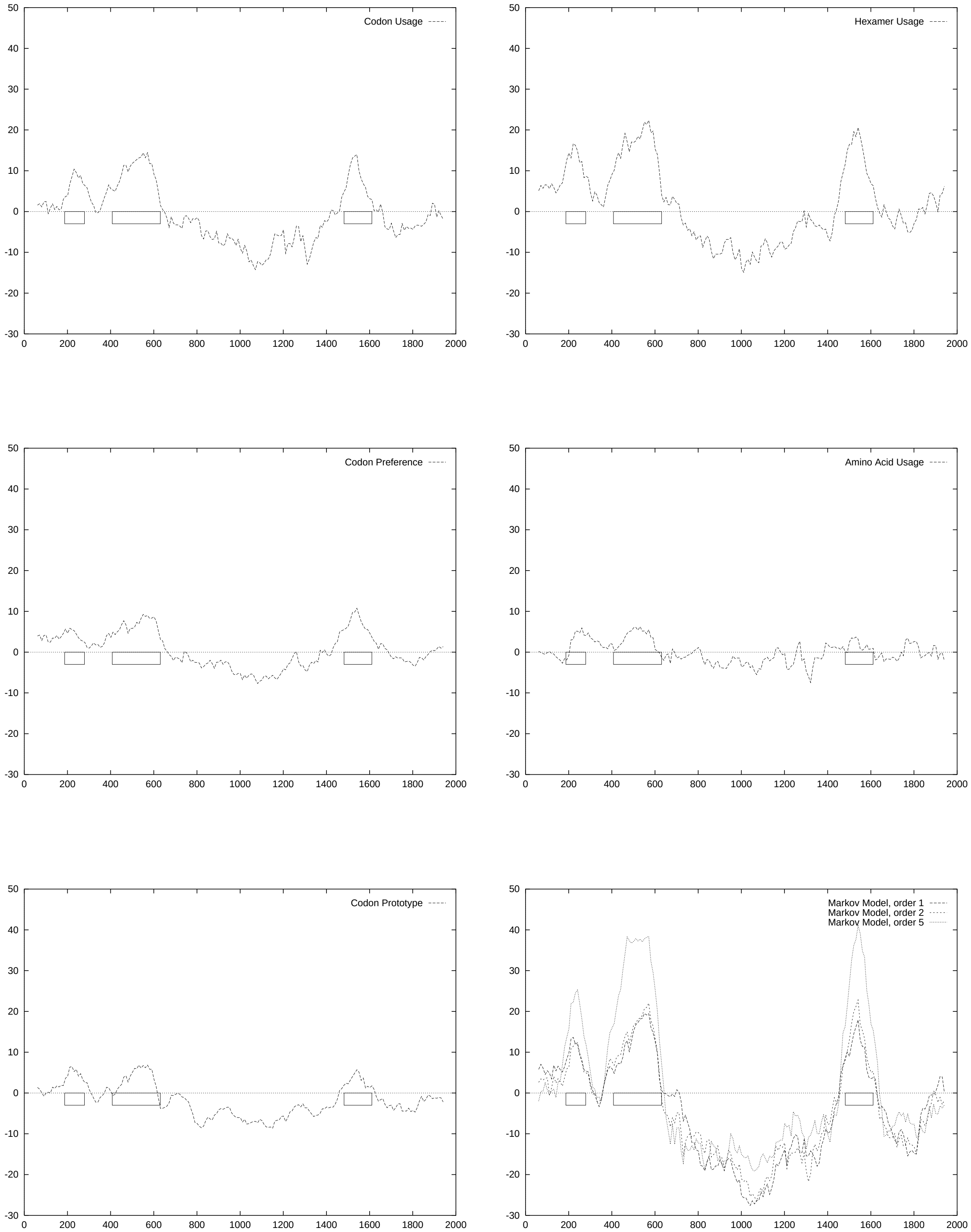

Figure 1: Values of the model based Coding Statistics along the 2000 bp human β-globin gene sequence, computed on an sliding window of length 120 and step 10.

