J. theor. Biol. (2000) **206,** 525–537
doi:10.1006/jtbi.2000.2144, available online at http://www.idealibrary.com on IDEAL®

Optimization of Coding Potentials Using Positional Dependence of Nucleotide Frequencies

Dirk Holste*, Ivo Grosse†‡, Sergey V. Buldyrev†, H. Eugene Stanley† and Hanspeter Herzel§

**Department of Theoretical Biophysics, Humboldt University Berlin, Invalidenstr. 42, D-10115, Berlin, Germany, † Center for Polymer Studies and Department of Physics, 590 Commonwealth Avenue, Boston University, Boston, MA 02215, U.S.A., ‡Institute for Molecular Biology and Biochemistry, Free University Berlin, Arnimallee 22, D-14195 Berlin, Germany and §Institute for Theoretical Biology, Humboldt University Berlin, Invalidenstr. 43, D-10115 Berlin, Germany*

(*Received* 22 *December* 1999, *Accepted in revised form* 17 *July* 2000)

We study the coding potential of human DNA sequences, using the positional asymmetry function (D_p) and the positional information function (I_q). Both D_p and I_q are based on the positional dependence of single nucleotide frequencies. We investigate the accuracy of D_p and I_q in distinguishing coding and non-coding DNA as a function of the parameters p and q, respectively, and explore at which parameters p_{opt} and q_{opt} both D_p and I_q distinguish coding and non-coding DNA most accurately. We compare our findings with classically used parameter values and find that optimized coding potentials yield comparable accuracies as classical frame-independent coding potentials trained on prior data. We find that p_{opt} and q_{opt} vary only slightly with the sequence length.

1. Introduction

Many sequence projects have turned from mapping to large-scale sequencing, including organisms from simple bacteria to complex vertebrates. Biochemical techniques on their own may not be adequate for annotating all genes in primary DNA sequences, and so they are customarily supported by computer-based predictions of genes (Fleischmann *et al.*, 1995; Nelson *et al.*, 1999). However, the reliable annotation of genes by statistical means remains a difficult problem in molecular biology (Fickett, 1996; Searls, 1998) as evidenced by the complete DNA sequences of human chromosomes 21 and 22 (Hattori *et al.*, 2000; Dunham *et al.*, 1999).

Genes of higher eukaryotes consist of coding regions (exons) that are interrupted by non-coding regions (introns). Exons and introns possess distinctive statistical patterns that distinguish coding and non-coding DNA. Conventional programs for gene-finding integrate heterogeneous types of biological information, referred to as the search by content and the search by signal. A third type of information refers to database similarity searchers. Gene-search by content is based on statistical general patterns of coding DNA regions. Gene search by signal is based on the detection of DNA binding sites and on other signals in the surrounding of a gene. In order to predict the most likely gene structure from a primary sequence, gene search by content is typically merged with the search by signal, using probabilistic models of DNA, discriminant analysis, or neural networks. Several statistical models have been applied in programs to the task of gene-identification, such as GeneID (Guigó *et al.*, 1992), GeneParser (Snyder & Stormo, 1993), GENMARK (Borodovsky & McIninch, 1993), GenLang (Dong & Searls, 1994), FGENEH (Solovyev *et al.*, 1994), GRAIL II (Xu *et al.*, 1994), MZEF (Zhang, 1997), GENSCAN (Burge & Karlin, 1997), GeneGenerator (Kleffe *et al.*, 1998), and GLIMMER (Salzberg *et al.*, 1998). The advantages and disadvantages of these programs and the application in conjunction have been evaluated (Gelfand, 1995; Fickett, 1996; Burset & Guigó, 1996; Claverie, 1997; Murakami & Takagi, 1998).

One well-known statistical pattern of exons is the existence of a reading frame and the unequal use of coding nucleotide triplets (codons). The reading frame induces a triplet periodicity in coding sequences, which is absent in non-coding sequences. The non-uniform codon usage gives rise to a different relative frequency $f(b|l)$ of each nucleotide $b =$ A, C, G, T, in a position $l \in (1, 2, 3)$ of the reading frame. Possible reasons for the non-uniformity of the codon usage are: (i) the non-uniform amino acid composition of proteins, (ii) the unequal number of codons encoding different amino acids, and (iii) the non-uniform distribution of synonymous codons encoding the same amino acid.

We study several coding potentials based on $f(b|l)$. The coding potential correlates with the likelihood that a certain region in DNA is protein-coding and builds the core of many gene-finding programs in order to find a rough location of open reading frames (ORFs). Evaluated coding potentials can be applied to DNA sequences without prior training. Beyond the statistical pattern $f(b|l)$ there exist further patterns, such as relationships between coding sequences and adjacent intergenic DNA (Bernardi, 2000). The inclusion of further biological information can improve accuracy in gene-finding (Guigó & Fickett, 1995; Burset & Guigó, 1996; Burge & Karlin, 1997).

We consider coding potentials which can be directly derived from a query sequence. A number of methods have been developed to calculate the coding potential based on $f(b|l)$, such as the prevalence for the occurrence of codons of the form purine–any nucleotide–pyrimidine (Shepherd, 1981), the non-uniform positional nucleotide usage (Fickett, 1982; Staden, 1984), the different G + C content (Bibb *et al.*, 1984), the detection of periodicities (Silverman & Linsker, 1986; Michel, 1986), the higher concentration of G in the first codon position (Trifonov, 1987), the positional dependence of entropy (Amalgor, 1985; Grosse *et al.*, unpublished data) or the correlation between nucleotide pairs (Grosse *et al.*, 2000a).

Many coding potentials are parameter-dependent. A systematic analysis of this parameter dependence has been left standing and this is the focus of this paper. We generalize classical coding potentials to the positional asymmetry function D_p and the positional information function I_q, and examine how accurately D_p and I_q can distinguish coding and non-coding DNA as a function of p and q, respectively. We search for values p_{opt} and q_{opt} for which both D_p and I_q yield the maximum accuracy, and compare p_{opt} and q_{opt} with classical parameters. At p_{opt} and q_{opt}, we find that D_p and I_q yield a comparable accuracy as traditional frame-independent coding potentials, most of which are trained on prior data and require a much higher number of input parameters. We base our studies on two standard data sets: (i) the benchmark data set of Fickett & Tung (1992) and (ii) the GenBank release 111.0 (Benson *et al.*, 1999). We also examine the dependence of the coding potentials studied here with respect to the A + T content of coding and non-coding DNA, since it has been shown that the accuracy of coding potentials can be affected by the A + T content (Guigó & Fickett, 1995; Burset & Guigó, 1996).

