Pergamon

Pattern Recognition, Vol. 29, No. 7, pp. 1187–1194, 1996
Elsevier Science Ltd

0031-3203/96 $15.00 + .00

0031-3203(95)00145-X

APPLICATION OF INFORMATION THEORY TO DNA SEQUENCE ANALYSIS: A REVIEW

RAMÓN ROMÁN-ROLDÁN,* PEDRO BERNAOLA-GALVÁN† and JOSÉ L. OLIVER‡

* Departamento de Física Aplicada, University of Granada, 18071-Granada, Spain
† Department of Applied Physics II, University of Málaga, Spain
‡ Institute of Biotechnology, University of Granada, Spain

(*Received* 19 *January* 1995; *in revised form* 15 *September* 1995; *received for publication* 16 *October* 1995)

Abstract—The analysis of DNA sequences through information theory methods is reviewed from the beginning in the 70s. The subject is addressed within a broad context, describing in some detail the cornerstone contributions in the field. The emerging interest concerning long-range correlations and the mosaic structure of DNA sequences is considered from our own point of view. A recent procedure developed by the authors is also outlined. Copyright © 1996 Pattern Recognition Society. Published by Elsevier Science Ltd.

Information theory DNA sequences Entropy Chaos-game representation

1. INTRODUCTION

In the words of Werner Ebeling and Mijail Volkenstein:[1] "...living beings are natural ordered and information-processing macroscopic systems originating from processes of self-organization and natural evolution... all processes in living systems originate from physical processes. Living beings are open thermodynamic systems which permanently exchange matter, energy, entropy and information with their surrounding...". Many other authors agree that living beings are characterized mainly by their ability to process information and thus they can be analyzed from this perspective. The physical support of such information is the DNA double helix, which plays a basic role in both the coding and the transmission to the next generation of all the information needed for living functions.

The above quotation suggests that the maintenance of living activity as an information processor: (a) is of physical nature and (b) affects more long-range phenomena, such as biological evolution or the origin of life. These problems are usually addressed from the double perspective of thermodynamics and information theory. Many authors have attempted to join these two approaches,[2–4] in trying to solve this problem.[5,6] Moreover, the processing of biological information has an artificial parallel: the processing of information by computers. Both types of massive information systems are needed from a joint analysis. This approach, centered on the so-called "Physics of Information", is complex and attractive, and is nowadays the subject of intense research [see reference (7) and refs therein], among which the cornerstone work may be the recent *Physics of Computation Workshop*.[8]

Here we focus on a more limited field. Nucleotide sequences are examined from an external point of view, as messages, without taking into account the detailed physical-chemical mechanisms for information processing. Protein synthesis is modeled as a system to process information, source plus channel (Section 2). A basic question is to obtain significant and reliable measures of parameters such as order, regularity, structure, complexity, etc. in a given DNA sequence. This would allow comparisons with other sequences (or with other segments of the same sequence), thus deriving results of interest to evolutionary studies (molecular phylogeny), identification of coding segments (finding genes, exons, transcription signals), etc. In general, the aim is a measure capable of indicating how far a natural sequence is from a random one.

The application of information theory to DNA sequences began in the 70s. Two periods can be distinguished, the first around 1970–1977, when the first publication appeared. Several authors[9–11] developed methods to estimate parameters such as information, redundancy or divergence in DNA sequences. The shared aim of all these studies was to obtain a quantitative expression of the complexity of these sequences. In Section 3, we describe the pioneering work of Gatlin, as well as subsequent modifications.

Despite the fact that DNA sequences contain all the relevant information for living beings, the above attempts do not completely succeed in obtaining a quantitative measure for such information. In some of these studies, DNA was virtually indistinguishable from a random sequence. The best exponent of this pessimistic point of view was the paper of Hariri *et al.*[12] After a relatively quiet pause, the second period (1987 to the present) can be characterized by renewed inte-

rest in the subject, aided by the great increase in sequence data generated by genome projects. Nucleotide-sequence data banks now contain chains long enough to overcome the major limiting factor in properly applying information theory to DNA sequences. Other techniques derived from signal theory, such as Fourier transform, autocorrelation, spectral analysis, random walks or chaotic dynamics, have also been applied. The most outstanding result was the finding of long-range correlation in DNA sequences.[13–15] Information-theory measures are also used in detecting long-range correlations, such as mutual information.[13,16]

