Pacific Symposium on Biocomputing 5:611-620 (2000)

AVERAGE MUTUAL INFORMATION OF CODING AND NONCODING DNA

IVO GROSSE, SERGEY V. BULDYREV, H. EUGENE STANLEY
Boston University, Center for Polymer Studies and Department of Physics, Boston, MA 02215, U.S.A.

DIRK HOLSTE, HANSPETER HERZEL
Humboldt University Berlin, Institute for Theoretical Biology and Theoretical Biophysics Group, Invalidenstr. 43, D–10115, Berlin, Germany

One basic problem in the analysis of DNA sequences is the recognition of protein-coding genes. Computer algorithms to facilitate gene identification have become important as genome sequencing projects have turned from mapping to large-scale sequencing, resulting in an exponentially growing number of sequenced nucleotides that await their annotation. Many statistical patterns have been discovered that are different in coding and noncoding DNA, but most of them vary from species to species, and hence require prior training on organism-specific data sets. Here, we investigate if there exist *species-independent* statistical patterns that are different in coding and noncoding DNA. We introduce an information-theoretic quantity, the average mutual information (AMI), and we find that the probability distribution functions of the AMI are significantly different in coding and noncoding DNA, while they are almost identical for different species. This finding suggests that the AMI might be useful for the recognition of protein-coding regions in genomes for which training sets do not exist.

1 Introduction

DNA carries the genetic information of many living organisms, and one goal of genome projects is to extract that information. One basic problem in the analysis of DNA sequences is the recognition of genes. Since experimental techniques alone are not appropriate for recognizing all genes in every genome, computer algorithms are useful for predicting the location of genes[1,2,3,4,5]. Gene-finding programs combine the search for biological signals with the analysis of statistical patterns that are different in coding and noncoding DNA. Many such patterns have been discovered[6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21], but most of them vary from species to species. This species dependence requires traditional gene identification programs be trained on organism-specific data sets before they can be applied to search for genes in un-annotated DNA sequences. Here we investigate whether there exist *species-independent* statistical patterns that are different in coding and noncoding DNA.

2 Mutual Information Function

In search for such patterns, we study for coding and noncoding DNA the *mutual information function*[22,23,24,25]

$$I(k) \equiv \sum_{i,j=1}^{4} P_{ij}(k) \cdot \log_2 \left(\frac{P_{ij}(k)}{p_i \cdot p_j} \right), \tag{1}$$

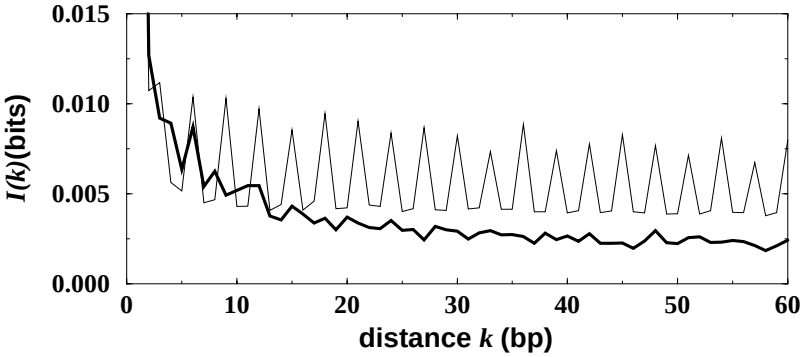
where p_i denotes the probability of finding the nucleotide $n_i \in \{A, C, G, T\}$, and $P_{ij}(k)$ denotes the probability of finding the pair of nucleotides n_i and n_j separated by a gap of length k. $I(k)$ quantifies the amount of information (in units of bits) that one obtains about nucleotide Y by learning the identity of nucleotide X a distance k away.

The following two examples may serve to illustrate the intuitive meaning of $I(k)$. Consider a random, uncorrelated sequence, in which each nucleotide occurs independently of any other nucleotide in the sequence. Intuitively it is clear that we cannot obtain any information from any nucleotide X about any nucleotide Y, so $I(k)$ should be zero for all distances k. Indeed $I(k) = 0$ for all k according to Eq. (1), since the statement that all nucleotides are statistically independent can be mathematically formulated by the set of equalities: $P_{ij}(k) = p_j \cdot p_j$ for all i, j, and k. From these equalities it follows that all the logarithms appearing in Eq. (1) are zero, and hence the sum in Eq. (1) is equal to zero.