2. Coding Potentials

Consider a moving window of length $3N$ base pairs (bp) along a DNA sequence and decompose the window into N non-overlapping triplets. Denote the number of occurrences of base b at a given triplet position l by $N(b|l)$ and define the relative frequencies by $f(b|l) = N(b|l)/N$. Let the frame dependence matrix $\mathbf{F}$ be the 4×3 (base × position) element matrix in which each element contains $f(b|l)$. Since for each l normalization constrains each column of $\mathbf{F}$ to $\sum_{b=1}^{4} f(b|l) = 1$, only 9 out of 12 numbers are independent. Separate the mean frequency $f(b) = \sum_{l=1}^{3} f(b|l)/3$

from $f(b|l) = f(b) + d(b|l)$ as the elements of the vector $\mathbf{f}$. Keep the residuals $d(b|l)$ as elements of the matrix $\mathbf{D}$, which represent deviations of positional base compositions from occurrences expected by chance. We visualize the notations in the following sketch, in which the data are obtained from the coding sequence of the human beta-myosin heavy chain (HUMBMYH7) gene:

$$\overbrace{\text{atg }\underbrace{\text{ggagattcggagatggcagtctttggggctgccgccccctacctgcgcaagtcagag}}_{\text{Moving window of size 54 bp}}\text{ ...... ttgaatgaggag}}^{\text{Coding sequences}}$$

$$\underbrace{\text{gga gat tcg gag ... ... aag tca}}_{N = 18 \text{ triplets of the above window}}$$

We compute the 12 numbers $N(b|l)$ from the above window and display the frequencies $f(b|l)$, $f(b)$, and $d(b|l)$ (rounding off to 0.01):

$$\mathbf{F} = \begin{pmatrix} 0.11 & 0.22 & 0.17 \\ 0.17 & 0.39 & 0.33 \\ 0.50 & 0.17 & 0.33 \\ 0.22 & 0.22 & 0.17 \end{pmatrix},$$

$$\mathbf{f} = \begin{pmatrix} 0.17 \\ 0.30 \\ 0.33 \\ 0.20 \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} -0.06 & 0.05 & 0.01 \\ -0.13 & 0.09 & 0.04 \\ 0.17 & -0.16 & -0.01 \\ 0.02 & 0.02 & -0.04 \end{pmatrix}.$$

These matrices show, e.g. the excess (lack) of G in the first (second) codon position and the high G+C content in the third codon position. Next, we briefly discuss classical coding potentials. We illustrate each concept by using the values of $\mathbf{F}$, $\mathbf{f}$, and $\mathbf{D}$ above.

- *Base composition asymmetry* (Fickett, 1982). The coding potential is calculated from a linear weighted sum over the ratios of the maximal and minimal values of $f(b|l)$ as

$$A \equiv \sum_{b=1}^{4} \left\{ W(b) \frac{\max_{l \in (1,2,3)} \{f(b|l)\}}{\min_{l \in (1,2,3)} \{f(b|l)\} + 1/N} + w(b) f(b) \right\}. \quad (1)$$

The weights $W(b)$ and $w(b)$ are calculated from training sets of exons and introns. In order to construct a training-independent coding potential, we simplify eqn (1) and set all $W(b) \equiv 1$ and $w(b) \equiv 0$. If we substitute the values from $\mathbf{F}$, we obtain $A = 6.26$.

- *Uneven positional base frequencies* (Staden, 1984). The coding potential is calculated from the sum over the deviations $d(b|l)$ of the positional frequencies from the mean

$$D_1 \equiv \sum_{b=1}^{4} \sum_{l=1}^{3} |d(b|l)|. \quad (2)$$

The introduction of the index "1" will become clear in the context later on. Substituting the values from $\mathbf{D}$, we obtain $D_1 = 0.80$. To contrast this outcome with non-coding DNA, we make the simplifying assumption that each base b shows no dependence on the position l, such that $f(b|1) = f(b|2) = f(b|3)$. Hence, all $d(b|l) = 0$ due to the absence of any frame dependence and $D_1 = 0$.

- *Positional asymmetry* (Fickett & Tung, 1992). The coding potential is computed from the sum over the positional spread of $f(b|l)$ from $f(b)$:

$$D_2 \equiv \sum_{b=1}^{4} \sum_{l=1}^{3} d^2(b|l). \quad (3)$$

This coding potential is closely related to the one proposed by Staden (1984). We obtain

$D_2 = 0.09$ for **D** and $D_2 = 0$ for non-coding sequences.

- *Fourier transform* (Silverman & Linsker, 1986; Michel, 1986). The square of the Fourier transform (the power spectrum) is calculated for each base b binary translated DNA sequences. Assigning base $\tilde{b}$ at the nth sequence position $U_n(\tilde{b}, b) = \delta_{\tilde{b}, b}$ (where $\delta_{\tilde{b}b} = 1$ if base $\tilde{b} = b$, and 0 otherwise), the power spectrum is computed as

$$P(f_m) \equiv \sum_{b=1}^{4} \left| \frac{\sum_{n=1}^{3N} U_n(\tilde{b}, b)\, \mathrm{e}^{-i2\pi n f_m}}{3N} \right|^2, \qquad (4)$$

where $f_m = m/3N$ with $m = 1, \dots, 3N/2$. Commonly, $m = N$ is used to calculate the coding potential $P(\frac{1}{3})$. The full spectrum can be used to include effects of statistical noise (Tiwari *et al.*, 1997).

Interestingly, it has been observed (Guigó, 1999) that $P(\frac{1}{3})$ and D_2 have the same coding potential. Since $\sum_{n=1}^{3N} U_n(\tilde{b}, b) = 3N f(b)$, it can be analytically shown that by using $\mathrm{e}^{-i2\pi/3}$ as weights, $P(\frac{1}{3})$ is up to a constant equal to D_2 (Grosse, unpublished). Consequently, $P(\frac{1}{3})$ can be calculated from **F**.

- *Positional information* (Grosse *et al.*, unpublished data). For each position l, the positional entropy $H_1(l)$ can be calculated from $f(b|l)$ (Amalgor, 1985). However, the accuracy of $H_1(l)$ is limited (Fickett & Tung, 1992). By normalizing the entropy, we define the positional information I_1 as the difference between the entropy of the mean values H_1 [calculated form $f(b)$] and the average $\langle H_1(l) \rangle_l$ as

$$\begin{aligned} I_1 &\equiv H_1 - \frac{1}{3}\sum_{l=1}^{3} H_1(l) \\ &= -\sum_{b=1}^{4} f(b) \log_2 f(b) \\ &\quad + \frac{1}{3}\sum_{l=1}^{3}\sum_{b=1}^{4} f(b|l)\ \log_2 f(b|l). \end{aligned} \qquad (5)$$

Furthermore, if we introduce $f(l)$ as the relative frequency of the position l and use the joint frequency $f(b, l) = f(b|l)\ f(l)$, we can express I_1 as the *mutual information* between positional nucleotides (Grosse *et al.*, unpublished data)

$$\begin{aligned} I_1 &\equiv H_1 - \sum_{l=1}^{3} f(l) H_1(l) \\ &= \sum_{b=1}^{4}\sum_{l=1}^{3} f(b, l)\ \log_2 \left(\frac{f(b, l)}{f(b) f(l)} \right). \end{aligned} \qquad (6)$$

I_1 provides an intuitive meaning to the coding potential. Its outcome can be interpreted as the average mutual information in base b about the position l measured in units of bits. We calculate $I_1 = 0.12$ (bits) for **F** and $I_1 = 0$ (bits) for non-coding DNA, since $f(b, l)$ factorizes to $f(b) f(l)$.