Cosmi *et al.* reported an entropic method for recognizing DNA patterns, in order to classify sequences,[17] which is briefly described in Section 4. Recently, the role of repeats on word entropies has been analysed.[18] We addressed this topic in Section 5. In Section 6, we present our recent (1993) and current research, also showing the nonrandomness of DNA sequences through entropic profiles derived from associate chaotic images.[19]

2. THE DNA–PROTEIN COMMUNICATION CHANNEL

All the information needed to control the biological reactions in cells and tissues, including protein synthesis, is contained in DNA. We are aware that this sentence expresses only a first-order approximation, since nucleic acids exhibit a complex chemistry and the string model of DNA (representation of DNA as a string of letters that represent the order in which the nucleotides occur in the molecule) does not express that chemistry, being a very approximate description of the molecule. The central question of the application of techniques such as information theory to the study of DNA may be then how good is this first approximation. The DNA molecule can be described schematically as a polynucleotide chain forming a double helix. There are four nucleotides (bases) and thus we can consider the DNA chain to be a message coming from a source that uses an alphabet of four symbols. On the other hand, proteins are also lineal chains of 20 different basic constituents (the amino acids). Each DNA sequence determines a unique protein chain although each protein can be coded by several somewhat different DNA sequences. The genetic code establishes the correspondence between the sequence of nucleotides in DNA and the sequence of amino acids in the protein. Since there are 20 amino acids and only four nucleotides, a combination of several nucleotides—just three (codon)—is needed to code each amino acid.

Genetic information always flows in an irreversible manner from nucleic acids to protein in all living beings (no mechanisms for back-translating proteins into nucleic acids are known) and, essentially, through the same basic mechanisms. It is not our aim here to describe the details of such mechanisms, except to note that biological information transfer can be well modeled as a communication channel. The input is the DNA sequence and the output the amino-acid chain in the protein. The information source and the transmission channel for the proposed communication system are described below.

2.1. *The information source*

The central concept in studies involving contiguous patterns of textual elements is an abstract device (called the source) that generates sequences of symbols (messages) chosen from a finite alphabet. Such symbol selection can take place according to a variety of random mechanisms. Particularly important are ergodic sources in which the random mechanism leads to "typical" (i.e. statistically homogeneous) messages with high probability (close to 1) and to "atypical" sequences with negligible probability. It is well known that probability distributions of textual elements (such as letters or biliterals) are preserved in all sufficiently long texts and thus we can assume that languages, and also DNA, can be modeled by ergodic sources. For DNA, the information source may be defined by:

(1) The alphabet: $\mathbf{B} = \{C, A, U, G\}$ (C, cytosine; A, adenine; G, guanine; U, thymine or uracil).

(2) In general, symbols in **B** are not emitted with the same frequency. The probability distribution on the alphabet is a property of the source:

$$\mathrm{p}(U) + \mathrm{p}(C) + \mathrm{p}(A) + \mathrm{p}(G) = 1. \tag{1}$$

(3) Bases are not independent in the genetic message. The source cannot be considered to be of the Bernouilli type, but rather must be modeled as a Markov source with a stochastic matrix:

$$[\mathrm{p}(B_i|B_j)], \quad \sum_i \mathrm{p}(B_i|B_j) = 1. \tag{2}$$

It is also assumed that this Markov source is stationary and ergodic, and thus probability distribution can be (1) derived from the conditional probabilities and (2) experimentally determined:

$$\mathrm{p}(B_i) = \sum_j \mathrm{p}(B_i|B_j)\mathrm{p}(B_j). \tag{3}$$

2.2. *The transmission channel DNA protein*

This corresponds to the genetic code shown in Table 1, heavily degenerated, and known since 1961. The channel is assumed to be stationary and memoryless; it can be represented by the random correspondence between the codon set $\mathbf{B}^3 = \{B_1, B_2, B_3\}$ and the amino-acid set $\mathbf{A} = \{A_i\}$:

$$\mathbf{B}^3 \xrightarrow{\mathrm{p}(A/B_1B_2B_3)} A.$$

For a channel without noise (mutations), the input/output probabilities are:

$$\mathrm{p}(A_i/B_1, B_2, B_3) = \begin{cases} 1, & \text{if the pair } (A_i/B_1, B_2, B_3) \text{ belongs to the code} \\ 0, & \text{if not.} \end{cases}$$

