Species independence of mutual information in coding and noncoding DNA

Ivo Grosse,[1] Hanspeter Herzel,[2] Sergey V. Buldyrev,[1] and H. Eugene Stanley[1]
[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2]*Institute for Theoretical Biology, Humboldt University, Invalidenstrasse, 43, 10115 Berlin, Germany*
(Received 29 October 1999)

We explore if there exist universal statistical patterns that are different in coding and noncoding DNA and can be found in all living organisms, regardless of their phylogenetic origin. We find that (i) the *mutual information function* $\mathcal{I}$ has a significantly different functional form in coding and noncoding DNA. We further find that (ii) the probability distributions of the *average mutual information* $\overline{\mathcal{I}}$ are significantly different in coding and noncoding DNA, while (iii) they are almost the same for organisms of all taxonomic classes. Surprisingly, we find that $\overline{\mathcal{I}}$ is capable of predicting coding regions as accurately as organism-specific coding measures.

PACS number(s): 87.10.+e, 02.50.−r, 05.40.−a

I. INTRODUCTION

DNA carries the genetic information of most living organisms, and the goal of genome projects is to uncover that genetic information. Hence, genomes of many different species, ranging from simple bacteria to complex vertebrates, are currently being sequenced. As automated sequencing techniques have started to produce a rapidly growing amount of raw DNA sequences, the extraction of information from these sequences becomes a scientific challenge. A large fraction of an organism's DNA is not used for encoding proteins [1]. Hence, one basic task in the analysis of DNA sequences is the identification of coding regions. Since biochemical techniques alone are not sufficient for identifying all coding regions in every genome, researchers from many fields have been attempting to find statistical patterns that are different in coding and noncoding DNA [2–6]. Such patterns have been found, but none seems to be species independent. Hence, traditional coding measures [7] based on these patterns need to be trained on organism-specific data sets before they can be applied to identify coding DNA. This training-set dependence limits the applicability of traditional coding measures, as many new genomes are currently being sequenced for which training sets do not exist.

II. MUTUAL INFORMATION FUNCTION

In search for *species-independent* statistical patterns that are different in coding and noncoding DNA, we study the *mutual information function* $\mathcal{I}(k)$, which quantifies the amount of information (in units of bits) that can be obtained from one nucleotide X about another nucleotide Y that is located k nucleotides downstream from X [8]. Within the framework of statistical mechanics $\mathcal{I}$ can be interpreted as follows. Consider a compound system (X,Y) consisting of the two subsystems X and Y. Let p_i denote the probability of finding system X in state i, let q_j denote the probability of finding system Y in state j, and let P_{ij} denote the joint probability of finding the compound system (X,Y) in state (i,j). Then the entropies of the systems X,Y, and (X,Y) are defined by

$$\mathcal{H}[X] \equiv -k_B \sum_i p_i \ln p_i ,$$

$$\mathcal{H}[Y] \equiv -k_B \sum_j q_j \ln q_j , \quad \text{and}$$

$$\mathcal{H}[X,Y] \equiv -k_B \sum_{i,j} P_{ij} \ln P_{ij} ,$$

where k_B denotes the Boltzmann constant. If X and Y are *statistically independent*, then $\mathcal{H}[X]+\mathcal{H}[Y]=\mathcal{H}[X,Y]$, which states that the Boltzmann entropy is *extensive*. If X and Y are *statistically dependent*, then the sum of the entro-

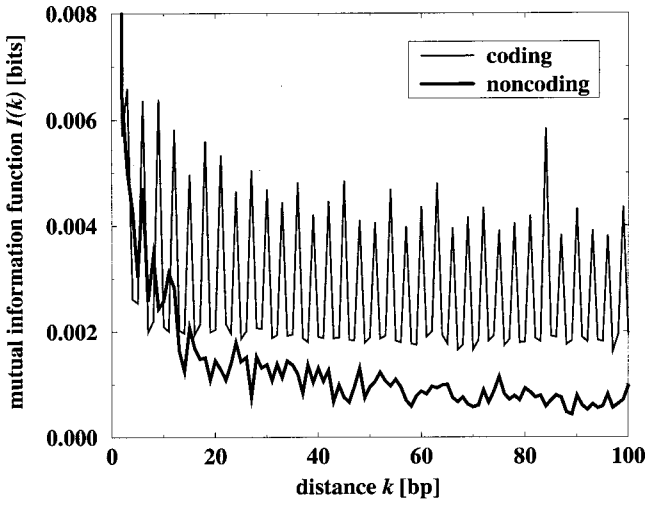

FIG. 1. Mutual information function, $\mathcal{I}(k)$, of human coding (thin line) and noncoding (thick line) DNA, from GenBank release 111 (Ref. [10]). We cut all human, non-mitochondrial DNA sequences into non-overlapping fragments of length 500 bp, starting at the 5′-end. We compute the mutual information function of each fragment, correct for the finite length effect (Ref. [13]), and display the average over all mutual information functions (of coding and noncoding DNA separately). We find that for noncoding DNA $\mathcal{I}(k)$ decays to zero as k increases, while for coding DNA $\mathcal{I}(k)$ shows persistent period-3 oscillations.