- *Average mutual information* (Grosse *et al.*, 2000a, b). The mutual information $I(k)$ as a function of the base pair $(\tilde{b}, b)$ separated by a distance k is used. Under the simplifying assumption that the DNA sequence consists of statistically independent codons (Herzel & Grosse, 1995), the frequency $f_k(\tilde{b}, b)$ of base pairs $\tilde{b}$ and b in a distance k becomes a function of **F**

$$f_k(\tilde{b}, b) = \frac{1}{3}\begin{cases} f(\tilde{b}|1) f(b|1) + f(\tilde{b}|2) f(b|2) + f(\tilde{b}|3) f(b|3), & k = 3, 6, 9, \dots, \\ f(\tilde{b}|1) f(b|2) + f(\tilde{b}|2) f(b|3) + f(\tilde{b}|3) f(b|1), & k = 4, 7, 10, \dots, \\ f(\tilde{b}|1) f(b|3) + f(\tilde{b}|2) f(b|1) + f(\tilde{b}|3) f(b|2), & k = 5, 8, 11, \dots. \end{cases}$$

If we transpose the subscripts $\tilde{b}$ and b, we have $f_k(\tilde{b}, b) = f_{k+1}(b, \tilde{b})$ for distances $k = 4, 7, 10, \dots$ Hence, $I(k)$ assumes only two values: the in-frame (out-of-frame) mutual information I_{in} (I_{out}) for $k = 3, 6, 9, \dots$ ($k = 4, 5, 7, 8, \dots$). The average is used to define the coding potential as

$$\bar{I} \equiv \frac{I_{in} + 2\, I_{out}}{3}. \qquad (7)$$

Equation (7) quantifies the average mutual information shared by $\tilde{b}$ and b given the distance k between $\tilde{b}$ and b is a multiple of 3. We calculate $\bar{I} = 0.005$ (bits) for **F** and $\bar{I} = 0$ (bits) for non-coding DNA.

3. Accuracy of Positional Information

In this section, we illustrate the application of coding potentials. We evaluate the positional information I_1 for the two different data sets used throughout this study and examine the statistical dependence of coding potentials on different A+T content.

For each sequence from biologically known sets of exons and introns, we compute I_1 and thus obtain the I_1-histograms for exons and introns. The I_1-histograms overlap due to the finite sequence length. We evaluate the accuracy of a coding potential as follows:

1. Let the true positives (negatives), TP (TN), denote the fraction of coding (non-coding) sequences correctly predicted as coding (non-coding).
2. One determines the threshold above which a sequence is predicted as coding by imposing equal relative errors on the prediction of exons and introns, $TP = TN$.
3. One quantifies the accuracy (Fickett & Tung, 1992) of a coding potential as $(TP+TN)/2$, ranging from 1/2 (no discrimination) to 1 (exact discrimination).

To compare the accuracy of I_1 with the accuracy of other coding potentials, we analyse the standard data set of human DNA established by Fickett & Tung (1992). Since I_1 does not require prior training on organism-specific data, we compute the accuracy of I_1 for both the training set ($A_{training}$) and the test set (A_{test}). In order to test the robustness of I_1, we use an additional data set B of all human sequences from GenBank release 111.0 (Benson *et al.*, 1999). We identify coding DNA using the "CDS" key word in the GenBank flatfile format. We obtain non-overlapping coding and non-coding sequences of length L bp by partitioning all human sequences in GenBank 111.0 longer than L into sequences of length L, starting at the 5′-end (cf. Table 1).

Figure 1 shows the I_1-histograms for sets $A_{training}$, A_{test}, and B. We find that both coding and non-coding DNA have unimodal I_1-histograms with distinct maxima. We find that for each data set the histograms are significantly different for coding and non-coding DNA. In each data set, I_1-histograms for non-coding DNA are centered at significantly smaller values than the I_1-histograms of coding DNA. Figure 1 also shows that the I_1-histograms for sets $A_{training}$, A_{test}, and B are similar. Hence, the accuracy of I_1 is similar when evaluated on different sets of human DNA sequences. In Table 2, we show the accuracy for I_1, A, $\bar{I}$, D_1, and D_2 for sets $A_{training}$, A_{test}, and B. Table 2(a) shows the accuracy for coding potentials for the data sets $A_{training}$, A_{test}, and B for three different sequence lengths. Table 2(b) shows the accuracy for eight frame-independent coding potentials evaluated in Fickett & Tung (1992) as being the most accurate for A_{test}. We find that D_2, I_2, and coding potentials for p (q) adjacent to $p = q = 2$ are as accurate as the most effective classical measures after prior training on $A_{training}$. Fickett & Tung (1992) evaluate the accuracy of the entropy $H_1(l)$ (Amalgor, 1985) for $L = 108$ bp to be 63%. The inset in Fig. 1 shows the significantly higher accuracy of I_1 of 76% (Grosse *et al.*, unpublished data).

Figure 1 shows for coding sequences of set B a small shift of the coding I_1-histogram

TABLE 1

The number of coding and non-coding sequences in (1000s) in the data sets $A_{training}$, A_{test}, and B

Data set	Sequence class	54 bp	108 bp	162 bp	1080 bp
$A_{training}$	Coding	20.5	7.1	3.5	—
	Non-coding	125.1	58.1	36.5	—
A_{test}	Coding	22.9	8.2	4.3	—
	Non-coding	122.1	57.0	35.6	—
B	Coding	595.2	282.9	178.5	4.5
	Non-coding	171.7	81.1	51.1	16.0

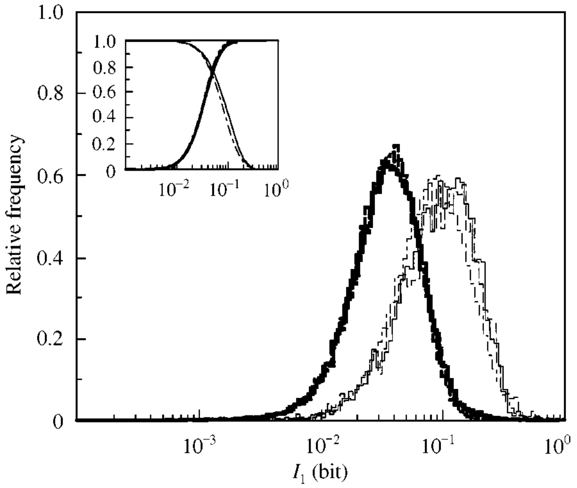

FIG. 1. Histograms of I_1 for human exons (——) and introns (——) of length 108 bp. The inset shows the cumulative histograms. While the values of I_1 vary from sequence to sequence, the I_1-histogram is almost the same for all three data sets. The I_1-histograms show an overlap of approximately 24% on $A_{training}$ and A_{test} and of 25% on set B: test set A (coding), ____; test set A (non-coding), ____; training set A (coding), _ _ _; training set A (non-coding), _ _ _; set B (coding), _ _ _; set B (non-coding), _ _ _.

towards non-coding, which leads to a slightly smaller accuracy of 75%. We examine whether the slight decrease in accuracy could be explained by A+T content variations between sets A and B. Figure 2 shows the A+T content of the sets A and B in conjunction with the mean and variance of I_1. A comparison of the top graphs shows that set B contains slightly more sequences with high A+T content than set A. The bottom graphs show that the accuracy of I_1 increases for sequences with low A+T content.