Table 1. Genetic code

GCU,GCC,GCA,GCG	Alanine
CGU,CGC,CGA,CGG,AGA,AGG	Arginine
AAU,AAC	Asparagine
GAU,GAC	Aspartic acid
UGU,UGC	Cysteine
CAA,CAG	Glutamic acid
GAA,GAG	Glutamine
GGU,GGG,GGA,GGC	Glycine
CAU,CAC	Histidine
AUU,AUC,AUA	Isoleucine
UUA,UUG,CUU,CUC,CUA,CUG	Leucine
AAA,AAG	Lysine
AUG	Methionine
UUU,UUC	Phenylalanine
CCU,CCC,CCA	Proline
UCU,UCC,UCA,UCG	Serine
ACU,ACC,ACA,ACG	Threonine
UGG	Tryptophan
UAU,UAC	Tyrosine
GUA,GUC,GUG,CUU	Valine
UAA,UAG,UGA	STOP

3. ORDER AND COMPLEXITY MEASURES

We next consider the original contribution of Gatlin,[9] together with the subsequent improvements by Sibbald.[20] For clarity, we have somewhat modified the presentation. The analysis of sequences takes into account the base composition as well as the ordering of bases, giving rise to dimers, trimers,..., n-tuples, n being as high as possible. A DNA sequence is considered under the following two complementary viewpoints.

3.1. *The order or regularity of the sequence*

As an extreme theoretical example, the most ordered sequence might be one as, for example, *AAAAAA*..., while one of the most disordered (complex) ones could be generated in a purely random way (an empirical assessment of this quality is far from trivial[21]). As a measure of order we can adopt the divergence between a sequence which is as random as possible and the natural sequence; such divergence is taken as the difference between the two corresponding Shannon entropies:

$$D_n = H_n^{(r)} - H_n^{(n)} = -\sum_{B_n} \frac{f_i f_j}{f_{i\otimes j}} \log \frac{f_i f_j}{f_{i\otimes j}} - \left(-\sum_{B_n} f_k \log f_k \right) \quad (4)$$

where f_s are relative frequencies and the indexes span the following sets:

- k, the B_n subsequences of length n;
- i, the subsequences of length $n-1$, obtained by deleting the first base of B_n;
- j, the subsequences of length $n-1$, obtained by deleting the last base of B_n;
- $i\otimes j$, the subsequences of length $n-2$, obtained by deleting the first and the last base of B_n.

For $n=1$ and a sequence of infinite length, $H_{expected} = 2\log 2 = 2$ bits. for any other level, $H_{expected}$ is the entropy of the theoretical distribution of the frequencies of n-tuples, conditioned by the observed ones for $n-1$, assuming an independent distribution. Therefore, $H_{expected}$ represents the maximum theoretical entropy at level n compatible with the observed sequence at the level $n-1$.

3.2. *The complexity*

Conceptually, this is the complement of the previous measure. Numerically, it would also be the complement of the previous measure if the analysis were carried out at only one level, but in such a case it would be superfluous. However, the multi-level study makes both measures useful. The complexity is defined by the divergences in a recurrent manner:

$$C_n = C_{n-1} - D_n. \quad (5)$$

Note that complexity increases with the entropies of the corresponding histograms:

$$C_n = C_{n-1} - D_n = (C_{n-1} - D_{n-1}) - D_n = \cdots = C_0 - \sum_{i=1}^{n} D_i = C_0 + \sum_{i=1}^{n} H_i^{(n)} - \sum_{i=1}^{n} H_i^{(r)}. \quad (6)$$

The first two divergences have a simple meaning in the context of the genetic message. While D_1 measures the distance from equiprobability, D_2 reflects the distance from the independence for contiguous bases. The *redundancy* R of the code (up to this level) is given[11] by:

$$2\cdot R = D_1 + D_2.$$

Figure 1 shows in cascade the multi-level measures of both divergence and complexity. Additivity and the

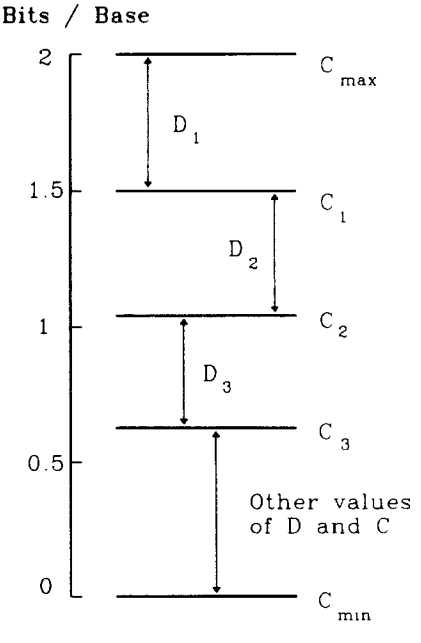

Fig. 1. Complexity scale for DNA.