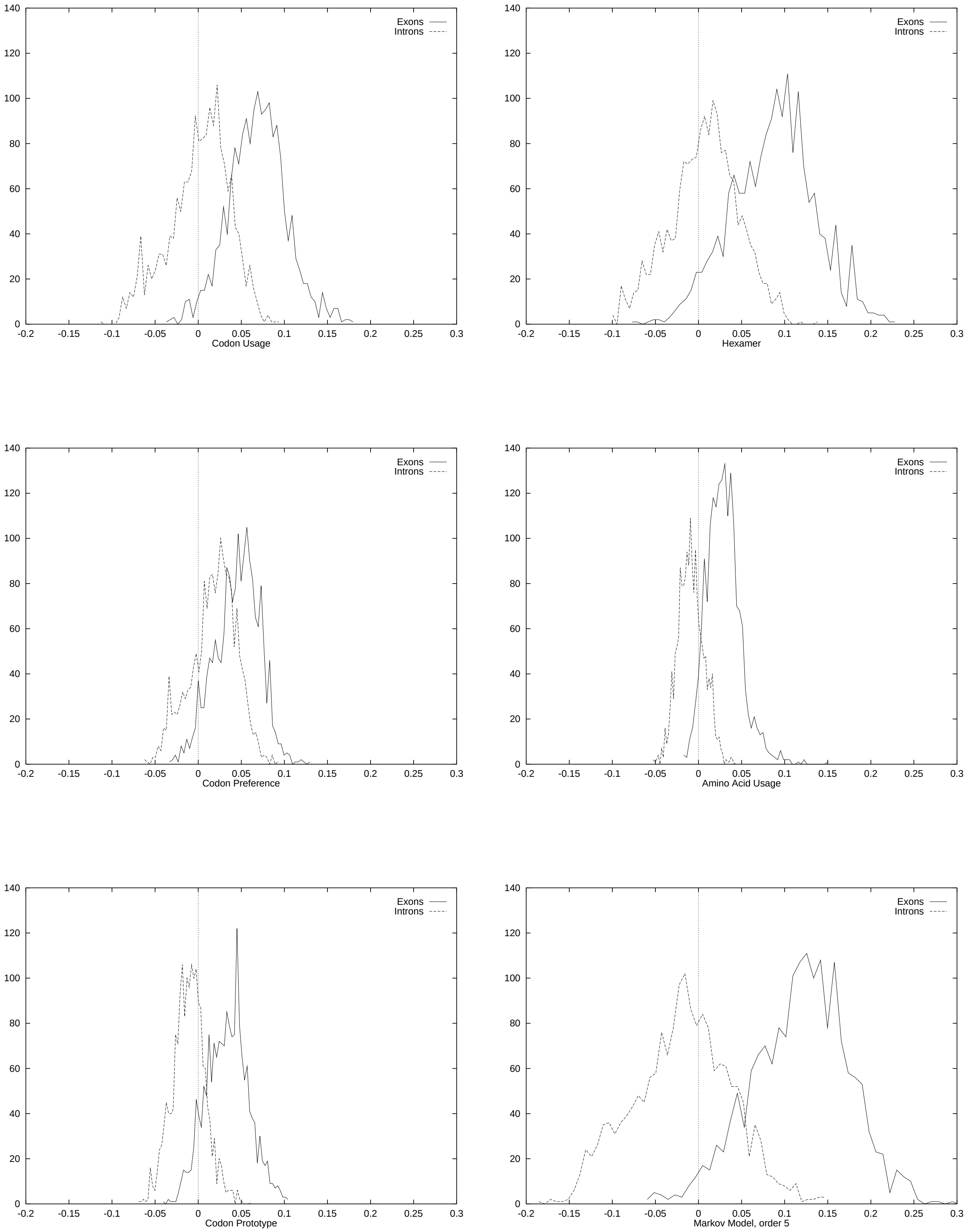

Figure 2: Distribution of the scores of the model based Coding Statistics in the set of 1761 human exons and 1753 human intron sequences. To plot them, the values of the Coding Statistics are divided by the length of the sequence.

3 Measures independent of a Model of Coding DNA

All the methods reviewed so far rely on a probabilistic model of what coding DNA is, under which the coding likelihood of DNA sequences is computed. To estimate the probabilities describing the coding model (of codons, amino acids, synonymous codons, hexamers, nucleotides at codon positions, ...) an non-biased sample of coding DNA is ideally required. However, for most species such a sample does not exist. Indeed for most eukariotic organisms (other than *Sacharomices cerevisiae* and maybe *Caenorhabditis elegans*) only an small fraction of the genes are known, and the set of known genes tend to be biased towards the highly expressed ones—which are likely to exhibit characteristic sequence features as, for instance, strong codon preference bias—. The situation is even worst in the case of the prokariotic genomes. Recent technological progress has made shotgun sequencing of whole prokariotic genomes (up to a few megabases) feasible. It is usually the case, thus, that no sequences of genes are known before the whole genome a prokariotic organism is sequenced. Coding measures not depending of an "a priori" model of coding DNA would, therefore, be very useful. A number of such measures have been proposed. In general, the underlying assumption is that coding DNA is less "random" or "homogeneous" than non coding DNA with respect to some feature related to codigness—codon usage, base composition—. Deviation from randomness or inhomogeneity can be measured independently of a reference model, and the resulting score correlated with coding function. Since there is no reference model, these scores do not have a direct probabilistic meaning, although their distribution can be empirically studied in known sets of coding and non-coding sequences.

Obviously, deviation from randomness or inhomogeneity may in the practice mean a number of different things. Fichant and Gautier (1987), for instance, measure, using Correspondence Analysis, the degree of homogeneity in codon usage between the three frames of the sequence problem. The assumption is that if the sequence is coding, codon usage will be markedly different in the coding frame than in the two other frames—and therefore it will exist inhomogeneity in codon usage between frames—, while if the sequence is not coding, codon usage will be the essentially the same in the three frames—and codon usage will be homogeneous between frames. While Fichant and Gautier (1987) measure is based on the usage of codons, Fickett and Tung (1982) and Staden (1984) propose measures independent of a reference model, based on the asymmetry in the base composition between codon positions. We discuss one such measure next.