It is a general feature that many coding potentials show a dependence on the A+T content (Guigó & Fickett, 1995). Although I_1 is *a priori*-independent of the A+T content and shows no systematic dependence when it is applied to computer-generated Markov sequences (data not shown), it does show a dependence for experimental, coding DNA. Figure 2 shows that I_1 is almost independent of the A+T content for non-coding sequences, whereas it decays with increasing A+T content for coding sequences. Hence, we find that one possible explanation for the decrease in accuracy is indeed the difference in the individual A+T content of the A and B. Another possible explanation for the slight decrease in accuracy could be due to newer annotated contiguous sequences in GenBank, a number which stems from gene-finding programs and may still be putative.

By calculating the correlation coefficient $C(X, Y)$, we quantify the linear statistical dependence of I_1 $(=X)$ on the A+T content $(=Y)$. Analogously, we calculate the uncertainty coefficient $U(X, Y)$ to quantify the nonlinear statistical dependence of X on Y (see Appendix A). Table 3(a) shows $C(X, Y)$ and Table 3(b) shows $U(X, Y)$ for $X = I_1$, A, $\bar{\mathrm{I}}$, D_1, and D_2 vs. $Y = A + T$ applied to the data sets $A_{training}$, A_{test}, and B. Table 3 shows that most coding potentials have no distinct correlation on the A+T content for introns, while it shows clear linear anti-correlations and in general higher nonlinear correlations of coding potentials on the A+T content for exons than for introns. We find both $C(X, Y)$ and $U(X, Y)$ for most coding potentials higher for set A and set B. One possible explanation is the overall higher A+T content for set B than for set A.

4. Optimization of Coding Potentials

In this section, we generalize the coding potentials D_1, D_2 to D_p and I_1 to I_q, and study the accuracy of D_p and I_q as a function of p and q, respectively.

Consider D_1 and D_2 as two selected quantities of the generalized coding potential D_p, which we define as the *positional asymmetry function*

$$D_p \equiv \sum_{b=1}^{4} \sum_{l=1}^{3} |d(b|l)|^p. \tag{8}$$

The parameter p can take on any real number. D_p recovers the coding potential of Staden (1984) for $p = 1$ and the coding potential of Fickett & Tung (1992) for $p = 2$.

We generalize I_1 as follows. Recall that I_1 can be defined as the difference $H_1 - \langle H_1(l) \rangle_l$. According to Rényi (1970), there exists a natural extension of the ordinary Shannon entropy H_1 (Shannon, 1948) to the generalized entropies H_q. We define the Rényi entropies of $f(b)$ as (Rényi, 1970)

$$H_q \equiv \frac{1}{1-q} \log_2 \left(\sum_{b=1}^{4} f^q(b) \right), \tag{9}$$

TABLE 2

Comparison of classical coding potentials with D_p and I_q *for p* (q) = 1, 2, 3, *and* 4

(a) Coding potentials using positional dependence of nucleotide frequencies

Coding measure	Set $A_{training}$			Set A_{test}			Set B		
	54 bp (%)	108 bp (%)	162 bp (%)	54 bp (%)	108 bp (%)	162 bp (%)	54 bp (%)	108 bp (%)	162 bp (%)
A	68.9	75.9	79.6	68.3	75.1	79.3	66.9	74.2	79.3
D_1	69.9	76.6	80.9	69.4	76.3	79.8	68.0	75.1	79.4
D_2	70.2	76.8	80.8	70.0	76.6	80.1	68.0	75.5	80.5
D_3	69.8	76.7	80.4	69.9	76.7	80.3	68.1	75.3	80.3
D_4	69.5	76.4	80.1	69.5	76.3	80.0	67.9	75.1	80.0
$\bar{I}$	69.7	76.4	80.6	69.6	76.1	80.1	67.6	75.2	80.3
I_1	69.2	76.6	80.7	69.0	75.9	80.0	67.1	75.1	80.2
I_2	70.6	77.2	81.1	70.2	76.9	80.6	69.1	76.2	80.9
I_3	69.6	76.6	80.4	68.9	75.2	79.2	68.4	75.3	79.9
I_4	68.7	75.1	78.8	67.9	73.6	77.2	67.6	73.9	78.2

(b) Most accurate coding potentials (frame-independent)

Coding potential	No. of input parameters	Set A_{test}		
		54 bp (%)	108 bp (%)	162 bp (%)
Hexamer	4096	70.5	73.1	74.2
Positional symmetry	12	70.2	76.6	80.6
Dicodon usage	4096	70.2	72.9	73.9
Fourier	8	69.9	76.5	80.8
Hexamer-1	4096	69.9	72.6	73.8
Hexamer-2	4096	69.9	72.6	73.8
Run	6	66.6	70.3	71.3
Codon usage	64	65.2	68.0	69.5

where the parameter q can take on any real number. For large (small) q the entropies H_q are dominated by the most (least) $f(b)$. For $q = 0$, simply the number of non-vanishing frequencies $f(b)$ is counted. Analogously, we define H_q for $f(b|l)$. We substitute H_q for H_1 in eqn (5) and define the *positional information function* as

$$I_q \equiv H_q - \frac{1}{3}\sum_{l=1}^{3} H_q(l) = \frac{\log_2(\sum_{b=1}^{4} f^q(b))}{1-q} - \frac{1}{3}\sum_{l=1}^{3} \frac{\log_2(\sum_{b=1}^{4} f^q(b|l))}{1-q}. \qquad (10)$$

Equivalently, we can write the above expression using $f(l)$ as

$$I_q \equiv \frac{1}{1-q}\sum_{l=1}^{3} f(l)\log_2\left(\frac{\sum_{b=1}^{4} f^q(b)\,f^q(l)}{\sum_{b=1}^{4} f^q(b,l)}\right). \qquad (11)$$

For $q \to 1$, I_q recovers the positional information I_1. The generalizations D_p and I_q are in the center of this study. Let us consider the mechanism of how a parameter change affects I_q (the mechanism for D_p is similar). The parameter q accounts for two essential features: (i) the different characteristic patterns in the frame dependence matrix **F** (cf. Section 2) and (ii) the in general higher values of $I_{q=1}$ for coding sequences than for non-coding sequences (cf. Fig. 1). The role of q is that it can change the relative weight that each element in **F** contributes to I_q. By varying q over its parameter range, we can weight characteristics such as the G + C content of isochores or G in the first codon position. Therefore, we conduct a systematic analysis of I_q (D_p) and test whether there is the possibility that the accuracy of I_q (D_p) could be higher for parameter values that are different from the classical values $q = 1$ ($p = 1, 2$).