alternate order simplify an understanding of its meaning. Before a DNA chain is observed, a complexity of 2 bits/base is available. The first measure (base composition) detects a divergence D_1 with respect to C_{max}, which leads to a complexity C_1 (in this case, the same as the entropy). At the second level, a divergence D_2 with respect to C_1 is detected, which lead to a complexity C_2, and so on. At any level n, the complexity C_n is residual in character, in the sense of being what remains after successive reductions due to the order introduced into the sequence from $n = 1$. However, at the same time, it also has the character of being available for additional reductions at higher levels.

Equiprobability and independence of the symbols (which imply maximum entropy per symbol) are the conditions under which the source can emit the maximum diversity of messages and thereby lead to the highest richness in information. Any deviation from these conditions decreases diversity, but increase the reliability of the message, measured by the redundancy. As a result, the system is able to detect and proof-read errors. DNA chains, as with messages in any other natural language or in any other artificial communication system, have developed a trade-off solution between what might be called *quantity* and *quality* of information. Two observations are noteworthy here: (1) in vertebrate genomes the redundancy is due mainly to D_2, while in the remaining genomes it is due to D_1;[9] (2) in agreement with the last observation, and by using protein sequences from different organisms, Reichert *et al.*[10] found the existence of linearity between R and D_2, claiming that this finding constitutes the "true arrow of time".

As mentioned above, the overall results of these attempts were not as encouraging as might be expected,[12] thus indicating that the first-order model may be insufficient and that there may be important information that the string model does not convey. It is hard to accept, however, that we are not able to express

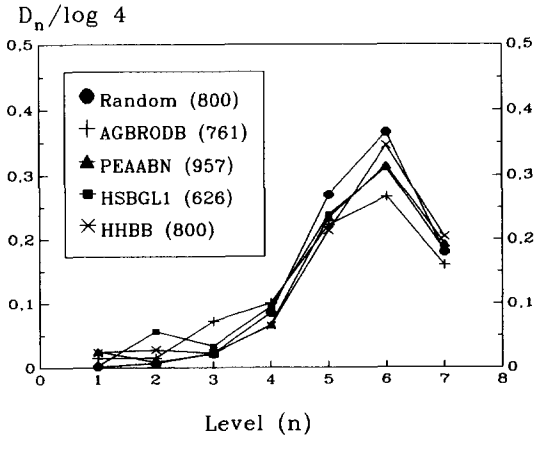

Fig. 2. Conditional divergence versus conditioning level for some short DNA sequences.

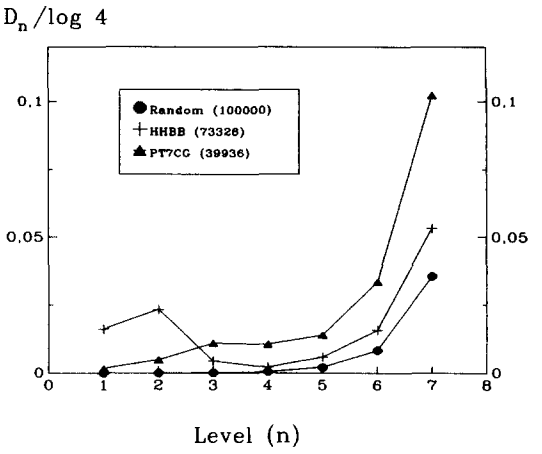

Fig. 3. Conditional divergence versus conditional level for two long DNA sequences and a random computer-generated one.

in a mathematical form all we know about genetic information. The failure may be due to the way in which we have approached the problem.

One of the reasons for the failure may be the relatively short lengths of the sequences used. The conditional divergences (divided by log 4, in order to be expressed in bits/base) of some DNA sequences with similar lengths to those used by Hariri *et al.* [see Fig. 3 in reference (12)] are shown in Fig. 2. We have also included a random sequence as well as a fragment of 800 bp from HHBB (total length = 73326 bp).

We cannot distinguish between both the different DNA sequences, nor between a given DNA sequence and a random one.