pies of the subsystems X and Y is *strictly greater* [9] than the entropy of the compound system (X,Y), i.e., $\mathcal{H}[X]+\mathcal{H}[Y] > \mathcal{H}[X,Y]$. The *mutual information* $\mathcal{I}[X,Y]$ is defined as the difference of the sum of the entropies of the subsystems and the entropy of the compound system,

$$\mathcal{I}[X,Y] \equiv \mathcal{H}[X]+\mathcal{H}[Y]-\mathcal{H}[X,Y].$$

If k_B is replaced by $1/\ln 2$, then $\mathcal{I}[X,Y]$ quantifies the amount of information in X about Y in units of bits [9]. Two obvious but noteworthy properties of $\mathcal{I}[X,Y]$ are (i) $\mathcal{I}[X,Y]=\mathcal{I}[Y,X]$, so the amount of information in X about Y is equal to the amount of information in Y about X, and (ii) $\mathcal{I}[X,Y]\geqslant 0$, so the amount of information is always nonnegative, and it is equal to zero if and only if X and Y are statistically independent. We choose $P_{ij}(k)$ to denote the joint probability of finding the pair of nucleotides n_i and n_j $(n_i,n_j \in \{A,C,G,T\})$ spaced by a gap of $k-1$ nucleotides, and we define $p_i \equiv \Sigma_j P_{ij}(k)$ and $q_j \equiv \Sigma_i P_{ij}(k)$. Then

$$\mathcal{I}(k) \equiv \sum_{i,j=1}^{4} P_{ij}(k) \log_2 \frac{P_{ij}(k)}{p_i q_j} \tag{1}$$

quantifies the degree of statistical dependence between the nucleotides X and Y spaced by a gap of $k-1$ nucleotides, and we study $\mathcal{I}$ as a function of k for coding and noncoding DNA of all eukaryotic organisms available in GenBank release 111 [10].

Figure 1 shows $\mathcal{I}(k)$ for human coding and noncoding DNA. We observe that for noncoding DNA $\mathcal{I}(k)$ decays to zero, whereas for coding DNA $\mathcal{I}(k)$ oscillates between two values, the *in-frame* mutual information $\mathcal{I}_{\text{in}}$ at distances k that are multiples of 3 and the *out-of-frame* mutual information $\mathcal{I}_{\text{out}}$ at all other values of k.

III. AVERAGE MUTUAL INFORMATION

The oscillatory behavior of $\mathcal{I}(k)$ in coding DNA is a consequence of the presence of the genetic code [which maps nonoverlapping nucleotide triplets (codons) to amino acids] and the nonuniformity of the codon frequency distribution. The fact that the codon frequencies are nonuniformly distributed in almost all organisms is well known to biologists, and arises because (i) the frequency distribution of amino acids is non-uniform, (ii) the number of synonymous codons [11] that encode one amino acid varies from 1 to 6, and (iii) the frequency distribution of synonymous codons is nonuniform [12].

A simple model that incorporates the nonuniformity of the codon frequency distribution, but neglects any other correlation, is the *pseudo-exon model* [13], which concatenates codons randomly chosen from a given probability distribution $(Q_{AAA},\ldots,Q_{TTT})$, where Q_{XYZ} denotes the probability of codon XYZ $(X,Y,Z \in \{A,C,G,T\})$. As the pseudo-exon model has been shown to reproduce the period-3 oscillations in genomic DNA [13], we use the model assumption of neglecting weak correlations between codons in order to express the joint probabilities $P_{ij}(k)$ in terms of the 12 *positional nucleotide probabilities* $p_i^{(m)}$ [14] of finding nucleotide n_i at position $m \in \{1,2,3\}$ in an arbitrarily chosen reading frame [15] as follows [3,13]:

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

It is clear that $P_{ij}(k)$ is invariant under shifts of the reading frame, because the expressions on the rhs of Eq. (2) are invariant under cyclic permutations of the upper indices (1,2,3). Since the second and third line on the rhs of Eq. (2) are identical after transposition of the lower indices (i,j), we obtain $P_{ij}(k=4,7,10,\ldots)=P_{ji}(k=5,8,11,\ldots)$, which implies that $\mathcal{I}(k)$ computed from $P_{ij}(k)$ of Eq. (2) will assume only two different values, $\mathcal{I}_{\text{in}}=\mathcal{I}(3,6,9,\ldots)$ and $\mathcal{I}_{\text{out}}=\mathcal{I}(4,5,7,8,10,11,\ldots)$.