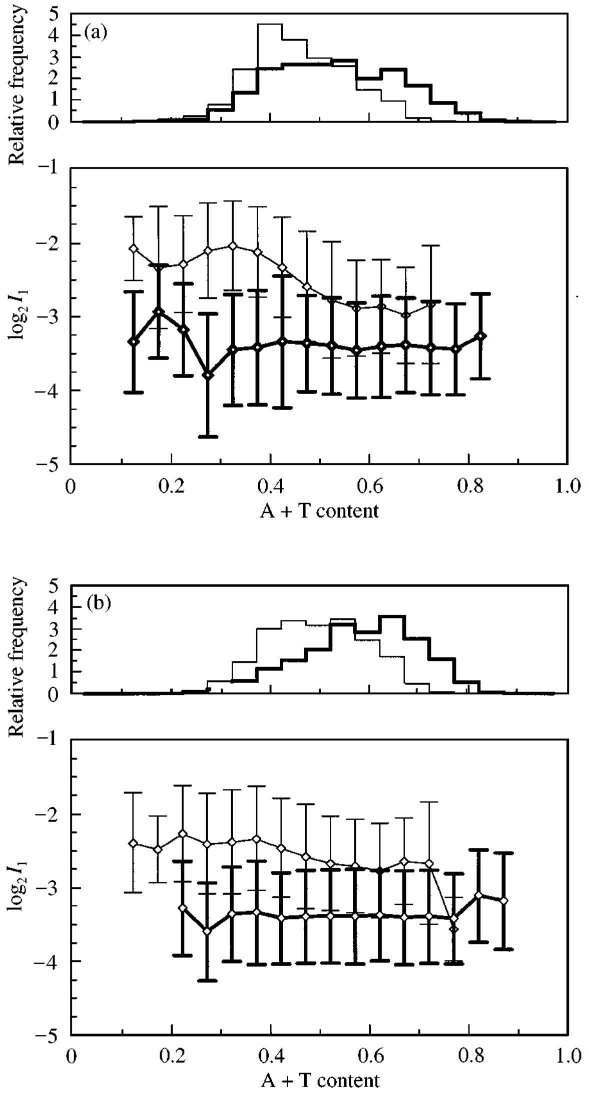

FIG. 2. Dependence of I_1 on the A+T content for exons (——) and introns (——). We analyse $\log_2 I_1$ for A_{test} (a) and B (b), because the I_1-histograms show broad tails (cf. Fig. 1). We compute the error bars by using subsets with 7000 coding and 7000 non-coding sequences. The top graphs show the A+T-histograms of exons (introns) with mean values 45% (52%) in set $A_{training}$ (data not shown), 43% (51%) in set A_{test}, and 47% (54%) in set B. The bottom graphs show the dependence of $\log_2 I_1$ on the A+T content by binning the A+T values to 20 bins and displaying the mean and standard deviation of $\log_2 I_1$ per bin. The overlap of the error bars indicates that the discrimination of exons and introns is less accurate for high A+T content.

To explore this possibility, we study the accuracy of D_p and I_q as a function of p (q) to determine the optimal parameter values p_{opt} and q_{opt} for $A_{training}$, A_{test}, and B. In order to compare the accuracy with benchmarked results, we choose the sequence lengths $L = 54$, 108, and 162 bp. Figure 3 (resp. Fig. 4) shows the accuracy for D_p (resp. I_q) as a function of $p\,(q) \in [-10, 10]$ for sequences of length 108 bp. We find that the accuracy of both D_p and I_q shows a strong dependence on p and q. Both D_p and I_q distinguish coding and non-coding DNA with significantly higher accuracy for $p, q > 0$ than for $p, q < 0$. As p increases beyond zero, the accuracy of D_p reaches its global maximum at about $p = 2$ and then only slightly decreases for $p \in [2, 10]$. The accuracy of I_q shows a clear maximum at about $q = 2$ and decays significantly as q increases. This finding is interesting as it states that the positional asymmetry D_2 (Fickett & Tung, 1992) is the most accurate coding potential among all D_p.

In Table 2, we compare the accuracy of D_p and I_q with the accuracy of other coding measures for sequences of length 54 and 162 bp. Table 2(a) shows that the accuracy by which D_2 and I_2 identify unannotated DNA is as high as the precision of those methods listed in Table 2(b) which are trained on the DNA to be analysed. Table 2(b) also shows the number of learned input parameters required for the discrimination (somewhat variable for Fourier and Run). Since D_p (I_q) depend only on the 12 frequencies $d(b|l)$ ($f(b|l)$), generalized coding potentials can be usefully studied for sequence lengths as short as 50 bp.

Finally, we examine for parameter values $p(q) \in [0, 4]$ the statistical dependence of D_p and I_q on the A+T content. Table 3 shows for D_p and I_q for $p(q) = 1, 2, 3$, and 4 in Table 3(a) linear and in Table 3(b) nonlinear statistical dependences of D_p and I_q on the A+T content. We find that the dependence of D_p and I_q on the A+T content is similar to the dependence of classical coding potentials on the A+T content. D_p shows no distinct correlations for non-coding DNA but clear anti-correlations for coding DNA. The statistical dependence of I_q on the A+T content is more complex. Table 3 shows for I_q less linear correlations to coding DNA than D_p to coding DNA but more nonlinear correlations to non-coding DNA than D_p to non-coding DNA, when q and p are increased.

TABLE 3

Dependence of classical coding potentials and D_p and I_q for $p(q) = 1, 2, 3,$ and 4 on the $A+T$ content. (a) shows linear and (b) shows nonlinear statistical dependences of coding potentials on the $A+T$ content for $A_{training}$, A_{test}, and B.