3.1 Measures based on base compositional bias between codon positions

3.1.1 Position Asymmetry.

The goal here is to measure how asymmetric is the distribution of nucleotides at the three triplet positions in the sequence problem. Both, Fickett and Tung (1982) and Staden (1984) calculate the asymmetry independently for each nucleotide (although using different formulas), and then combine the values into a single score. Staden (1984) simply sums the four values. Fickett and Tung (1982), in the widely used TESTCODE program, considers in addition the frequencies of the four nucleotides. Each of the asymmetries and frequencies is used to make an estimate of coding likelihood—which makes this measure dependent of a sample set of known coding and non-coding sequences—, and the separate estimates are all combined using a linear weighted sum. Here we also compute the asymmetry independently for each nucleotide, simply as the variance of the frequency of the nucleotide at the three codon positions as suggested in Fickett and Tung (1992), and sum the four values obtained into a single score. Let $f_S(b,r)$ be the (relative) frequency of nucleotide b at codon position r in the sequence problem S, as calculated from one of the three decompositions of S in codons (any of them). Let

$$f_S(b) = \sum_{r=1}^{3} (f_S(b,r))/3$$

be the average frequency of nucleotide b at the three codon positions, and let's define the asymmetry in the distribution of nucleotide b, as the variance of this frequency

$$\text{asym}(b) = \sum_{i=1}^{3} (f_S(b,i) - f_S(b))^2$$

Note that the value of asym(b) is independent of the frame in S in which the codons are defined. Therefore, only one value of asymmetry needs to be computed along the sequence problem (and not one for each frame, as we have been doing so far). Then we compute the Position Asymmetry of the sequence, PA(S) as

$$PA(S) = \text{asym}(A) + \text{asym}(C) + \text{asym}(G) + \text{asym}(T)$$

Table 4, and Figures 5 and 6 show the results obtained when PA is computed in our test sequences.

3.2 Measures based on periodic correlations between nucleotide positions

Given a DNA sequence, we can compute how many times nucleotide i is followed by nucleotide j at a distance of k nucleotides, $N_{ij}(k)$. For instance, if the sequence

$$S = \texttt{AGGACGGGATCA}$$

then $N_{GA}(1) = 2$, $N_{AT}(0) = 1$, $N_{GG}(0) = 3$, $N_{AA}(7) = 2$, and so on. Figure 3 shows the absolute frequency of the pair $\texttt{A} \cdots \texttt{A}$ with k nucleotides between the two `A`'s occurring in the 200 first base pairs of the sequences in our test sets of human exons and introns. As it is possible to see, a clear periodic pattern arises from the set of exon sequences. The nucleotide `A` is more likely to be found at distance $k = 2, 5, 8, \cdots$ from another

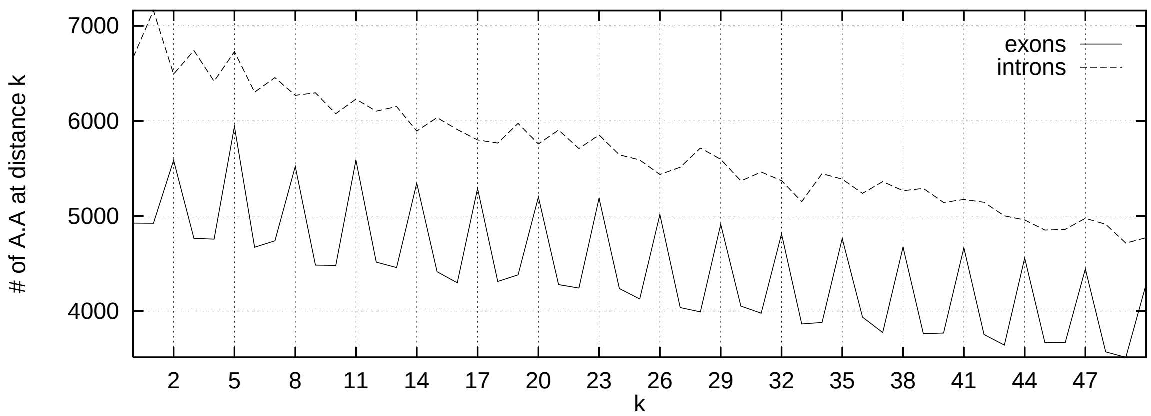

Figure 3: Periodic structure in DNA sequences. The absolute frequency of the pair A $\cdots$ A with k (from 0 to 5) nucleotides between the two A's in the 200 first base pairs of the sequences in the set of 1761 human exons and 1753 human introns. A clear period-3 pattern appears in coding regions, which is absent in non-coding regions. Due to the finite size of the sequences (200 bp) the periodic pattern vanishes at longer distances k. A similar periodic pattern appears in coding regions for the other fifteen possible pairs of nucleotides.

A than at other distances. Note that nucleotide pairs at a distance of $k = 2, 5, 8, \cdots$ nucleotides, are at the same codon position, whereas nucleotide pairs at other distances, are not. Such a periodic pattern reflects correlations between nucleotide positions along coding sequences (that is, the probability of finding a nucleotide at a given position in a coding sequences is not independent of the nucleotide occurring at some other—even distant—position). The correlations arise, in turn, because of the asymmetry in base composition at the three codon positions in coding sequences (Gutiérrez et al., 1994). The periodic pattern, which is characteristic of the 16 pairs of nucleotides, and not only of the pair A $\cdots$ A, is absent in the intronic sequences.

A number of coding statistics have been devised based in measuring the periodic structure (or the correlation structure) of DNA sequences. We discuss three such measures next. Konopka (1994) in the Position Asymmetry Index, compares the probability of pairs of the same nucleotide to appear at a distance $k = 2, 5, 8, \cdots$... (that is, at the same codon position) in the query sequence with the probability of these pairs to appear at other distances (different codon positions), Herzel and Große (1995) in the Average Mutual Information, compare the correlations of all pairs of nucleotides at the same codon positions with the correlations at different codon positions. Finally, Tiwari et al. (1997) use the relative peak at the frequency 1/3 in the Fourier spectrum of the sequence.