When larger sequences are used, some differences can be found. Figure 3 shows the profiles obtained for the complete length of the human sequence HHBB (73326 bp), a viral sequence (PT7GG, 39936 bp) and a random sequence of 100.000 bp in length. Nevertheless, the figures obtained for higher level values (4–7) were not reliable and it can be shown that they can vary with the length of the sequence.

4. CHARACTERIZATION OF SEQUENCES

Cosmi *et al.*[17] attempted to classify sequences according to their codon-usage frequencies. These authors used a statistical method based on maximum entropy techniques of multivariate statistics. The method was applied to the analysis of nucleotide sequences from eukaryotes. Protein-coding sequences are often interrupted by intervening untranslated sequences (introns). Dividing coding segments from their introns, these authors obtained two mutually exclusive classes. The percent of correctly attributed sequences is 26% for coding and 35% for introns with a small percent of error (1%). Thus, the method could be used to distinguish between coding and noncoding sequences.

5. THE ROLE OF REPEATS

The role of DNA repeats on word entropies has been recently analysed by Herzel *et al.*[18] They constructed hypothetical model sequences composed by equidistributed symbols with randomly interspersed repeats, showing that the entropy of DNA sequences measuring the information content is much lower than suggested by earlier empirical studies.

6. A RECENT CONTRIBUTION

Oliver *et al.*[19] have recently described a method to assign entropic profiles to DNA sequences; in such profiles an entropic measure is plotted against different subsequence lengths.

A combination of two different techniques was used. In 1990, Jeffrey[22] proposed a powerful method to analyse DNA sequences: the chaos-game representation (CGR). This method, based on a technique from chaotic dynamics, produces a square, fractal-like picture of gene sequences, visually revealing previously unknown structures. This provides a graphic way of displaying both statistical and sequential properties of DNA sequences. The densities of points in subsquares of 4^{-m} in size correspond to the frequencies of oligomers of m in length. On the other hand, certain concepts from the field of multiresolution-information for digital images recently developed by Román *et al.*[23] were applied to the CGR images of DNA chains. First, the entropy of the histogram of gray levels was translated into the entropy of the histogram of point densities in CGRs; secondly, different resolutions in images were translated into different subsquare sizes of the CGRs, which in turn correspond to different oligomer lengths.

6.1. *Iterated function system*

The aspect of chaos theory applied in this work is described by Barnsley[24] as an *iterated function system* (IFS):

$$\{(\mathbf{X}, d);\quad w_n, n = 1, 2, \ldots, N\},$$

where $(\mathbf{X}, d)$ is a metric space and w_n is the n-function of the form $w_n: \mathbf{X} \to \mathbf{X}$, such that $\forall x, y \in \mathbf{X}$, $d(w_n(x), w_n(y)) \leq s d(x, y)$, s being the contraction factor, $0 \leq s \leq 1$.

For simplicity, the IFSs are usually of the form:

$$\{\mathbb{R}^M, w_n,\quad n = 1, 2, \ldots, N\}.$$

The analysis of DNA sequences is carried out as follows: (A) The number of functions is $N = 4$, the number of different bases in the sequence; for $M = 2$, we have $X = \mathbb{R}^2$. (B) Each w_n function has the form $x_{i+1} = s_n(x_i + c_n)$ and all the contraction factors s_n are fixed to 0.5. (C) The constants c_n of the functions are assigned to the corners of a square. (D) In a random algorithm (*chaos game*), the functions w_n operate according to a given probability distribution. In this work, it is the DNA sequence that drives the operation order of the system functions; that is, each function is associated to a base in the sequence. Plotting the points determined by the successive pairs (x, y), we obtain the CGR attractor.

6.2. *Entropic measures on the histogram of densities*

By normalizing the density set, we obtain a distribution whose entropy, for each m, coincides with the one mentioned above. Instead, the authors obtain the normalized histogram $\mathbf{Q}$ of densities (the number of cells in the CGR versus the density) and, also for each m, either the entropy (H-m profile) or the relative entropy with respect to the reference histogram Z (D-m profile), being:

$$\mathrm{H}(\mathbf{Q}) = \sum_{j=0}^{N} q_j \cdot \log q_j \tag{7}$$

$$\mathrm{D}(\mathbf{Q} \| \mathbf{Z}) = \sum_{j=0}^{N} q_j \cdot \log \frac{q_j}{z_j}, \tag{8}$$

where q_i is the observed relative frequency of cells with i points and z_i the expected one.