(a) Correlation coefficient

Data set	Sequence class	A	D_1	D_2	D_3	D_4	$\bar{I}$	I_1	I_2	I_3	I_4
$A_{training}$	Coding	− 0.39	− 0.31	− 0.31	− 0.31	− 0.31	− 0.35	− 0.36	− 0.20	− 0.08	0.00
	Non-coding	0.01	− 0.02	− 0.03	− 0.03	− 0.03	0.00	0.00	− 0.04	− 0.05	− 0.06
A_{test}	Coding	− 0.42	− 0.33	− 0.35	− 0.36	− 0.35	− 0.37	− 0.38	− 0.22	− 0.07	− 0.02
	Non-coding	− 0.03	− 0.05	− 0.05	− 0.05	− 0.05	− 0.04	− 0.04	− 0.05	− 0.03	− 0.02
B	Coding	− 0.23	− 0.20	− 0.20	− 0.20	− 0.20	− 0.21	− 0.22	− 0.14	− 0.07	− 0.03
	Non-coding	0.03	− 0.04	− 0.04	− 0.04	− 0.04	− 0.01	0.00	− 0.13	− 0.20	− 0.23

(b) Uncertainty coefficient

Data set	Sequence class	A	D_1	D_2	D_3	D_4	$\bar{I}$	I_1	I_2	I_3	I_4
$A_{training}$	Coding	0.06	0.04	0.04	0.04	0.04	0.05	0.05	0.03	0.03	0.03
	Non-coding	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04
A_{test}	Coding	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.03	0.02	0.02
	Non-coding	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04
B	Coding	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.01	0.02	0.02
	Non-coding	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04

We randomly choose for each class a subset of 7000 sequences of exons and introns of length 108 bp. We (a) rank the data and calculate $C(X, Y)$, and we (b) distribute the data on an array covered by $M = 4 \times 4$ bins and calculate the $U(X, Y)$ rounding off to $|0.01|$.

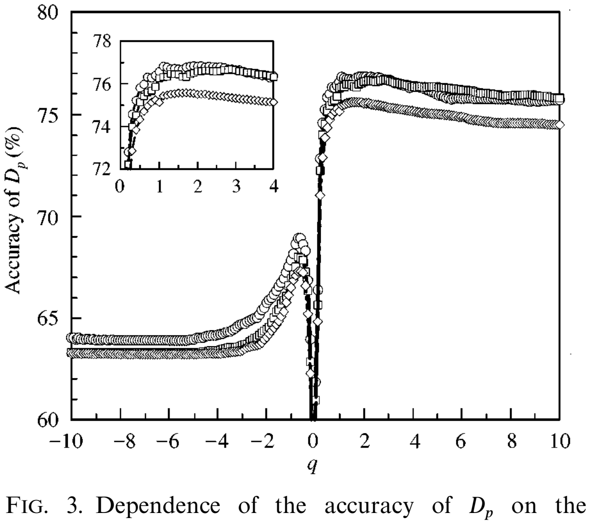

FIG. 3. Dependence of the accuracy of D_p on the parameter $p \in [-10, 10]$ for sets $A_{training}$, A_{test}, and B. The inset shows the region $p \in [0, 4]$. The accuracy of D_p shows a strong dependence on the parameter p, dropping to nearly 50% (no discrimination) when passing through $p = 0$. The accuracy of D_p shows two maxima, one for $p < 0$ and one for $p > 0$, and the accuracy of D_p is almost constant for $p > p_{opt}$: training set A, ○—○; test set A, □—□; set B, ◇- -◇.

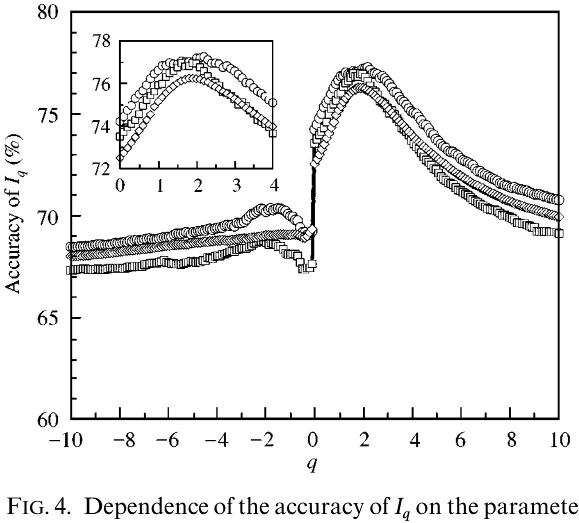

FIG. 4. Dependence of the accuracy of I_q on the parameter q for sets $A_{training}$, A_{test}, and B. The inset shows the region $q \in [0, 4]$. The accuracy of I_q shows a clear maximum at q_{opt} and decays significantly for $q > q_{opt}$: training set A, ○—○; test set A, □—□; set B, ◇- -◇.

5. Length Dependence of Accuracy

We investigate the length dependence of the accuracy of D_p and I_q, and examine the dependence $p_{opt}(N)$ and $q_{opt}(N)$ on N. We partition all human sequences of set B longer than 1080 bp into sequences of non-overlapping coding and non-coding sequences of length 1080 bp and study the accuracy for sequence sizes ranging from 27 to 1080 bp. We allow only such sequence sizes which are integer fractions of 1080 bp, so that for each L exactly the same number of nucleotides are studied and no nucleotide is left over when partitioning the sequences of 1080 bp into the sequences of sizes L. We keep for each length the overall number of triplets N within the set constant. Figure 5 and 6 show that both $p_{opt}(N)$ and $q_{opt}(N)$ are almost independent of N. As fluctuations increase for small sequence lengths ($N \leqslant 20$), the accuracy does not abruptly change with changing $p(q)$ to values adjacent to p_{opt} (q_{opt}).

We also find that the sharp peak of the accuracy of I_q becomes less pronounced with increasing N (cf. Fig. 6). As such, the use of q_{opt} plays a more important role when using small sequence sizes. This situation is relevant for human DNA sequences, which have an average exon length of 150 bp (Deutsch & Long, 1999).

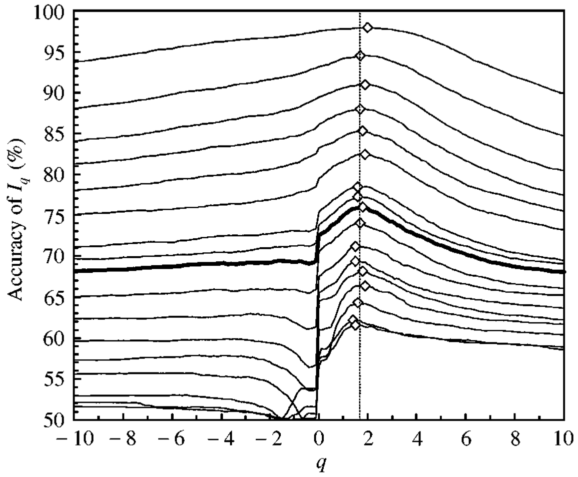

FIG. 6. Accuracy of I_q as a function of the parameter q for different sequence lengths for set B (cf. Fig. 5). The value q_{opt} (◇) at which the accuracy of I_q assumes its maximum does not undergo large variations when varying the sequence length. The ---- shows the mean value $\langle q_{opt} \rangle = 1.7 \pm 0.2$ of all optimal q_{opt}: set B, ——; set B (108 bp), ▬; maxima, ◇.