3.2.1 Periodic Asymmetry Index

Given a sequence S, Konopka (1994) considers three distinct probabilities, the probability P_{in} of finding pairs of the same nucleotide at distances $k = 2, 5, 8, \cdots$, the probability P^1_{out} of finding pairs of the same nucleotide at distances $k = 0, 3, 6, \cdots$, and the probability P^2_{out} of finding pairs of the same nucleotide at distances $k = 1, 4, 7, \cdots$. Because of the 3-base periodic pattern, in coding regions P_in will be larger than the other two probabilities, while in non-coding regions the three probabilities will be similar. The tendency to cluster homogeneous di-nucleotides in a 3-base periodic pattern can be measured by the Periodic Asymmetry Index

$$PAI(S) = \frac{\max(P_{in}, P^1_{out}, p^2_{out})}{\min(P_{in}, P^1_{out}, p^2_{out})}$$

which can be taken as an indicative of the coding potential of the sequence S; In fact, Konopka (1994) computes the Periodic Asymmetry Index in an slightly different way, he computes the tendency to cluster di-nucleotides in a 2-base periodic pattern (suggested to be characteristic of intronic sequences) and computes the Periodic Asymmetry Index as the ratio of the two tendencies (2-base over 3-base periodicity).

Table 4, and Figures 5 and 6 show the results obtained when PAI is computed in our test sequences.

3.2.2 Average Mutual Information

Given a sequence S, let $P_{ij}(k)$ be the probability in the sequence of the pair of nucleotides i and j at a distance of k nucleotides. These probabilities can be estimated by the absolute frequencies $N_{i,j}(k)$ above (see Li (1997) for considerations regarding the estimation of these probabilities). The correlation between nucleotide i and nucleotide j at a distance of k nucleotides can be calculated (Li, 1997) as:

$$\rho_{ij}(k) = P_{ij}(k) - p_i p_j$$

where p_i and p_j are the probabilities of nucleotides i and j in S. Thus, for each distance k, sixteen different individual correlations can be calculated. A measure that summarizes all individual correlations at a given distance k is the Mutual Information function (Shannon, 1948):

$$I(k) = \sum_{i,j \in \{A,C,G,T\}} P_{ij}(k) \log\left(\frac{P_{ij}(k)}{p_i p_j}\right)$$

$I(k)$ quantifies the amount of information that can be obtained from one nucleotide about another nucleotide at a distance k. Figure 4 shows the Mutual Information Function computed on the first 200 bp of the sequences in the set of human exons and introns. The 3-base periodic pattern in coding sequences becomes obvious. $I(k)$ has larger values for $k = 2, 5, 8, \cdots$. The pattern is absent in non-coding sequences.

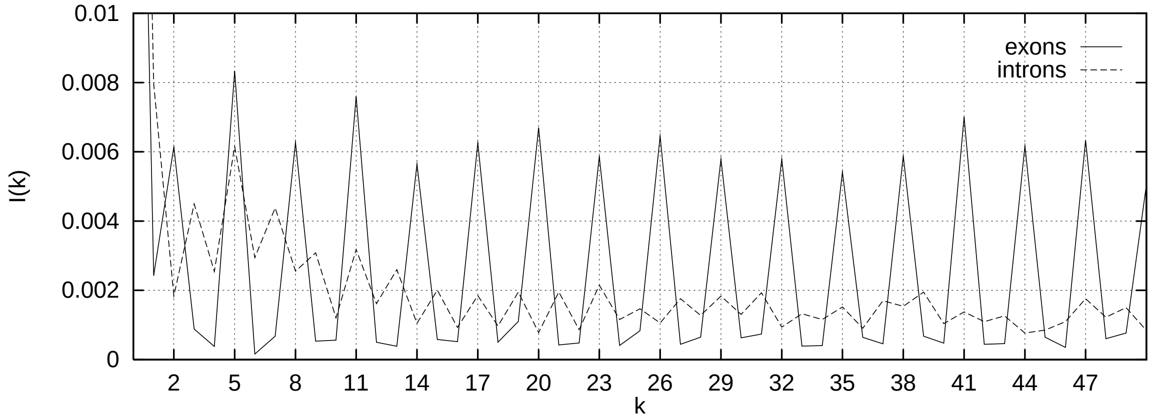

Figure 4: The Mutual Information function computed for distances from $k = 0$ to $k = 50$ in the 200 first bp from the sequences in set of 1761 human exons and in the set of 1753 human introns.

In coding DNA, thus, $I(k)$ oscillates between two values, while in non-coding DNA, $I(k)$ is rather flat. Herzel and Große (1995) use this fact to construct a coding statistic. They call the two values between which $I(k)$ oscillates in coding DNA, the *in-frame* mutual information I_{in} at distances $k = 2, 5, 8, \cdots$, and the *out-of-frame* mutual information I_{out} at $k = 4, 5, 7, 8, \cdots$. They show that I_{in} and I_{out} can be computed directly from the probabilities $f_S(b, r)$ of the nucleotide b to appear at codon position r, as estimated from S. In order to reduce the pair of numbers I_{in} and I_{out} to a single quantity, they compute the Average Mutual Information as (Groβe et al., 1998)

$$AMI = \frac{I_{in} + 2I_{out}}{3}$$

Table 4, and Figures 5 and 6 show the results obtained when AMI is computed in our test sequences.

Correlation structures in DNA sequences—such as those measured in AMI— may reveal biologically relevant large scale heterogeneity in genomic sequences, other than coding. For a review of correlation structures in DNA sequences, and their biological implication see (Li, 1997).