In order to construct a coding measure that can predict whether a single sequence is coding or noncoding, we focus on the presence (absence) of the period-3 oscillation in coding (noncoding) DNA, and neglect any other statistical pattern in $\mathcal{I}(k)$, such as the decay of $\mathcal{I}(k)$ in noncoding DNA and the decay of the envelope of $\mathcal{I}(k)$ in coding DNA. Based on Eq. (2), we are able to express, for each single DNA sequence, the maxima and minima of the $\mathcal{I}(k)$ oscillations, $\mathcal{I}_{\text{in}}$ and $\mathcal{I}_{\text{out}}$, in terms of $p_i^{(m)}$ as follows: we sample from each sequence the 12 frequencies $p_i^{(m)}$, compute $P_{ij}(k)$ from $p_i^{(m)}$ by using Eq. (2), and then compute

$$\mathcal{I}_{\text{in}}=\mathcal{I}(3) \quad \text{and} \quad \mathcal{I}_{\text{out}}=\mathcal{I}(4)=\mathcal{I}(5) \tag{3}$$

by plugging $P_{ij}(k)$ and $p_i=q_i=(p_i^{(1)}+p_i^{(2)}+p_i^{(3)})/3$ into Eq. (1). For the sake of obtaining a simple coding measure with a natural and intuitive interpretation, we compute from $\mathcal{I}_{\text{in}}$ and $\mathcal{I}_{\text{out}}$ the *average mutual information*

$$\overline{\mathcal{I}} \equiv \mathcal{P}_{\text{in}} \cdot \mathcal{I}_{\text{in}} + \mathcal{P}_{\text{out}} \cdot \mathcal{I}_{\text{out}}, \tag{4}$$

where $\mathcal{P}_{\text{in}}=\frac{1}{3}$ and $\mathcal{P}_{\text{out}}=\frac{2}{3}$ denote the occurrence probabilities of $\mathcal{I}_{\text{in}}$ and $\mathcal{I}_{\text{out}}$. The value of $\overline{\mathcal{I}}$ quantifies the *average* amount [16] of information one obtains about a nucleotide X by learning both the identity of any other nucleotide Y in the same DNA sequence and whether the distance k between X and Y is a multiple of 3. We compute $\overline{\mathcal{I}}$ from each single sequence fragment [17] with the goal to distinguish coding from noncoding DNA. Due to the presence of the genetic code we expect that $\overline{\mathcal{I}}$ will be typically greater in coding than in noncoding DNA.

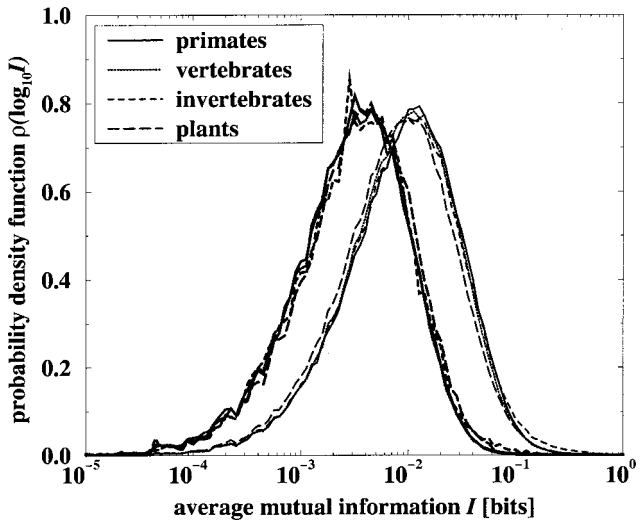

FIG. 2. $\overline{\mathcal{I}}$ distributions of coding DNA (thin lines) and noncoding DNA (thick lines) from all eukaryotic DNA sequences in GenBank release 111 (Ref. [10]). We cut all DNA sequences into nonoverlapping fragments of length 54 bp (Ref. [17]), starting at the 5′-end. We compute $\overline{\mathcal{I}}$ of each DNA fragment and show the $\overline{\mathcal{I}}$ histograms for coding and noncoding DNA, for each of the 4 disjoint taxonomic sets (primates, nonprimate vertebrates, invertebrates, plants) separately. We find that (i) for all taxonomic sets $\rho_n(\overline{\mathcal{I}})$ is centered at significantly smaller values than $\rho_c(\overline{\mathcal{I}})$, while (ii) $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$ of different taxonomic sets are almost identical. The close similarity of the $\overline{\mathcal{I}}$ distributions for different taxonomic orders, phyla, and kingdoms illustrates the species independence of $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$.