As a second example consider a sequence in which each nucleotide occurring with equal probability $p_i = 1/4$ is determined by the previous nucleotide. In this case we will be able to *determine* the identity of nucleotide Y by learning the identity of X. Intuitively we say we obtain an information of 2 bits about Y by learning the identity of X. Indeed $I(k) = 2$ by Eq. (1), so again Eq. (1) agrees with our intuition. For quaternary sequences $I(k)$ always ranges from 0 to 2, and for most DNA sequences $I(k)$ is close to 0, which states that in a typical DNA sequence the information in nucleotide X about nucleotide Y is small. If $I(k)$ is monotonically decreasing with k, it means that the information in nucleotide X about nucleotide Y gets smaller as the distance k between X and Y increases.

Fig. 1 shows $I(k)$ for coding and noncoding DNA of animals extracted from GenBank[26] release 111. We find that for noncoding DNA $I(k)$ decays to zero, while for coding DNA $I(k)$ oscillates between two values, the *in-frame* mutual information, I_{in}, at distances k that are multiples of 3, and the *out-of-frame* mutual information, I_{out}, at other k. The period-3 oscillations in coding DNA originate from the presence of the genetic code, and from the nonuniformity of the codon frequency distribution[27,28,29].

Figure 1: Mutual information function, $I(k)$, of coding DNA (thin line) and noncoding DNA (thick line) from the set of animal DNA in GenBank 111. We extract from files `gbpri1.seq`, `gbpri2.seq`, `gbpri3.seq`, `gbrod.seq`, `gbmam.seq`, `gbvrt.seq`, `gbinv1.seq`, and `gbinv2.seq`, all protein-coding exons and all introns with a minimum length of 300 bp, and we cut these sequences into non-overlapping fragments of length 300 bp, starting at the 5'-end. We compute the mutual information function of each fragment, and display the average over all mutual information functions (of exons and introns separately). While $I(k)$ for noncoding DNA decays monotonically to zero as k increases, $I(k)$ of coding DNA shows persistent period-3 oscillations.



3 Average Mutual Information (AMI)

The decay of $I(k)$ for noncoding DNA and the decay of the envelope of $I(k)$ for coding DNA indicate the existence of long-range correlations. However, even when neglecting those correlations, the remaining statistical patterns are strong enough to distinguish coding from noncoding DNA. Neglecting weak codon-codon correlations, the joint probabilities $P_{ij}(k)$ can be computed in terms of the 12 *positional nucleotide probabilities*, $p_i^{(m)}$, of finding the nucleotide n_i at position $m \in \{1,2,3\}$ in an arbitrarily chosen reading frame as follows[6,30]:

$$P_{ij}(k) = \frac{1}{3} \cdot \begin{cases} p_i^{(1)} \cdot p_j^{(1)} + p_i^{(2)} \cdot p_j^{(2)} + p_i^{(3)} \cdot p_j^{(3)} & \text{for} \quad k = 3,6,9,... \\ p_i^{(1)} \cdot p_j^{(2)} + p_i^{(2)} \cdot p_j^{(3)} + p_i^{(3)} \cdot p_j^{(1)} & \text{for} \quad k = 4,7,10,... \\ p_i^{(1)} \cdot p_j^{(3)} + p_i^{(2)} \cdot p_j^{(1)} + p_i^{(3)} \cdot p_j^{(2)} & \text{for} \quad k = 5,8,11,... \end{cases} . \quad (2)$$

Since the second and the third line in Eq. (2) differ only by a permutation of subscripts i and j, $I(k)$ assumes only two different values, namely I_{in} and I_{out}. We sample the $p_i^{(m)}$ from each sequence, compute $P_{ij}(k)$ from $p_i^{(m)}$ by using Eq. (2), and then compute $I_{\text{in}} \equiv I(3)$ and $I_{\text{out}} \equiv I(4) = I(5)$ from $P_{ij}(k)$.