The *reference* is a theoretical source producing sequences in a purely random way; that is, each sequence should be a sample from a succession of equiprobable random variables. If so, the resulting histogram obtained by averaging on several of these sequences goes toward the binomial distribution as the number of sequences goes to ∞.

Any departure from random behavior results in a rise in the relative entropy (always positive), whereas the effect on histogram entropy is a deviation from that corresponding to the binomial distribution. Figures higher than the binomial value may be associated with the presence of structure or complexity in the sequence, while lower figures correspond to excessively uniform sequences.

To obtain the H-m profiles, the density scale was adjusted by grouping histogram classes in order to obtain an entropy of 0.25 bits for the theoretical reference. Higher entropies reveal the presence of order and regularity in the analysed sequence. This method would hide random deviations because of the excessive sequence uniformity, but such uniformity would not be expected in DNA sequences.

6.3. *Examples and results*

H-m profiles were obtained in both random and DNA sequences for the necessary range of resolutions. Figure 4 shows the entropic profile for a DNA sequence (HUMHBB) compared with several simulated random ones of the same length (73326 nucleotides). DNA and random sequences look very different, since the first departs markedly from randomness for all resolution levels from 1 to 5.

Histogram entropies of most DNA sequences showed a maximum for $m = 2$, the exceptions being the β-globin regions from the human (HUMHBB) and mouse (MMBGCXD) genomes, the bacterial sequence

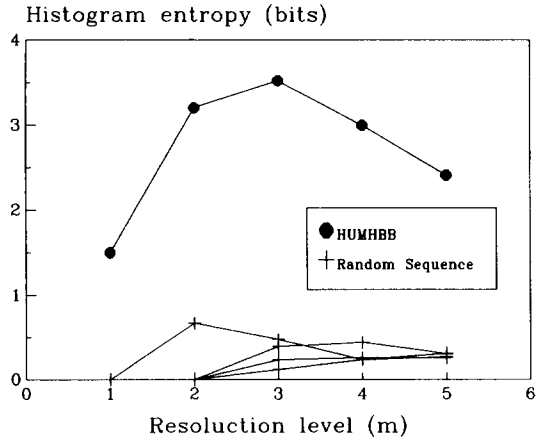

Fig. 4. The entropic profiles of a DNA sequence (HUMHBB) compared with those from several random sequences.

ECUNC and the *Xenopus* mitochondrial genome, all showing a maximum for $m = 3$.

Therefore, our results seem to suggest the existence of a maximum of dependence at the di- and trinucleotide levels, the dependence decreasing at higher resolution levels; thus, randomness of DNA sequences seems to be greater as the resolution level (i.e. the oligomer length considered) increases beyond $m = 3$.

We have also used D-m profiles in a later, more general work,[25] including, besides DNA sequences, other types of sequences.

6.4. *Current research*

We summarize here recent addenda[26] to the above described research topic.

- For low resolutions, the histogram scale is quite fine-grained: the range is $\{0, 1, \ldots, N\}$ for all resolutions, while the number of nonzero values may be at best 4^m (total number of cells in the GCR). The reference histogram, derived from a theoretical distribution, is not affectd by this restriction, taking nonzero values over the entire range. This effect is more apparent at low resolutions, since, when m rises, 4^m may reach similar or even higher values than the number of significant values in the reference histogram (for the binomial histogram, this number may be estimated as 2μ, $\mu = N/4^m$ being the average number of points per cell).

A solution proposed for this problem[19] is an adjustment of the histogram scale, but this approach presents some drawbacks: (1) the entropy limit for which a given sequence is considered random is chosen arbitrarily and (2) random deviation due to excessive uniformity are hidden.

Instead, we propose the elaboration of a *splitting CGR* as follows: Given a maximum resolution M_{max}, for $m < M_{max}$ we delineate $4^{M_{max}-m}$ CGRs by taking nonoverlapping sequence segments of length $N4^{m-M_{max}}$ CGR. The histogram for the sequence was then the average of the histograms for each segment. In this way, the maximum number of nonzero values is the same for all resolutions ($4^{M_{max}}$) and, therefore, the comparisons to the reference histogram are equally reliable (the average number of points per cell is also te same).

Since different segments in the sequence are analysed separately, splitting CGRs enables a distinction between sequences with the same global nucleotide composition, but different subsequence order (frequencies).