3.2.3 Fourier Spectrum

Periodic correlations in DNA sequences can also be examined by means of Fourier Analysis. The partial spectrum of a DNA sequence S of length l corresponding to nucleotide b is defined as (Li et al., 1994; Silverman and Linsker, 1986):

$$S_b(f) = \frac{1}{N^2} \left(\sum_{j=1}^{l} U_b(S_j) e^{2\pi i f j} \right)^2$$

where $U_b(S_j) = 1$ if $S_j = b$, and it is 0 otherwise, and f is the discrete frequency, $f = k/l$, with $k = 1, 2, \cdots, l/2$. The total Fourier Spectrum of the DNA sequence is the sum of the four partial Spectra:

$$S(f) = \sum_{b \in \{A,C,G,T\}} S_b(f)$$

DNA coding regions reveal the characteristic periodicity of 3 as a distinct peak at frequency $f = 1/3$. No such "peak" is apparent for non-coding sequences (Tsonis et al., 1991), (Tiwari et al., 1997). We have computed the Fourier Spectrum at $f = 1/3$ ($S(1/3)$) in our test sequences. Results appear in Table 4, and Figures 5 and 6. As it is possible to see from Figure 5, the Fourier Spectrum profile in the human β globin gene sequences is identical (save scale) to the Position Asymmetry profile. This indicates that there is a one to one correspondence between Fourier Spectrum and Position Asymmetry, and that one measure can be directly derived from the other. In fact, this can be shown analytically (Ivo Große, personal comunication). The relation between the two measures depends on the length of the sequence, as the dissimilar distributions of the Position Asymmetry and Fourier Spectrum scores indicates in the set of 1761 human exons and 1753 human introns, which have variable length.

Actually, in order to obtain a cleaner signal, Tiwari et al. (1997) build the ratio of the Fourier Spectrum at $f = 1/3$ over the average of the total spectrum of the sequence, ($\overline{S}$), which can be computed from the frequencies of the nucleotides along the sequence.

		exon sequence			intron sequence	
		coding frame	non coding frames		frame 1	frame 2	frame 3
Codon Usage		24.06	-16.13	-3.16	-14.36	-23.74	-19.67
Hexamer Usage		27.62	-11.64	-6.51	-20.90	-27.56	-22.07
		39.98	-14.58	-8.46	-26.73	-27.81	-25.87
Codon Preference		15.97	-1.32	7.24	-7.96	-12.70	-14.93
Amino Acid Usage		8.17	-14.87	-10.17	-6.15	-10.69	-4.57
Codon Prototype		9.87	-11.23	-10.30	-11.45	-17.44	-14.49
Markov Model	order 1	29.92	-2.69	-3.31	-35.44	-42.40	-41.73
	order 2	34.73	-18.26	-7.77	-29.61	-41.76	-40.05
	order 5	72.69	-21.38	13.56	-37.63	-30.99	-36.40
Position Asymmetry		0.0957			0.0211		
Periodic Asymmetry Index		1.159			1.009		
Average Mutual Information		0.00681			0.000344		
Fourier Spectrum		2.278			0.892		

Table 4: Values of different coding statistics in the 223 bp long second coding exon of the human β-globin gene, and in a 223 bp long sequence from the middle of the second intron of the same gene