We find that $\ln(I_{\rm in})$ and $\ln(I_{\rm out})$ are almost linearly dependent, and are thus highly correlated (correlation coefficient $C = 0.96$ for both coding and noncoding DNA). This simplifies the question of how to combine $I_{\rm in}$ and $I_{\rm out}$ into a single quantity, as almost any combination will yield approximately the same accuracy. For the sake of easy interpretation, we choose a simple linear combination and define the *average mutual information*[31]

$$\mathrm{AMI} \equiv I_{\rm in} \cdot P_{\rm in} + I_{\rm out} \cdot P_{\rm out} \ , \tag{3}$$

where $P_{\rm in} \equiv 1/3$ and $P_{\rm out} \equiv 2/3$ denote the occurrence probabilities of $I_{\rm in}$ and $I_{\rm out}$. When choosing $P_{\rm in}$ and $P_{\rm out}$ in this way, the AMI quantifies the average amount of information (in units of bits) that one obtains about nucleotide Y by learning both the identity of nucleotide X and if the distance k between X and Y is a multiple of 3.

The practical implementation of the algorithm looks as follows:

1. Count the number of occurrences of nucleotide $n_i \in \{A, C, G, T\}$ in position $m \in \{1, 2, 3\}$ of an arbitrarily chosen reading frame in a given DNA sequence of length[a] N. Denote that number by $N_i^{(m)}$.
2. Divide $N_i^{(m)}$ by $N/3$, the total number of nucleotides occurring in position m, and define the *positional nucleotide frequency* $p_i^{(m)} \equiv 3 \cdot N_i^{(m)}/N$. Note that the positional nucleotide frequencies are normalized to 1, that is $\sum_{i=1}^{4} p_i^{(m)} = 1$ for all m.
3. Compute $P_{ij}(3)$ and $P_{ij}(4)$ from $p_i^{(m)}$ by using Eq. (2).
4. Define $p_i \equiv \sum_{m=1}^{3} p_i^{(m)}/3$, which is the overall, normalized frequency of nucleotide n_i.
5. Compute $I(3)$ and $I(4)$ from $P_{ij}(3)$, $P_{ij}(4)$, and p_i by using Eq. (1). Define $I_{\rm in} \equiv I(3)$ and $I_{\rm out} \equiv I(4)$ as well as $P_{\rm in} \equiv 1/3$ and $P_{\rm out} \equiv 2/3$.
6. Compute the average mutual information (AMI) by using Eq. (3).

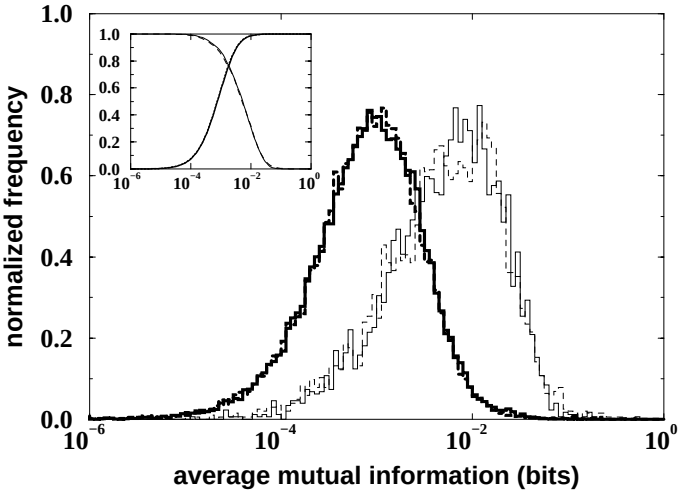
The source code is available upon request from `ivo@bu.edu`.

4 Accuracy

Fig. 2 shows the AMI histograms for coding and noncoding human DNA sequences of length 108 bp from the data sets of Fickett and Tung[1]. Since the AMI does not require prior training, we show the AMI histograms for both the training and the test set. We find that for both data sets the AMI distributions are significantly different for coding and noncoding DNA.

[a] For the sake of simplicity, assume N be a multiple of 3.