- The histogram entropy depend's on sequence length and thus sequences of different lengths cannot be properly compared. The solution here is a pseudo-normalization. Figure 5 is a plot of both maximum (Gibbs distribution,[11,26]) and binomial entropies versus sequence length. Furthermore, the entropies corresponding to both a random sequence with different probabilities for each symbol and a DNA chain are also shown. The relative positions of the two last values are maintained constant in respect to the maximum and the binomial entropies. Thus, we propose the following measure of sequence complexity:

$$h = \frac{H(\mathbf{Q}) - H_{bin}}{H_{max} - H_{bin}},$$

which will be named *normalized entropy*, a measure of entropy *independent of sequence length*. It is not a strict normalization, since the negative values ($H(\mathbf{Q}) < H_{bin}$) prove to be somewhat dependent on sequence length. On the other hand, the possibility of a sequence showing ($H(\mathbf{Q}) < H_{bin}$) may be also interesting, since it means a *random excess* which is highlighted through this representation.

- There are two causes by which the histogram entropy computed from the splitting CGR may differ from the binomial entropy: (a) the relative frequencies of the different subsequences are not the same (*compositional heterogeneity*) and (b) the subsequence distribution over the sequence varies (*spatial hetero-*

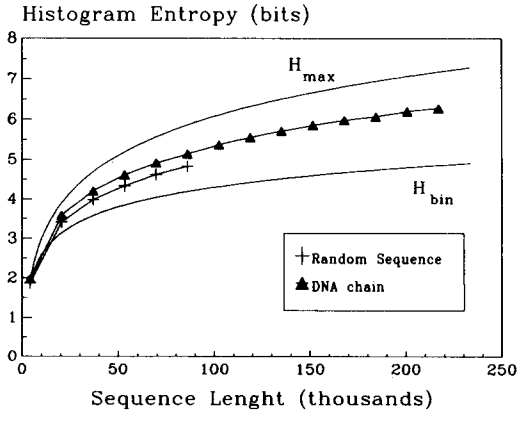

Fig. 5. Histogram entropy of the splitting CGR ($m = 2$) versus sequence length.

geneity). The solution we propose for the first problem is to use as the reference histogram, for each resolution, that obtained under the assumption that the probability for each subsequence is the corresponding observed relative frequency. It can be then shown that this histogram corresponds to the nonweighted average of binomial distributions.

Now, nonzero relative entropies, as well as the differences between histogram entropy and the reference histogram, may be attributed only to spatial heterogeneity. Note also that $H(\mathbf{Q}) < H_{ref}$ prove to mean an excess of homogeneity and that $H(\mathbf{Q}) > H_{ref}$ means an excess of heterogeneity. Such an analysis may be useful in searching for long-range correlations in DNA sequences.[13–15,27,28]

- On the other hand, it may also be useful to know the part of complexity which can be attributed to the restrictions introduced at each resolution. To discount the structure dragged from previous resolutions, the most random histogram at length m, which is compatible with the known constraints at *length* $m' < m$, is used as a reference.

7. CONCLUSIONS

As can be seen in this review, the information theory methods are fully suitable for DNA sequence analysis, in the present state of art. Mainly, it is due to their ability for handling symbolic sequences.

We thus encourage both information theory and DNA sequence researchers to develop new applications, mainly in the current subject of long-range correlations and compositional patchiness.

Acknowledgements—This work was partially supported by grants TIC91-646 to R.R.R. and PB93-1152-CO2-01 to J.L.O. from the DGICYT of the Spanish Government.