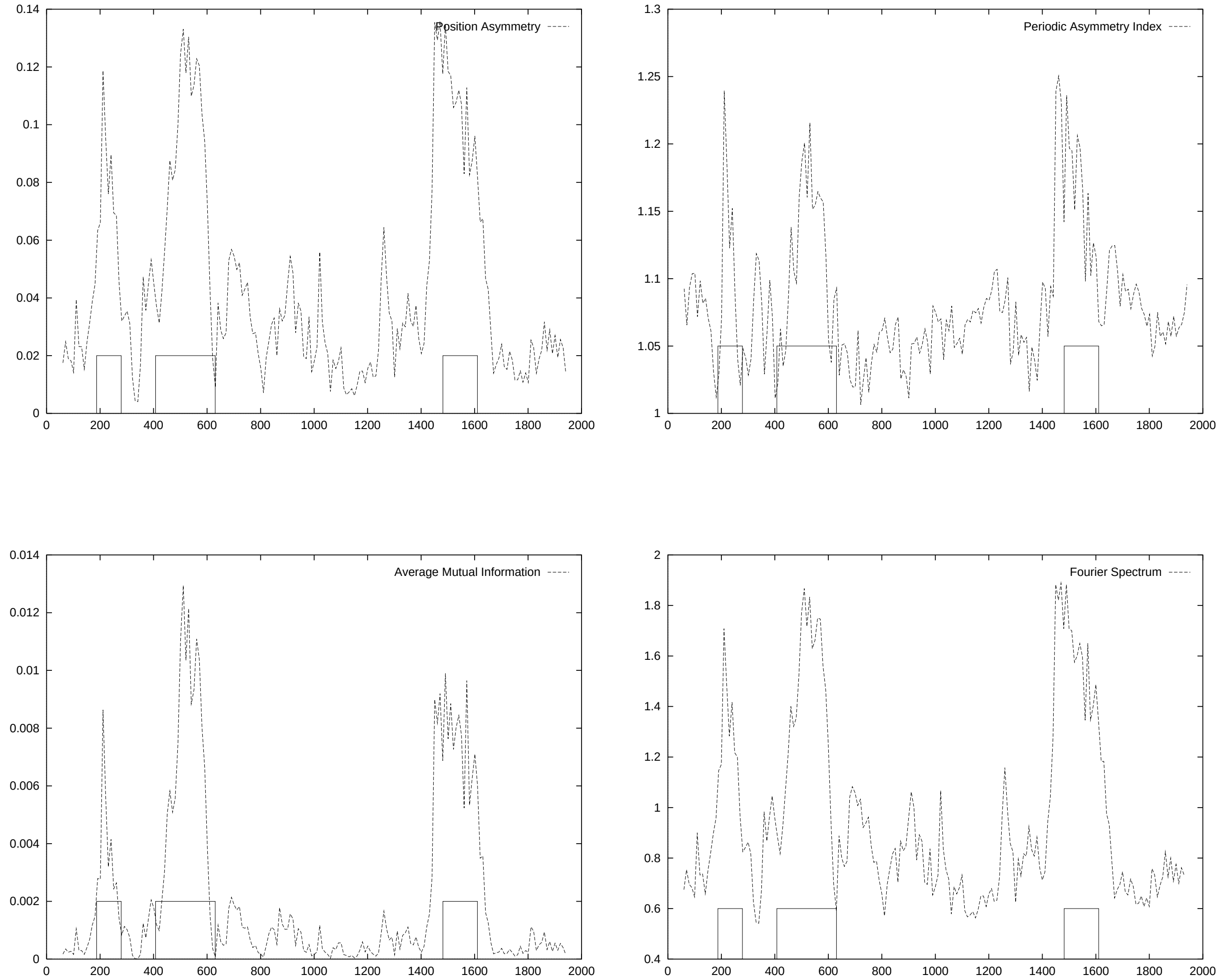

Figure 5: Values of the model independent Coding Statistics along the 2000 bp human β-globin gene sequence, computed on an sliding window of length 120 and step 10.

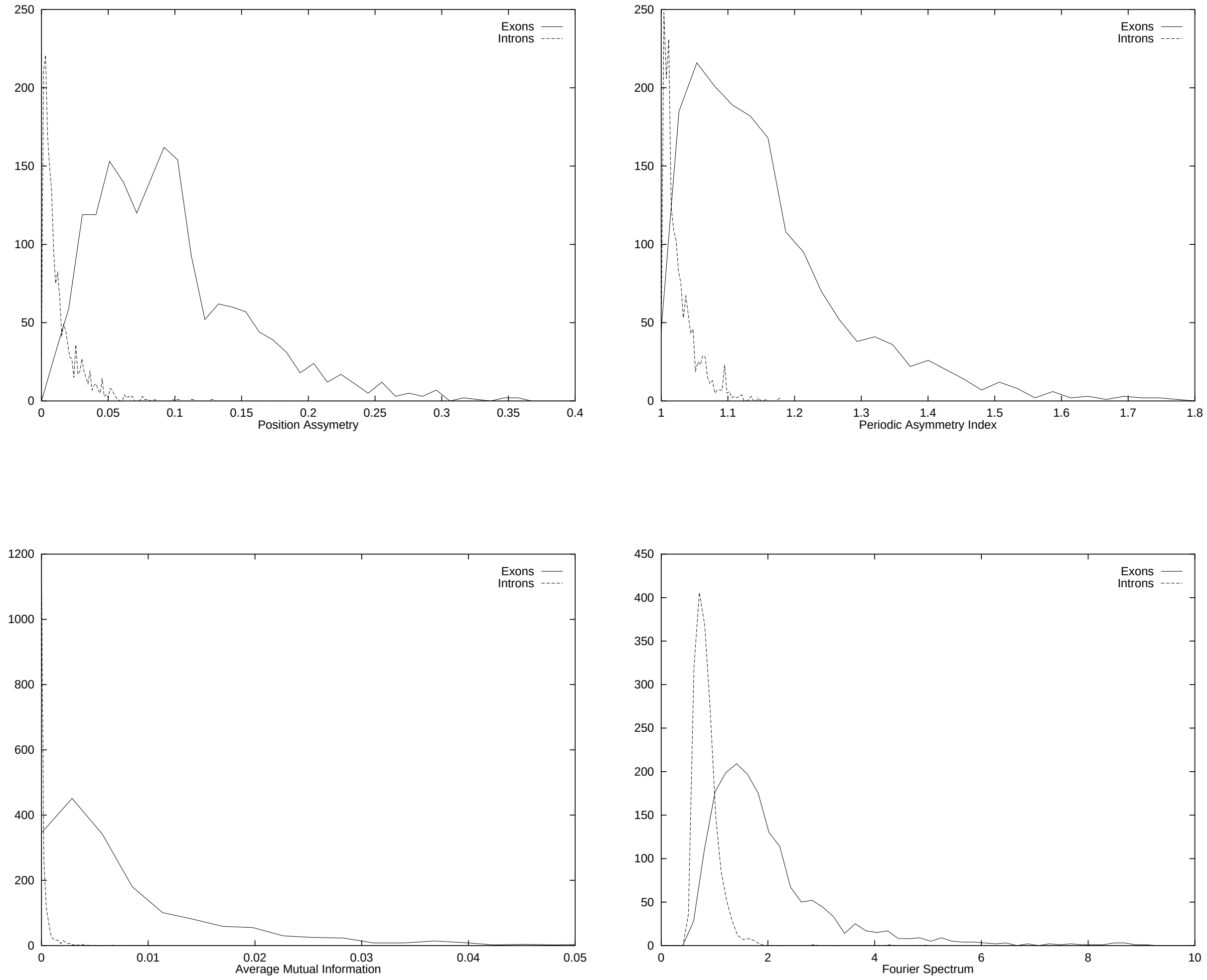

Figure 6: Distribution of the scores of the model independent Coding Statistics in the set of 1761 human exons and 1753 human intron sequences. To plot them, the values of the Coding Statistics are divided by the length of the sequence.