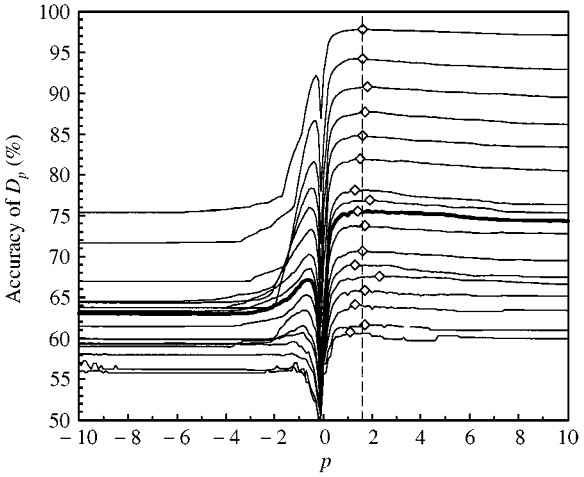

FIG. 5. Accuracy of D_p as a function of the parameter p for different sequence lengths for set B (from bottom to top: 27, 30, 36, 45, 54, 60, 72, 90, 108, 120, 135, 180, 216, 270, 360, 540, and 1080 bp). The accuracy of D_p increases with increasing sequence length. The value p_{opt} (◇) at which the accuracy of D_p assumes its maximum does not undergo large variations when varying the sequence length. The broken line shows the mean value $\langle p_{opt} \rangle = 1.6 \pm 0.3$ of all optimal p_{opt}: set B, ——; set B (108 bp), ▬; maxima, ◇.

6. Conclusions

This study addressed the problem of possible new parameters that distinguish coding and non-coding DNA more accurately. We generalized the coding potentials D_1 and D_2 to the positional asymmetry function, D_p, and the coding potential I_1 to the positional information function, I_q. We study the accuracy of D_p and I_q for $p(q) \in [-10, 10]$, and we search for those values p_{opt} and q_{opt} for which the accuracy is maximal. By plotting the accuracy as a function of $p(q)$, we find that both functions show a strong dependence on p and q, respectively (cf. Figs 3 and 4).

We find that D_p and I_q can enhance the contribution of strong characteristics—such as the preference of G in the first codon position and the high G+C content in the third codon position—and can thus distinguish coding and non-coding DNA significantly more accurate for p, $q > 0$ than for p, $q < 0$. It is an interesting result that a “quadratic weighting” (i.e., $p = q = 2$) of $d(b|l)$ or $f(b|l)$ maximizes the accuracy of D_p and I_q. In this parameter region, D_p or

I_q are as accurate as frame-independent coding potentials bench-marked in Fickett & Tung (1992) as being the most accurate for sequence lengths $L = 54$, 108, and 162 bp. This finding is intriguing, since most conventional coding potentials require training on organism-specific data sets whereas both D_p and I_q do not. It also indicates that there are differences in coding and non-coding DNA which are related to generalized entropies (distances) I_q (D_p) of $q = 2$ ($p = 2$). We emphasize that we restrict ourselves in this study to the analysis of coding potentials based on the frame dependence of single nucleotides. Beyond single nucleotides, one could also define generalized coding potentials based on di- and trinucleotide frequencies. While coding potentials like Hexamer exploit oligonucleotide frequencies, Table 2 shows that D_2 and I_2 distinguish coding and non-coding DNA with similar accuracy as Hexamer. Moreover, since both p_{opt} and q_{opt} show only a slight dependence on the sequence size N, there is no need for fine-tuning.

In order to gain some insight into the robustness of generalized coding potentials, we apply D_p and I_q to two different data sets A and B, and find that the accuracy differs only slightly (cf. Fig. 1). Since the data sets A and B show different variations in the A+T content, we also study the linear and nonlinear statistical dependences of coding potentials on the A+T content in sets A and B. The conjunction of linear and nonlinear analyses shows for most coding potentials no distinct correlations on the A+T content for non-coding DNA, while it shows clear linear anti-correlations and in general higher nonlinear correlations for coding DNA than for non-coding DNA. The applied combination of linear and nonlinear analysis can be easily used to detect statistical dependences of other coding potentials. We stress that D_p and I_q have the advantage that they can be applied without prior training, and that they can be supplemented by the search for biological signals such as start and stop codons, splice sites, promoter segments and so forth.

In conclusion, our study shows that differences in the coding potential of coding and non-coding DNA can be optimally extracted, and that this generalized approach could provide a firmer grounding for the application of coding potentials to recognize coding sequences in novel DNA for which training sets are not yet available.

We thank Lev Levitin and Olaf Weiss for stimulating discussions, the referees for valuable comments, and the Deutsche Forschungsgemeinschaft (DFG), the Graduate Programme "Dynamics and Evolution" (DFG-GK 268), and the National Institutes of Health for financial support.

REFERENCES

AMALGOR, H. (1985). Nucleotide distribution and the recognition of coding regions in DNA sequences: an information theory approach. *J. theor. Biol.* **117,** 127–136.

BENSON, D. A., BOGUSKI, M. S., LIPMAN, D. J., OSTELL, J., OUELETTE, B. F. F., RAPP, B. A. & WHELLER, D. L. (1999). GenBank. *Nucleic Acids Res.* **27,** 12–17.

BERNARDI, G. (2000). Isochores and the evolutionary genomics of vertebrates. *GENE* **241,** 3–17.

BIBB, M. L., FINDLAY, P. R. & JOHNSON, M. W. (1984). The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *GENE* **30,** 157–166.

BORODOVSKY, M. & MCININCH, J. (1993). GENMARK: parallel gene recognition for both DNA strands. *Comput. Chem.* **17,** 123–133.

BURGE, C. & KARLIN, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94.

BURSET, M. & GUIGÓ, R. (1996). Evaluation of gene structure prediction programs. *Genomics* **34,** 353–367.

CLAVERIE, J.-M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6,** 1735–1744.

DEUTSCH, M. & LONG, M. (1999). Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27,** 3219–3228.

DONG, S. & SEARLS, D. B. (1994). Gene structure prediction by linguistic methods. *Genomics* **23,** 540–551.

DUNHAM, I., SHIMIZU, N., ROE, B. A., CHISSOE, S. *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402,** 489–495.

FICKETT, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* **10,** 5303–5318.

FICKETT, J. W. (1996). Finding genes by computer: the state of the art. *Trends Genet.* **12,** 316–320.

FICKETT, J. W. & TUNG, C.-S. (1992). Assessment of protein coding measures. *Nucleic Acids Res.* **20,** 6441–6450.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J., DOUGHERTY, B. A., MERRICK, J. M. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512.

GELFAND, M. S. (1995). Prediction of function in DNA sequence analysis. *J. Comput. Biol.* **2,** 87–115.

GROSSE, I., BULDYREV, S. V., STANLEY, H. E., HOLSTE, D. & HERZEL, H. (2000a). Average mutual information of coding and non-coding DNA. *Pacific Symp. Biocomput. 2000* **5,** 611–620.

GROSSE, I., HERZEL, H., BULDYREV, S. V. & STANLEY, H. E. (2000b). Species independence of mutual information in coding and non-coding DNA. *Phys. Rev. E* **61,** 5624–5629.