Figure 2: AMI distributions of data sets `humg108a` (solid lines) and `humg108b` (dashed lines) of Fickett and Tung[1] for coding DNA (thin lines) and noncoding DNA (thick lines). In both data sets the AMI distribution of noncoding DNA is centered at significantly smaller values than the AMI distribution of coding DNA. The cumulative distribution functions of the AMI presented in the inset show that the AMI allows a discrimination of coding and noncoding DNA with an accuracy of approximately 76%.



In order to compare the accuracy by which the AMI can distinguish coding from noncoding DNA with the accuracy of traditional coding measures, we use the standard benchmark test and data sets of Fickett and Tung[1]. Table 1 shows the accuracy of the top 8 phase-independent coding measures as ranked in Fickett and Tung[1] and the accuracy of the AMI computed on exactly the same data sets. We find that the AMI is as accurate as many of the traditional coding measures, which are trained on organism-specific data sets[1], in contrast to the AMI, which does not require prior training.

5 Species Independence

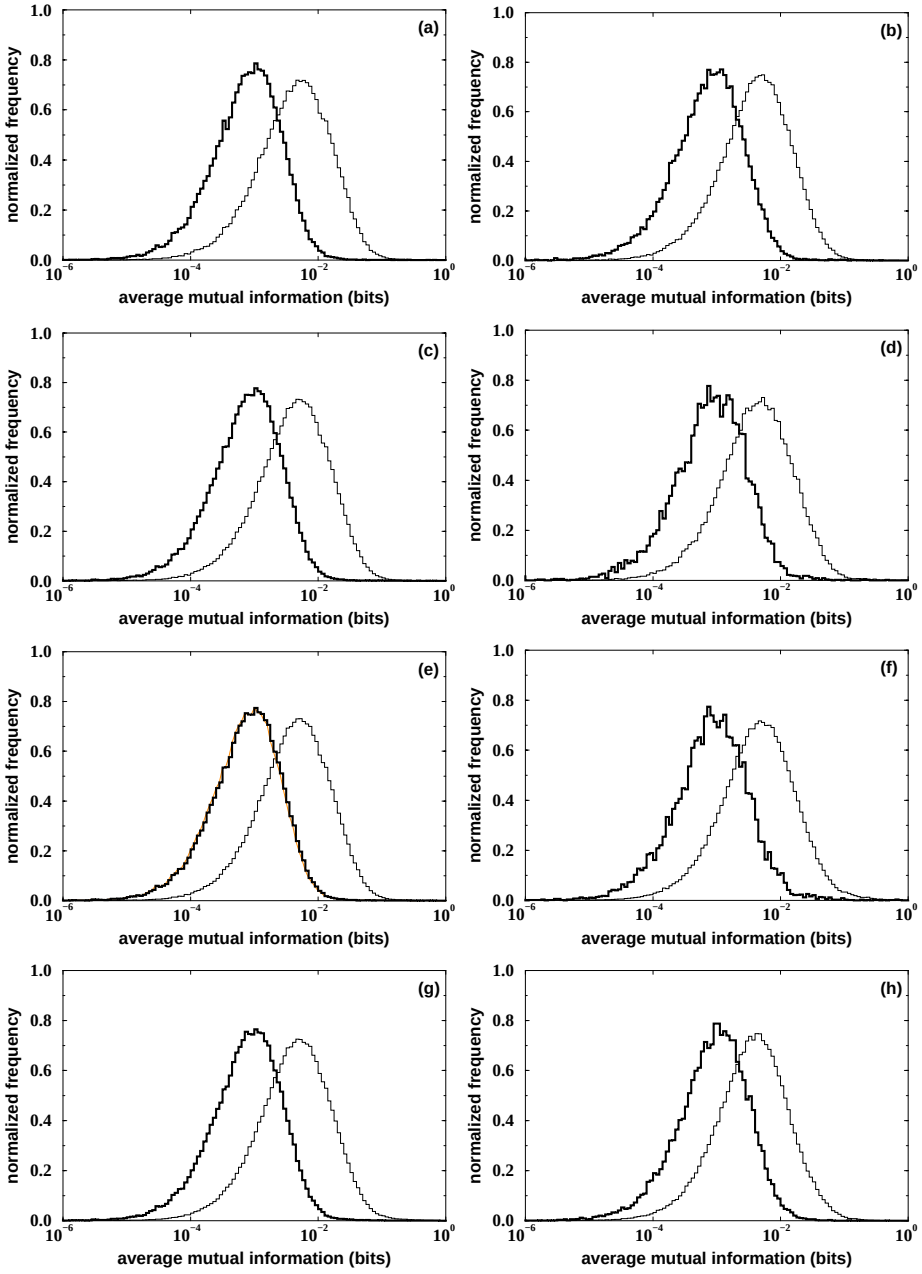
After having found that—without prior training—the AMI can distinguish coding from noncoding DNA as accurately as traditional coding measures, the question arises if the probability distribution functions of the AMI are species-independent. Fig. 3 shows the histograms of the AMI for 8 taxonomic sets, obtained from GenBank[26] release 111. We find that for all studied organisms the AMI distributions are significantly different for coding and noncoding DNA.