Program	Authors	WWW address
GENEMODELER	Fields and Soderlund, 1990	
GENEID	Guigó et al., 1992	`www1.imim.es/geneid.html`
		`geneid@darwin.bu.edu`
SORFIND	Hutchinson and Hayden, 1992	
GENEPARSER	Snyder and Stormo, 1993	`beagle.colorado.edu/ eesnyder/GeneParser.html`
GENEMARK	Borodovski and McIninch, 1993	`intron.biology.gatech.edu/~genmark`
		`genmark@ford.gatech.edu`
GENVIEW	Milanesi et al., 1993	`www.itba.mi.cnr.it/webgene`
GREAT	Gelfand and Roytberg, 1993	
GRAIL II / GAP	Xu et al., 1994	`avalon.epm.ornl.gov/gallery.html`
		`grail@ornl.gov`
FGENEH	Solovyev et al., 1994	`dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html`
		`service@bchs.uh.edu`
GENELANG	Dong and Searls, 1994	`cbil.humgen.upenn.edu/~sdong/genlang_home.html`
		`genlang@cbil.humgen.upenn.edu`
XPOUND	Thomas and Skolnick, 1994	
GENIE	Kulp et al., 1996	`www-hgc.lbl.gov/inf/genie.html`
PROCRUSTES	Gelfand et al., 1996	`www-hto.usc.edu/software/procrustes/`
MZEF	Zhang, 1997	`www.cshl.org/genefinder`
GENSCAN	Burge and Karlin, 1997	`gnomic.stanford.edu/GENSCANW.html`
MORGAN	Salzberg et al., 1997	`www.cs.jhu.edu/labs/compbio/morgan.html`
VEIL	Henderson et al., 1997	`www.cs.jhu.edu/labs/compbio/veil.html`

Table 5: List of Gene Identification programs, and Internet access. e-mail server address is provided when different from the WWW address.

4 Coding Statistics in Gene Identification Programs

A number of gene identification programs for prediction of gene structure in large genomic regions are currently available. Table 5 shows a list of available programs and Internet sites to access them. See Fickett (1996), Guigó (1997a) and Claverie (1997) for recent reviews, Burset and Guigó (1996) for a comparative benchmark, and the WWW document maintained by Wential Li at `linkage.rockefeller/edu/wli/gene/list.html` for an up-to-date list of references. At the core of all such programs, there exists one or more coding statistics related to the measures reviewed here. Indeed, currently more powerful programs are entirely built on Hidden Markov Models (GenScan (Burge and Karlin, 1997), Genie (Kulp et al., 1996), Veil (Henderson et al., 1997)), which can be seen as a generalization of the Markov Models discussed here.

A general strategy among gene identification programs is to integrate the output of a number of coding statistics. Thus, to name just a few examples, the popular Grail program (Uberbacher and Mural, 1991) uses a neural network to integrate a number of coding statistics, mostly related to Hexamer Usage and Position Asymmetry in base composition. Solovyev et al. (1994) in the Fgeneh program use linear discriminant analysis, while Dong and Searls (1994) in GenLang use linguistic methods. Although increased accuracy in the gene predictions is obtained in this way, because coding statistics are all essentially measuring codon usage bias in one way or another, their output is strongly correlated (Fickett and Tung, 1992). Indeed, Table 6 shows the correlation between the scores of the coding statistics reviewed here in the set of human exon and intron sequences. As it can be seen, the coding statistics are strongly correlated. The only two statistics truly uncorrelated are Codon Preference and Amino Acid Usage in exonic sequences, as otherwise expected. It appears, thus, that some combination of just two statistics, one measuring correlation between positions within a codon, and the other measuring dependence between codons along the query sequence could produce the most discriminant output.

	CU	HU	CPre	AAU	CPro	MM-1	MM-2	MM-5	PA	PAI	AMI	FOU
Codon Usage		0.908	0.769	0.492	0.803	0.876	0.909	0.772	0.590	0.558	0.529	0.562
Hexamer Usage	0.925		0.822	0.287	0.802	0.912	0.928	0.869	0.585	0.544	0.510	0.559
Codon Preference	0.927	0.932		-0.069	0.637	0.833	0.838	0.723	0.456	0.415	0.421	0.438
Amino Acid Usage	0.738	0.626	0.537		0.351	0.238	0.262	0.199	0.392	0.391	0.355	0.383
Codon Prototype	0.822	0.820	0.782	0.623		0.810	0.799	0.673	0.708	0.659	0.611	0.662
Markov Model, k=1	0.943	0.941	0.952	0.604	0.875		0.969	0.801	0.543	0.502	0.465	0.512
Markov Model, k=2	0.974	0.944	0.952	0.665	0.853	0.976		0.831	0.535	0.494	0.459	0.507
Markov Model, k=5	0.919	0.932	0.913	0.621	0.816	0.928	0.950		0.465	0.428	0.400	0.435
Position Assymetry	0.326	0.381	0.318	0.355	0.392	0.321	0.320	0.344		0.953	0.937	0.979
Periodic Assymetry Index	0.299	0.363	0.283	0.369	0.266	0.256	0.276	0.292	0.531		0.924	0.912
Average Mutual Information	0.225	0.267	0.216	0.247	0.314	0.217	0.228	0.251	0.873	0.469		0.911
Fourier Spectrum	0.381	0.455	0.373	0.432	0.414	0.363	0.370	0.401	0.914	0.713	0.726	

Table 6: Correlation between the different coding statistics in our set of exonic (upper triangle of the table) and intronic (lower triangle) sequences. The scores of the model dependent coding statistics and of the Fourier Spectrum have been divided by the length of the sequence to compute the correlations.