GUIGÓ, R. (1999). DNA composition, codon usage and exon prediction. *Nucleic Acid and Protein Databases* (Bishop, M. ed.), New York: Academic Press.

GUIGÓ, R. & FICKETT, J. W. (1995). Distinctive sequence features in protein coding, genic non-coding, and intergenic human DNA. *J. Mol. Biol.* **253,** 51–60.

GUIGÓ, R., KNUDSEN, S., DRAKE, N. & SMITH, T. (1992). Prediction of gene structure. *J. Mol. Biol.* **226,** 141–157.

HATTORI, M., FUJIYAMA, A., TAYLOR, T. D., WATANABE, H., YADA, T., PARK H. S., TOYADA, A., ISHII, K., TOTOKI, Y. *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* **405,** 311–319.

HERZEL, H. & GROSSE, I. (1997). Correlations in DNA sequences: the role of protein coding sequences. *Phys. Rev. E* **55,** 800–810.

KLEFFE, J., HERMAN, K., VAHRSON, W., WITTIG, B. & BRENDEL, V. (1998). GeneGenerator: a flexible algorithm for gene prediction and its application to maize sequences. *Bioinformatics* **14,** 232–243.

MICHEL, C. J. (1986). New statistical approach to discriminate protein coding from noncoding DNA regions in DNA sequences and its evaluation. *J. theor. Biol.* **120,** 223–236.

MURAKAMI, K. & TAKAGI, T. (1998). Gene recognition by combination of several genefinding programs. *Bioinformatics* **14,** 665–675.

NELSON, K. E., CLAYTON, R. A., GILL, S. R., GWINN, M. L., DODSON, R. J., HAFT, D. H., HICKEY, E. K., PETERSON, J. D., NELSON, W. C., KETCHUM, K. A., *et al.* (1999). Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima. Nature* **399,** 323–329.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLIN, W. T. & FLANNERY, B. P. (1992). *The Art of Scientific Computing: Numerical Recipes in C*, 2nd Edn. U.S.A.: Cambridge, Cambridge University Press.

RÉNYI, A. (1970). *Probability Theory*. Amsterdam: North-Holland.

SACHS, L. (1984). *Applied Statistics*. New York: Springer-Verlag.

SALZBERG, S. L., DELCHER, A. L., KASIF. S. & WHITE, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **15,** 544–548.

SEARLS, D. B. (1998). Grand challenges in computational biology. *Computational Methods in Molecular Biology* (Salzberg, S. L., Searls, D. B. & Kasif, S., eds). Amsterdam, The Netherlands: Elsevier Science B. V.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell. Syst. Tech. J* **27,** 379–423.

SHEPHERD, J. W. C. (1981). Method to determine the reading frame of a protein from the purine-pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci. U.S.A.* **78,** 1590–1600.

SILVERMAN, B. D. & LINSKER, R. (1986). A measure of DNA periodicity. *J. theor. Biol.* **118,** 295–300.

SNYDER, E. E. & STORMO, G. D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res.* **21,** 607–613.

SOLOVYEV, V. V., SALAMOV, A. A. & LAWRENCE, C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **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 Res.* **12,** 551–556.

TIWARI, S., RAMACHANDRAN, S., BHATTACHARYA, A., BATTACHARYA, S. & RAMASWAMY, R. (1997). Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci.* **13,** 263–270.

TRIFONOV, E. N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. *J. Mol. Biol.* **194,** 643–652.

XU, Y., EINSTEIN, J. R., MURAL, R. J., SAHA, M. & UBERBACHER, E. C. (1994). An improved system for exon recognition and gene modeling in human DNA sequences. *Proc. 2nd Int. Conf. Intell. Systems Mol. Biol.* Menlo Park, CA: AAAI Press.

ZHANG, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 565–568.

APPENDIX A

In the appendix, we discuss the calculation of the correlation coefficient, $C(X, Y)$, and the uncertainty coefficient, $U(X, Y)$.

A.1. CALCULATION OF CORRELATION COEFFICIENT

We quantify linear statistical dependences by calculating the correlation coefficient $C(X, Y)$ of two samples of measurement X and Y. $C(X, Y)$ ranges from -1 to 1, where $(-)$ 1 corresponds to perfect sample (anti-)correlation, and 0 is the value for linearly independent samples.

For two data samples of size S, we calculate the mean values $\bar{x} = \langle x \rangle_S$ and $\bar{y} = \langle y \rangle_S$ by $\langle x \rangle_S = \sum_{s=1}^{S} x_s/S$ and $\langle y \rangle_S = \sum_{s=1}^{S} y_s/S$, and the covariance $\sigma_{XY}^2 = \langle (x - \bar{x})(y - \bar{y}) \rangle_S$. One defines the correlation coefficients $C(X, Y)$ as (e.g. Sachs, 1984)

$$C(X, Y) = \frac{\sigma_{XY}^2}{\sigma_{XX}\sigma_{YY}}.$$

We use rank numbers instead of measurements, so that $C(X,Y)$ becomes independent of monotonic scaling. We obtain rank numbers through $X' = \{r_i(x)\}$, where $r_i(x)$ is the rank of the i-th element of the original data sample permutated according to $r_i(x) = \# \{j | x_j \leqslant x_i, 1 \leqslant j \leqslant S\}$.

A.2. CALCULATION OF UNCERTAINTY COEFFICIENT

We use the uncertainty coefficient $U(X, Y)$ to quantify nonlinear statistical dependences. We calculate the entropy of a M-bin discretized measurement by $H(X) = -\sum_{m=1}^{M} p_m \log_2 p_m$. One defines the uncertainty coefficient $U(X, Y)$ as (e.g. Press *et al.*, 1992)

$$U(X, Y) = 2\,\frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)}.$$

$U(X, Y)$ is the mutual information of measurement X about measurement Y (Shannon, 1948) normalized to the marginal entropies. We compute the entropies $H(X)$, $H(Y)$, and the joint entropy $H(X, Y)$ as follows:

1. Distribute X and Y on an array consisting of $M \times M$ bins. Choose the bins in such a way that the marginal probabilities $Pr(X = x_m) = p_m$ and $Pr(Y = y_n) = p_n$ are uniform. We have $H(X) = H(Y) = \log_2 M$, since $p_m = p_n = 1/M$.
2. Determine the joint probabilities, $Pr\{X = x_m, Y = y_n\} = p_{m,n}$ from the distribution of X and Y on the array of M^2 bins.
3. Compute $U(X, Y) = 2 - H(X, Y)/\log_2 M$, because $H(X) = H(Y) = \log_2 M$.

$U(X, Y)$ ranges from 0 to 1, where 0 corresponds to the case in which X and Y are statistically independent, and 1 to the case in which X and Y are functionally dependent. $U(X, Y)$ is related to $C(X, Y)$, and it can be analytically shown that weak correlations lead to $U(X, Y) \propto C^2(X, Y)$ (Herzel & Grosse, 1997).