IV. ACCURACY OF THE AVERAGE MUTUAL INFORMATION

First, we investigate how accurately $\overline{\mathcal{I}}$ can distinguish coding from noncoding DNA. The *accuracy* $\mathcal{A}$ is defined as follows: Denote by $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$ the probability density functions of $\overline{\mathcal{I}}$ for coding and noncoding DNA (see Fig. 2). Define the overlap integral $\mathcal{O}(\overline{\mathcal{I}}) \equiv \int \mathcal{M}(\overline{\mathcal{I}}) d\overline{\mathcal{I}}$, where $\mathcal{M}(\overline{\mathcal{I}})$ denotes the maximum of the two values $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$ at position $\overline{\mathcal{I}}$. In statistical terms, $\mathcal{O}(\overline{\mathcal{I}})$ can be expressed as the sum of T_p and T_n, $\mathcal{O}(\overline{\mathcal{I}}) = T_p + T_n$, where $T_p(T_n)$ denotes the fraction of true positives (true negatives) over all positives (all negatives) [18]. Hence, the accuracy, defined by $\mathcal{A}(\overline{\mathcal{I}}) \equiv \mathcal{O}(\overline{\mathcal{I}})/2$, ranges from from $\frac{1}{2}$ (no discrimination) to 1 (perfect discrimination) [19].

We use the standard data set and benchmark test from Ref. [5] and compare the accuracy of $\overline{\mathcal{I}}$ to the accuracy of all of the 21 coding measures evaluated in Ref. [5]. We find that the accuracy of $\overline{\mathcal{I}}$ [$\mathcal{A}(\overline{\mathcal{I}}) = 0.69,\ 0.76,\ 0.81$ for human DNA sequences of lengths $N = 54,\ 108,\ 162$ bp] is higher than the accuracy of many of the 21 traditional coding measures from Ref. [5]. In particular, $\mathcal{A}(\overline{\mathcal{I}})$ is comparable to the accuracy of the hexamer measure H, [$\mathcal{A}(H) = 0.70,\ 0.73,\ 0.74$], which is the most accurate of the 21 frame-independent [15] coding measures from Ref. [5]. This finding is interesting, because H (like all other 20 traditional coding measures) is trained on species-specific data sets, and $\overline{\mathcal{I}}$ is not. If the $\overline{\mathcal{I}}$ distributions turn out to be species independent, then $\overline{\mathcal{I}}$ could be used without prior training to distinguish coding from noncoding DNA in all species, regardless of their taxonomic origin [20].

TABLE I. Means (variances) of $\log_{10}\overline{\mathcal{I}}$ for coding and noncoding DNA of 6 taxonomic sets. While the means of $\log_{10}\overline{\mathcal{I}}$ are significantly different in coding and noncoding DNA, they are almost the same for all taxonomic sets. Also the variances of $\log_{10}\overline{\mathcal{I}}$ are almost the same for all taxonomic sets, supplementing the visual finding from Fig. 2 that the $\overline{\mathcal{I}}$-distributions are nearly species independent.

	Noncoding	Coding
Primates	−2.52 (0.31)	−2.04 (0.30)
Nonprimate vertebrates	−2.54 (0.39)	−2.06 (0.30)
Vertebrates	−2.53 (0.34)	−2.05 (0.30)
Invertebrates	−2.50 (0.33)	−2.04 (0.32)
Animals	−2.52 (0.34)	−2.05 (0.31)
Plants	−2.48 (0.31)	−2.09 (0.31)

V. SPECIES INDEPENDENCE OF THE AVERAGE MUTUAL INFORMATION

Next, we investigate the species independence of $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$. Figure 2 shows the $\overline{\mathcal{I}}$ distributions for coding and noncoding DNA sequences from species of different taxonomic orders, phyla, and kingdoms. We find that the $\overline{\mathcal{I}}$ distributions are significantly different for coding and noncoding DNA, while they are almost identical for all taxonomic sets. In order to supplement this qualitative finding by a quantitative analysis, we present in Table I the means and variances of $\log_{10}\overline{\mathcal{I}}$ [21]. Table I shows that the means are significantly different for coding and noncoding DNA, and that the means and variances are almost the same for all species. This finding is in agreement with the visual finding based on Fig. 2 that the $\overline{\mathcal{I}}$ distributions are species independent and significantly different in coding and noncoding DNA.