Table 1: Accuracy of 8 coding measures and the AMI. We compare the accuracies of the best 8 phase-independent coding measures as evaluated by Fickett an Tung[1] to the accuracy of the AMI for three sets of coding and noncoding human DNA sequences of lengths 54 bp, 108 bp, and 162 bp. We find that, on all three length scales, the accuracy of the AMI (without prior training) is comparable to the accuracy of traditional coding measures (after prior training).

	Coding Measure	54 bp	108 bp	162 bp
1.	Hexamer	70.5%	73.1%	74.2%
2.	Position Asymmetry	70.2%	76.6%	80.6%
3.	Dicodon Usage	70.2%	72.9%	73.9%
4.	Fourier	69.9%	76.5%	80.8%
5.	Hexamer-1	69.9%	72.6%	73.8%
6.	Hexamer-2	69.9%	72.6%	73.8%
7.	Run	66.6%	70.3%	71.3%
8.	Codon Usage	65.2%	68.0%	69.5%
9.	AMI	69.2%	76.1%	80.7%

We also find that for both coding and noncoding DNA the AMI distributions are virtually the same for all studied organisms. This species-independence of the AMI distributions is interesting because the AMI is a function of the codon usage, which is known to be species dependent[27,28,29,32,33,34,35].

Figure 3: AMI distributions of exons (thin line) and introns (thick line). We extract from the GenBank 111 files `gbpri1.seq`, `gbpri2.seq`, `gbpri3.seq`, `gbrod.seq`, `gbmam.seq`, `gbvrt.seq`, `gbinv1.seq`, `gbinv2.seq`, `gbpln1.seq`, and `gbpln2.seq` all protein-coding exons and all introns with a minimum length of 108 bp, and we cut these sequences into non-overlapping fragments of length 108 bp, starting at the 5'-end. We use a binary tree to categorize all eukaryotic DNA sequences into animals and plants, vertebrates and invertebrates, mammals and non-mammalian vertebrates, as well as primates and non-primate mammals. Specifically, we define primates $\equiv$ pri1 + pri2 +pri3, non-primate mammals $\equiv$ rod + mam, mammals $\equiv$ primates + non-primate mammals, non-mammalian vertebrates $\equiv$ vrt, vertebrates $\equiv$ mammals + non-mammalian vertebrates, invertebrates $\equiv$ inv1 + inv2, animals $\equiv$ vertebrates + invertebrates, and plants $\equiv$ pln1 + pln2. We compute for 8 different sets of organisms the AMI of all DNA sequences of equal length 108 bp and show the histograms of the corresponding AMI values in panels (a)–(h), which are organized as follows: the top row of the two panels compares (a) primates with (b) non-primate mammals; the second row compares (c) mammals with (d) non-mammalian vertebrates; the third row compares (e) vertebrates with (f) invertebrates; and the bottom row compares (g) animals with (f) plants. For all taxonomic classes, the AMI distribution of noncoding DNA is centered at significantly smaller values than the AMI distribution of coding DNA. The absence of significant differences between the histograms of different taxonomic categories states that the AMI distributions are species independent across the studied taxonomic classes for both coding and noncoding DNA.