REFERENCES

1. W. Ebeling and M. V. Volkenstein, Entropy and the evolution of biological information, *Phys. A* **163**, 398–402 (1990).
2. D. R. Brooks and E. O. Willey, *Evolution as Entropy: Toward a Unified Theory of Biology*, 2nd edn. University of Chicago Press, Chicago (1988).
3. J. S. Wicken, *Evolution, Thermodynamics, and Information.* Oxford University Press, New York (1987).
4. B. H. Weber, D. J. Depew and J. D. Smith (ed.), *Entropy, Information and Evolu- tion.* M.I.T. Press, Massachusetts (1988).
5. L. Brillouin, *Science and Information Theory.* Academic Press, New York (1962).
6. C. M. Caves, Information and entropy, *Proc. Phys. Comput. Workshop.*
7. R. Landauer, Information is physical, *Phys. Today* 23–29 (May 1991).
8. Physics of Computation Workshop. Douglas Matzke Dallas (Texas, U.S.A.) proceedings not published (October 1992).
9. L. Glatin, *Information Theory and the Living System.* Columbia University Press, New York (1972).
10. T. A. Reichert, D. N. Cohen and A. K. C. Wong, An application of information theory to genetic mutations and the matching of polypeptide sequences, *J. Theoret. Biol.* **42**, 245–261 (1973).
11. S. Guiasu, *Information Theory With Applications.* McGraw-Hill, New York (1977).
12. A. Hariri, B. Weber and J. Olmsted III, On the validity of Shannon-information calculations for molecular biological sequences, *J. Theoret. Biol.* **147**, 235–254 (1988).
13. W. Li and K. Kaneko, Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence, *Europhys. Lett.* **17**(7), 655–660 (1992).
14. C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simon and H. E. Stanley, Long-range correlation in nucleotide sequences, *Nature* **356**, 168–170 (1992).
15. R. F. Voss, Evolution of long-range fractal correlations and 1/f noise in DNA base sequences, *Phys. Rev. Lett.* **68**(25), 3805–3808 (1992).
16. S. F. Altschul, Amino acid substitution matrices from an information theoretic perspective, *J. Mol. Biol.* **219**, 555–565 (1991).
17. C. Cosmi, V. Cuomo, M. Ragosta and M. F. Macchiato, Characterization of nucleotide sequences using maximum entropy techniques, *J. Theoret. Biol.* **147**, 423–432 (1990).
18. H. Herzel, W. Ebeling and A. O. Schmitt, Entropies of biosequences: The role of repeats, *Phys. Rev. E* **50**, 5061–5071 (1994).
19. J. L. Oliver, P. Bernaola-Galvan, J. Guerrero-Garcia and R. Roman-Roldan, Entropic profiles of DNA sequences through caos game derived images, *J. Theoret. Biol.* **160**, 457–470 (1993).
20. P. R. Sibbald, S. Banerjee and J. Maze, Calculating higher order DNA sequence information measures, *J. Theoret. Biol.* **136**, 457–482 (1989).
21. F. A. James, Review of pseudo-random number generators, *Comput. Phys. Commun.* **60**(3), 329–344 (1990).
22. H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res.* **18**, 2163–2170 (1990).
23. R. Roman-Roldan, J. Martinez-Aroza and J. J. Quesada-Molina, Multiresolution information analysis for images, *Signal Process.* **24**, 77–91 (1991).
24. M. F. Barnsley, *Fractals Everywhere.* Springer-Verlag, New York (1988).
25. R. Roman-Roldan, P. Bernaola-Galvan and J. L. Oliver, Entropic features for sequence pattern through iterated function systems, *Pattern Recognition Lett.* **15**, 567–573 (1994).
26. P. Bernaola-Galvan, Contributions to sequence analysis using measures of entropic profiles. University of Granada. Internal report (September 1994).
27. W. Li, T. G. Marr and K. Kaneko, Understanding long-range correlation in DNA sequences. Preprint.
28. C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simon, H. E. Stanley and A. L. Goldberger. Mosaic organization of DNA nucleotides, *Phys. Rev. E* **49**(2), 1685–1689 (1994).

About the Author—RAMÓN ROMÁN-ROLDAN received his Ph.D. in physics in 1972 and presently he is Professor of Applied Physics at the University of Granada (Spain). He worked in Electronic Circuits and Systems for Instrumentation. His current field of interest is in Digital Image (Filtering, Segmentation) and DNA sequence analysis, mainly by using Information-Theoretic-based techniques. He now leads a team of researchers, also carrying out collaborative work with colleagues at other Departments in some interdisciplinar subjects.

About the Author—PEDRO A. BERNAOLA-GALVÁN is graduated in Physics (University of Granada, 1991). He now is elaborating his Doctoral Thesis on the application of Information Theory to sequence analysis. He teaches Thermodynamics at the University of Málaga, Spain.

About the Author—JOSÉ L. OLIVER graduated in biology in 1974 (University of Granada), obtaining a Ph.D. in Genetics in 1977 (University of Madrid). His fields of interest are theoretical biology and molecular evolution; he has addressed the analysis of molecular variability in natural populations, the evolution of gene expression after gene duplication and some problems of molecular phylogeny. More recently, he has been interested in the analysis of the large-scale structure of the genome. Professor J. L. Oliver now teaches Genetics and Biocomputing in Granada, having published more than 50 papers in the main journals of the field.