VI. UNDERSTANDING THE SPECIES INDEPENDENCE FOR NONCODING DNA

In search for a possible origin of the observed species independence, we attempt to develop simple models that are able to reproduce the $\overline{\mathcal{I}}$ distributions for coding and noncoding DNA.

We first present a model that reproduces the $\overline{\mathcal{I}}$ distributions for noncoding DNA. For a random, uncorrelated sequence of arbitrary composition (p_1, p_2, p_3, p_4), we can derive the asymptotic form of the probability density function $\rho(\overline{\mathcal{I}})$ as follows: Taylor-expand $\mathcal{I}(k)$ about $P_{ij}(k) - p_i p_j$, i.e., express $\mathcal{I}(k)$ by the power series $\Sigma_{i,j}\Sigma_{\ell=0}^{\infty} a_{ij\ell}[P_{ij}(k) - p_i p_j]^{\ell}$, and truncate the Taylor series after the quadratic term ($\ell = 2$). The constant term ($\ell = 0$) vanishes because $\mathcal{I}(k) = 0$ at $P_{ij}(k) = p_i p_j$, and the linear terms ($\ell = 1$) vanish because $\mathcal{I}(k)$ achieves its minimum at $P_{ij}(k) = p_i p_j$, which causes the first derivatives of $\mathcal{I}(k)$ to vanish at $P_{ij}(k) = p_i p_j$. Hence, the first nonvanishing terms in the

Taylor-series expansion are the quadratic terms ($\ell=2$), and we obtain

$$\mathcal{I}(k) \propto \frac{1}{\ln 2} \sum_{i,j} \frac{[P_{ij}(k)-p_i p_j]^2}{2p_i p_j}, \tag{5}$$

where the symbol $\propto$ indicates that we neglect terms of $O[(P_{ij}-p_i p_j)^3]$. Substituting $P_{ij}(k)$ (for $k=3,4,5$) by the expressions on the rhs of Eq. (2) and expressing $\overline{\mathcal{I}} \equiv [\mathcal{I}(3)+\mathcal{I}(4)+\mathcal{I}(5)]/3$ in terms of $p_i^{(m)}$ yields

$$\overline{\mathcal{I}} \propto \frac{1}{\ln 2}\left[\sum_{i,m} \frac{(p_i^{(m)}-p_i)^2}{2p_i}\right]^2. \tag{6}$$

For a random, uncorrelated sequence the probability density function of $N\Sigma_{i,m}(p_i^{(m)}-p_i)^2/p_i$ converges, for asymptotically large sequence length N, to a χ^2 distribution with 6 degrees of freedom [22]. Hence, we obtain that $\rho(\overline{\mathcal{I}})$ converges, for asymptotically large N, to

$$\rho(\overline{\mathcal{I}}) = \frac{(N\sqrt{\ln 2})^3}{4} \cdot \sqrt{\overline{\mathcal{I}}} \cdot e^{-N\sqrt{\ln 2}\sqrt{\overline{\mathcal{I}}}}. \tag{7}$$

Figure 3(a) shows $\rho(\overline{\mathcal{I}})$ from Eq. (7) and the $\overline{\mathcal{I}}$ histograms for human noncoding DNA for $N=54$, 108, and 162 bp. We find that (i) the $\overline{\mathcal{I}}$ distributions for noncoding DNA collapse after rescaling with a factor of N^2, and that (ii) the $\overline{\mathcal{I}}$-distributions can be approximated by Eq. (7). The agreement of the theoretical with the experimental $\overline{\mathcal{I}}$-distributions states that the species independence of the $\overline{\mathcal{I}}$ distributions for noncoding DNA may be attributed to the absence of the genetic code in noncoding DNA of all living species.

VII. UNDERSTANDING THE SPECIES INDEPENDENCE FOR CODING DNA

We now test if the species independence of the $\overline{\mathcal{I}}$ distributions for coding DNA may be reproduced by a simple model that incorporates the presence of a reading frame. We generate a random, uncorrelated sequence where the probability of obtaining nucleotide n_i at position m is given by $p_i^{(m)}$ averaged over the entire set of DNA sequences for which the model is constructed[23]. Figure 3(b) shows the $\overline{\mathcal{I}}$ histograms for the model sequences and for human coding DNA sequences of length $N=54$ bp. We find that the $\overline{\mathcal{I}}$ distribution of the model sequences is significantly different from the $\overline{\mathcal{I}}$ distribution of human coding DNA sequences. We perform the same analyses for different organisms, ranging from simple bacteria to complex vertebrates, as well as for different N, and we find that in all cases the modeled $\overline{\mathcal{I}}$ distributions cannot reproduce the $\overline{\mathcal{I}}$ distributions of experimental, coding DNA. This result shows that the presence of a reading frame in coding DNA is not sufficient to reproduce the $\overline{\mathcal{I}}$ distributions of experimental, coding DNA, and thus cannot explain the observed species independence for coding DNA. This finding leads us to the conclusion that there must exist additional correlations or inhomogeneities [24] in coding DNA, which are responsible for the observed species-independence of the $\overline{\mathcal{I}}$ distributions.