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
normalized frequency
average mutual information (bits)
0.0
0.2
0.4
0.6
0.8
1.0
10^{-6}
10^{-4}
10^{-2}
10^{0}

6 Quantification of Species Independence

In order to quantitatively compare the "species independence" of the AMI to the "species independence" of the codon usage, we introduce a quantity that we call the *degree of species dependence* (DSD). Define x_i and y_i $(i = 1, \ldots, M)$ to be the usage frequencies of the $M = 64$ codons for two non-overlapping sets of 1024 DNA sequences of length 108 bp. Denote by

$$\chi^2(X, Y) \equiv \sum_{i=1}^{M} \frac{(x_i - y_i)^2}{x_i + y_i} - (M + 1) \tag{4}$$

the normalized "distance" between two histograms $X \equiv (x_1, \ldots, x_M)$ and $Y \equiv (y_1, \ldots, y_M)$. Let A_c, A_n, B_c, and B_n denote the four possible histograms for coding and noncoding DNA from the taxonomic groups A and B. We define the DSD to be the ratio of the average distance between species and the average distance between coding and noncoding DNA,

$$\mathrm{DSD} \equiv \frac{\chi^2(A_c, B_c) + \chi^2(A_n, B_n)}{\chi^2(A_c, A_n) + \chi^2(B_c, B_n)} \,. \tag{5}$$

We analyze the degree of species dependence of the codon usage on four taxonomic levels by comparing primates with non-primate mammals, mammals with non-mammalian vertebrates, vertebrates with invertebrates, and animals with plants. We randomly partition the set of all GenBank-111 sequences into non-overlapping blocks of 1024 sequences, and compare all possible combinations of these blocks. Table 2 shows the average DSD over these combinations.

Column 1 of Table 2 shows that the degree of species dependence of the codon usage is quite small (0.01) when primates are compared to non-primate mammals. This states that the codon usage is not identical in primates and non-primate mammals, but it is so similar that the codon usage differences between primates and non-primate mammals is about 100 times smaller than the differences between exons and introns. When we compare vertebrates to invertebrates, the degree of species dependence increases to about 0.69, which states that the differences between species are approximately 2/3 as large as the differences between exons and introns. The data from column 1 are consistent with the well-known fact that the codon usage is species dependent[27,28,29,32,33,34,35].

Next, we analyze the degree of species dependence of the AMI by discretizing the continuous AMI distributions as follows: when comparing two AMI distributions X and Y (see Fig. 3), we map the AMI values into $M = 64$ bins in such a way that each bin $i \in \{1, \ldots, M\}$ contains the same number

Table 2: The degree of species dependence of the codon usage and the AMI. Column 1 displays the DSD of the codon usage; the value of 0.01 in row 1 states that the codon usage differences between primates and non-primate mammals are only 1% of the differences between coding and noncoding DNA. When DNA is analyzed from species belonging to different taxonomic classes, phyla, or kingdoms (rows 2, 3, and 4), the DSD becomes larger, which quantifies the well-known fact that the codon usage is strongly species dependent. Columns 2 displays the degree of species dependence of the AMI, which we compute in the same way (and for the same sets of sequences) as for the codon usage. The degree of species dependence of the AMI never exceeds 0.02, quantifying the finding from Fig. 3 that the AMI distributions are species independent.

class	of	organism	codon usage	AMI
primates	–	non-primate mammals	0.01	0.01
mammals	–	non-mammalian vertebrates	0.10	0.01
vertebrates	–	invertebrates	0.69	0.01
animals	–	plants	0.58	0.02

of data points $x_i + y_i$. We then compute the DSD of these discretized AMI distributions X and Y for the same blocks of 1024 sequences of length 108 bp as we used to calculate the DSD of the codon usage distributions.

We find (column 2 of Table 2) that the AMI differences between primates and non-primate mammals are about 100 times smaller than the AMI differences between exons and introns. It is surprising that the degree of species dependence remains of the order of 0.01 when mammals are compared to non-mammalian vertebrates, or when vertebrates are compared to invertebrates. Even when DNA from animals is compared to DNA from plants, the AMI yields a degree of species dependence of only 0.02. The data from column 2 are in agreement with the observation, based on Fig. 2 and Fig. 3, that the AMI distributions are species independent.