References

Borodovsky, M. and McIninch, J. 1993. Genmark: Parallel gene recognition for both dna strands. *Computer and Chemistry* 17:123–13.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic dna. *Journal of Molecular Biology* 268:78–94.

Burset, M. and Guigó, R. 1996. Evaluation of gene structure prediction programs. *Genomics* 34:353–357.

Claverie, J. M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics* 6:1735–1744.

Claverie, J.-M., Sauvaget, I., and Bougueleret, L. 1990. k-tuple frequency analysis: From intron/exon discrimination to t-cell epitope mapping. *Methods in Enzymology* 183:237–252.

Dong, S. and Searls, D. B. 1994. Gene structure prediction by linguistic methods. *Genomics* 23:540–551.

Fichant, G. and Gautier, C. 1987. Statistical method for predicting coding regions in nucleic acid sequences. *Nucleic Acids Research* 4:287–295.

Fickett, J. W. 1982. Recognition of protein coding regions in dna sequences. *Nucleic Acids Research* 10:5303–5318.

Fickett, J. W. 1996. Finding genes by computer: the state of the art. *Trends in Genetics* 12:316–320.

Fickett, J. W. and Tung, C. S. 1992. Assessment of protein coding measures. *Nucleic Acids Research* 20:6441–6450.

Gelfand, M. S. 1995. Prediction of function in dna sequence analysis. *Journal of Computational Biology* 1:87–115.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pave, A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research* 8:49–62.

Gribskov, M., Devereux, J., and Burgess, R. B. 1984. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Research* 12:539–549.

Groβe, I., Herzel, H., Buldyrev, S. V., and Stanley, H. E. 1998. A species-independent measure for distinguishing coding and noncoding dna. *submitted* .

Guigó, R. 1997a. Computational gene identification. *Journal of Molecular Medicine* 75:389–393.

Guigó, R. 1997b. Computational gene identification: An open problem. *Computers and Chemistry* 21:215–222.

Gutiérrez, G., Oliver, J., and Marín, A. 1994. On the origin of the periodicity of three in protein coding dna sequences. *Journal of theoretical Biology* 167:413–414.

Henderson, J., Salzberg, S., and Fassman, K. H. 1997. Finding genes in dna with a hidden markov model. *Journal of Computational Biology* 4:127–141.

Herzel, H. and Groβe, I. 1995. Measuring correlations in symbol sequences. *Physica A* 216:518–542.

Ikemura, T. 1985. Codon usage and trna content in unicellular and multicellular organisms. *Molecular Biology and Evolution* 2:13–34.

Konopka, A. K. 1994. Structure and Methods: VI. Human Genome Initiative and DNA Recombination, chapter Towards Mapping Functional Domains in Indiscriminantly Sequenced Nucleic Acids: A Computational Approach. Adenine Press, Guilderland, New York.

Kulp, D., Haussler, D., Reese, M., and Eeckman, F. H. 1996. A generalized hidden markov model for the recognition of human genes in dna. *In* D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith (eds.), Intelligent Systems for Molecular Biology, pp. 134–142, Menlo Park, California. AAAI press.

Li, W. 1997. The study of correlation structures of dna sequences: a critical review. *Computer and Chemistry* 21:257–271.

Li, W., Marr, T. G., and Kaneko, K. 1994. Understanding long-range correlations in dna sequences. *Physica D* 75:392–416.

McCaldon, P. and Argos, P. 1988. *Proteins: Structure, Function and Genetics* 4:99–122.

Mural, R. J., Mann, R. C., and Uberbacher, E. C. 1991. *In* C. C. Cantor and H. A. Lim (eds.), Proceedings of the First International Conference on Electrophoresis, Supercomputing and the Human Genome, pp. 164–172. World Scientific Co.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell Syst. The Bell System Technical Journal* 27:379–423.

Shepherd, J. C. 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proceedings National Academy Sciences USA.* 78:1596–1600.

Silverman, B. D. and Linsker, R. 1986. A measure of dna periodicity. *Journal of theoretical Biology* 118:295–300.

Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. 1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research* 22:5156–5163.

Staden, R. 1984. Measurements of the effects that coding for a protein has on a dna sequence and their use for finding genes. *Nucleic Acids Research* 12:551–567.

Staden, R. and McLachlan, A. 1982. Codon preference and its use in identifying protein coding regions in long dna sequences. *Nucleic Acids Research* 10:141–156.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., and Ramaswamy, R. 1997. Prediction of probable genes by fourier analysis of genomic sequences. *Computer Applications in the Biosciences* 13:263–270.

Tsonis, A. A., Elsner, J. B., and Tsonis, P. A. 1991. Periodicity in dna coding sequences: implications in gene evolution. *Journal of Theoretical Biology* 151:323.

Uberbacher, E. C. and Mural, R. J. 1991. Locating protein-coding regions in human dna sequences by a multiple sensor-neural network approach. *Proceedings National Academy Sciences USA* 88:11261–11265.