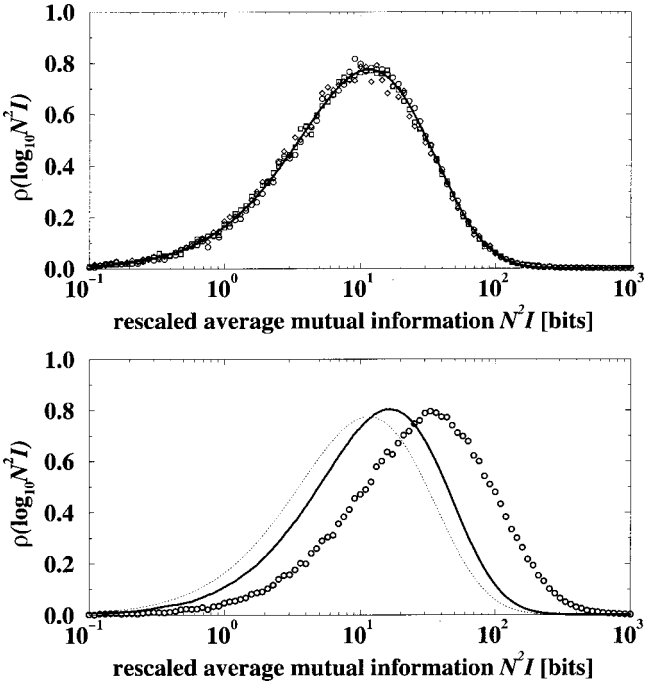

FIG. 3. Rescaled $\overline{\mathcal{I}}$ distributions of model and experimental, coding and noncoding DNA (Ref. [10]). Fig. 3(a) shows the histograms of $\log_{10} N^2\overline{\mathcal{I}}$ for human noncoding DNA for $N=54$ bp (○), 108 bp (□), and 162 bp (◇), and the corresponding χ^2 probability density function with 6 degrees of freedom (thick line). In addition to the observation (Fig. 2) that the $\overline{\mathcal{I}}$ distributions are almost identical for different species, we find that (i) the rescaled $\overline{\mathcal{I}}$ distributions collapse for all taxonomic sets and for all N, and that (ii) they agree with the χ^2 probability density function. Hence, the species independence of the $\overline{\mathcal{I}}$ distributions for noncoding DNA may be explained by the absence of a reading frame in noncoding DNA of all species. Figure 3(b) shows the histograms of $\log_{10} N^2\overline{\mathcal{I}}$ for human coding DNA sequences of length $N=54$ bp (○), the probability density function for model sequences (thick line), and the central χ^2 probability density function (thin dotted line). We find that (i) the modeled $\overline{\mathcal{I}}$ distribution (thick line) is indeed shifted to higher $\overline{\mathcal{I}}$ values than the $\overline{\mathcal{I}}$ distribution of noncoding DNA (thin dotted line), but that (ii) the $\overline{\mathcal{I}}$ distribution of the model sequences (thick line) is significantly different from the $\overline{\mathcal{I}}$ distribution of human coding DNA (○). The significant difference between the modeled and the experimental $\overline{\mathcal{I}}$ distribution states that the presence of a reading frame is not sufficient to explain the species independence of the $\overline{\mathcal{I}}$ distributions of coding DNA (Fig. 2).

VIII. CONCLUSIONS

We reported the finding of a species-independent statistical quantity, the average mutual information $\overline{\mathcal{I}}$, whose probability distribution function is significantly different in coding and noncoding DNA. We showed that $\overline{\mathcal{I}}$ can distinguish coding from noncoding DNA as accurately as traditional coding measures, which all require prior training on species-specific DNA data sets. The capability of $\overline{\mathcal{I}}$ to distinguish coding from noncoding DNA without prior training and irrespective of its phylogenetic origin suggests that $\overline{\mathcal{I}}$ might be useful to identify coding regions in genomes for which training sets do not exist. In an attempt to understand the origin of

the observed species independence of $\overline{\mathcal{I}}$, we found that the species independence of $\rho_n(\overline{\mathcal{I}})$ may result from the absence of a reading frame in noncoding DNA. We derived analytically the $\overline{\mathcal{I}}$ distribution for an ensemble of random, uncorrelated sequences of arbitrary composition, and we showed that this distribution is consistent with the observed $\overline{\mathcal{I}}$ distribution of noncoding DNA for all species and all sequence lengths N. For coding DNA, we could show that the presence of a reading frame in coding DNA sequences is not sufficient to reproduce the observed $\overline{\mathcal{I}}$ distributions of coding DNA. This finding makes it tempting to conjecture that additional correlations or inhomogeneities are a vital and species-independent ingredient of coding DNA sequences of any living organism.