This species independence, in connection with the finding that the accuracy of the AMI is comparable to the accuracy of traditional coding measures, suggests that the AMI might possibly be useful for the recognition of protein-coding regions in genomes for which training sets do not exist.

References

1. Fickett, J. W. & Tung, C.-S. (1992) *Nucleic Acids Res.* **20**, 6441–6450.
2. Gelfand, M. S. (1995) *J. Comp. Biol.* **2**, 87–115.
3. Burset, M. & Guigo, R. (1996) *Genomics* **34**, 353–367.
4. Fickett, J. W. (1996) *Comput. Chem.* **20**, 103–118.
5. Claverie, J.-M. (1997) *Hum. Mol. Gen.* **6**, 1735–1744.

6. Staden, R. & McLachlan, A. D. (1982) *Nucleic Acids Res.* **10**, 141–156.
7. Fickett, J. W. (1982) *Nucleic Acids Res.* **10**, 5303–5318.
8. Amalgor, H. (1985) *J. theor. Biol.* **117**, 127–136.
9. Guigo, R., Knudsen, S., Drake, N. & Smith, T. F. (1992) *J. Mol. Biol.* **226**, 141–157.
10. Borodovski, M. & McIninch, J. (1993) *J. Mol. Biol.* **268**, 1–17.
11. Gelfand, M. S. & Roytberg, M. A. (1993) *BioSystems* **30**, 173–182.
12. Dong, S. & Searls, D. B. (1994) *Genomics* **23**, 540–551.
13. Solovyev, V. V., Salomov, A. A. & Lawrence, C. B. (1994) *Nucleic Acids Res.* **22**, 5156–5163.
14. Thomas, A. & Skolnick, M. H. (1994) *IMA J. Math. Appl. Med. Biol.* **11**, 149–160.
15. Snyder, E. E. & Stormo, G. D. (1995) *J. Mol. Biol.* **248**, 1–18.
16. Xu, Y. & Uberbacher, E. C. (1997) *J. Comput. Biol.* **4**, 325–338.
17. Tiwari, S. et al. (1997) *Comput. Appl. Biosci.* **13**, 263–270.
18. Zhang, M. Q. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 565–568.
19. Burge, C. & Karlin, S. (1997) *J. Mol. Biol.* **268**, 78–94.
20. Salzberg, S., Delcher, A., Kasif, S. & White, O. (1998) *Nucleic Acids Res.* **15**, 544–548.
21. Kleffe, J. et al. (1998) *Bioinformatics* **14**, 232–243.
22. Shannon, C. E. (1948) *Bell Syst. Techn. J.* **27**, 379–423, 623–656.
23. Kullback, S. (1968) *Information Theory and Statistics* (Dover, New York).
24. Altschul. S. F. (1991) *J. Mol. Biol.* **219**, 555–565.
25. Dewey, T. G. (1997) *Fractals in Molecular Biophysics* (Oxford University Press, Oxford).
26. GenBank, release 111, 15 April 1999.
27. Fiers, W. & Grosjean, H. (1979) *Nature* **277**, 328–328.
28. Ikemura, T. (1981) *J. Mol. Biol.* **146**, 1–21.
29. Sharp, P. M. & Li, H. (1987) *Nucleic Acids Res.* **15**, 1281–1295.
30. Herzel, H. & Grosse, I. (1995) *Physica A* **216**, 518–542.
31. Grosse, I., Herzel, H., Buldyrev, S. V. & Stanley, H. E. (1999) *Species Independence of Mutual Information in Coding and Noncoding DNA*, submitted.
32. Bulmer, M. (1987) *Nature* **325**, 728–730.
33. Bernardi, G. (1989) *Ann. Rev. Genet.* **23**, 637–661.
34. Nakamura, Y. et al. (1996) *Nucleic Acids Res.* **24**, 214–215.
35. Karlin, S. & Mrazek, J. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 10227–10232.