ACKNOWLEDGMENTS

We thank D. Beule, C. DeLisi, J. W. Fickett, R. Guigo, K. Hermann, D. Holste, J. Kleffe, L. Levitin, W. Li, K. A. Marx, A. O. Schmitt, T. F. Smith, E. Trifonov, Z. Weng, and M. Q. Zhang for valuable discussions, and NIH, NSF, and DFG for financial support.

[1] B. Lewin, *Genes VI* (Oxford Univ. Press, Oxford, 1997); H. Lodish *et al.*, *Molecular Cell Biology* (Freeman, New York, 1995); B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland Publishing, New York, 1994).

[2] J. W. Fickett, Nucleic Acids Res. **10**, 5303 (1982).

[3] R. Staden and A. D. McLachlan, Nucleic Acids Res. **10**, 141 (1982).

[4] R. Guigo, S. Knudsen, N. Drake, and T. F. Smith, J. Mol. Biol. **226**, 141 (1992); M. Borodovski and J. McIninch, *ibid.* **268**, 1 (1993); M. S. Gelfand and M. A. Roytberg, BioSystems **30**, 173 (1993); S. Dong and D. B. Searls, Genomics **23**, 540 (1994); V. V. Solovyev, A. A. Salomov, and C. B. Lawrence, Nucleic Acids Res. **22**, 5156 (1994); A. Thomas and M. H. Skolnick, IMA J. Math. Appl. Med. Biol. **11**, 149 (1994); E. E. Snyder and G. D. Stormo, J. Mol. Biol. **248**, 1 (1995); Y. Xu and E. C. Uberbacher, J. Comput. Biol. **4**, 325 (1997); S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, Comput. Appl. Biosci **13**, 263 (1997); M. Q. Zhang, Proc. Natl. Acad. Sci. USA **94**, 565 (1997); C. Burge and S. Karlin, J. Mol. Biol. **268**, 78 (1997); J. Kleffe, Bioinformatics **14**, 232 (1998).

[5] J. W. Fickett and C.-S. Tung, Nucleic Acids Res. **20**, 6441 (1992).

[6] J. W. Fickett, Comput. Chem. (Oxford) **20**, 103 (1996); M. Burset and R. Guigo, Genomics **34**, 353 (1996); J.-M. Claverie, Hum. Mol. Genet. **6**, 1735 (1997); R. Guigo, *DNA Composition, Codon Usage, and Exon Prediction*, in Bishop (ed.) ''Genetics Databases'' (Academic Press, New York, 1999), pp 53–79.

[7] A *coding measure* is a function f that maps a statistical pattern $\vec{x}$ to a real number $y \equiv f(\vec{x})$ such that the probability distribution functions of y are different in coding and noncoding DNA. Typically, $\vec{x}$ is high dimensional, and f depends on many empirical parameters. Typically, these parameters vary significantly from species to species. Hence, these parameters must be fitted by empirical analyses of species-specific data sets. The process of fitting the parameters is called training of the coding measure.

[8] The mutual information function is similar to, but different from, autocorrelation functions (Ref. [13]). Its main advantage over correlation functions is that it does not require any mapping of symbols to numbers, which affects the analysis of symbolic sequences by correlation functions, because correlation functions are not invariant under changes of the map. Moreover, the mutual information function is capable of detecting *any* deviation from statistical independence, whereas—by definition—correlation functions measure only linear dependences. Hence, we use the mutual information function in our analysis of DNA sequences.

[9] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[10] We use all eukaryotic DNA sequences from GenBank release 111 (D. A. Benson, M. S. Boguski, D. J. Lipman, J. Ostell, B. F. Ouellette, B. A. Rapp, and D. L. Wheeler, Nucleic Acids Res. **27**, 12 (1999), ftp://ncbi.nlm.nih.gov/genbank/).

[11] There are $4^3=64$ codons, 3 of which are stop codons, and 61 of which encode 20 amino acids. Hence, the genetic code is *degenerate*, i.e., there are (many) amino acids that are encoded by more than one codon. All codons that encode the same amino acid are called *synonymous codons*.

[12] T. Ikemura, J. Mol. Biol. **146**, 1 (1981); P. M. Sharp and H. Li, Nucleic Acids Res. **15**, 1281 (1987); M. Bulmer, Nature (London) **325**, 728 (1987); G. Bernardi, Annu. Rev. Genet. **23**, 637 (1989); Y. Nakamura *et al.*, Nucleic Acids Res. **24**, 214 (1996).

[13] W. Li, J. Stat. Phys. **60**, 823 (1990); H. Herzel and I. Grosse, Physica A **216**, 518 (1995); Phys. Rev. E **55**, 800 (1997).

[14] Mathematically, $p_i^{(m)}$ can be defined in terms of Q_{XYZ} as follows: $p_i^{(1)} \equiv \Sigma_{Y,Z} Q_{n_i YZ}$, $p_i^{(2)} \equiv \Sigma_{X,Z} Q_{X n_i Z}$, and $p_i^{(3)} \equiv \Sigma_{X,Y} Q_{XY n_i}$.

[15] Since the genetic code is a nonoverlapping triplet code, there are three frames in which a DNA sequence can be translated into an amino acid sequence. In the cell, only one of the three *reading frames* encodes the proper amino acid, but in our statistical analysis the choice of the reading frame is *arbitrary* in the sense that $P_{ij}(k)$ is *invariant* under shifts of the reading frame.

[16] In terms of the mutual information function $\mathcal{I}(k)$ for the pseudo-exon model, the average mutual information $\overline{\mathcal{I}}$ can be expressed as $\overline{\mathcal{I}} = \lim_{N\to\infty} \Sigma_{k=1}^{N} \mathcal{I}(k)/N$.

[17] We choose the length to be 54 bp in order to allow a comparison with the standard data set created in Ref. [5], which consists of sequences of length 54 bp.

[18] Here, true positives (true negatives) refer to correctly-predicted coding (noncoding) sequences, and positives (negatives) refer to all coding (noncoding) sequences. Hence, T_p (T_n) denotes the fraction of correctly predicted coding (noncoding) sequences over all coding (noncoding) sequences. Mathematically, T_p and T_n are defined by $T_p \equiv \int \theta[\rho_c(\overline{\mathcal{I}}) - \rho_n(\overline{\mathcal{I}})] \rho_c(\widetilde{\mathcal{I}}) d\overline{\mathcal{I}}$ and $T_n \equiv \int \theta[\rho_n(\overline{\mathcal{I}}) - \rho_c(\overline{\mathcal{I}})] \rho_n(\widetilde{\mathcal{I}}) d\overline{\mathcal{I}}$, where θ denotes the Heavyside function, i.e., $\theta(x) \equiv 1$ for $x \geq 0$ and $\theta(x) \equiv 0$ for $x < 0$.

[19] If $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$ were identical, $\mathcal{O}(\overline{\mathcal{I}})$ would be equal to 1. If $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$ were completely disjoint (nonoverlapping), $\mathcal{O}(\overline{\mathcal{I}})$ would be equal to 2.

[20] It is clear that $\overline{\mathcal{I}}$ can be computed from sequences of any length N (which does not need to be a multiple of 54 bp). We present the accuracy of $\overline{\mathcal{I}}$ for $N=54$ bp, $N=108$ bp, and $N=162$ bp because these are the three length scales on which all of the 21 coding measures in Ref. [5] are evaluated.

[21] In Figs. 2 and 3 and in Table I we take the logarithm of $\overline{\mathcal{I}}$ because (i) the $\overline{\mathcal{I}}$ distributions have a broad tail (ranging over several orders of magnitude), and (ii) they are sharply peaked at $\overline{\mathcal{I}}=0$. Consequently, the moments of $\overline{\mathcal{I}}$ are dominated by large values of $\overline{\mathcal{I}}$ and not by the bulk of the distribution. Hence, we display the density and compute the moments of $\log_{10}\overline{\mathcal{I}}$ rather than those of $\overline{\mathcal{I}}$.

[22] The mathematical proof can be found in: H. Cramer, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, 1946). An intuitive heuristic argument of why the number of degrees of freedom is equal to 6 is that there are $4+3-1$ independent linear constraints that the $4\times3=12$ numbers $p_i^{(m)}-p_i$ must satisfy. Hence, the number of degrees of freedom is $4\times3-(4+3-1)=6$.

[23] For the probabilities $p_i^{(m)}$ we choose the total number of nucleotides n_i in position m of the biological reading frame divided by the total number of nucleotides from exactly the same set of coding human sequences to which the model sequences are compared.

[24] By *correlations or inhomogeneities* we mean that the probability distributions $p_i^{(m)}$ are not constant, but vary along the DNA sequence from gene to gene and also within a gene. These variations of the probability distributions $p_i^{(m)}$ seem to be a typical feature of coding DNA of any living organism.