J. theor. Biol. (1993) 160,457-470

Entropic Profiles of DNA Sequences Through Chaos-game-derived Images

J. L. Oliver +, P. Bernaola-Galván ++, J. Guerrero-García $ AND R. Román-Roldán ++

t *Department of Genetics,*
$ *Department of Applied Physics and*
0 *Department of Theoretical Physics, Faculty of Sciences,*
University of Granada, Spain

(Received on 10 *October* 1991, *Accepted in revised form on* 11 *July* 1992)

A new method to determine entropic profiles in DNA sequences is presented. It is based on the chaos-game representation (CGR) of gene structure, a technique which produces a fractal-like picture of DNA sequences. First, the CGR image was divided into squares 4^{-m} in size (m being the desired resolution), and the point density counted. Second, appropriate intervals were adjusted, and then a histogram of densities was prepared. Third, Shannon's formula was applied to the probability-distribution histogram, thus obtaining a new entropic estimate for DNA sequences, the histogram ***entropy,*** a measurement that goes with the level of constraints on the DNA sequence. Lastly, the entropic profile for the sequence was drawn, by considering the entropies at each resolution level, thus providing a way to summarize the complexity of large genomic regions or even entire genomes at different resolution levels. The application of the method to DNA sequences reveals that entropic profiles obtained in this way, as opposed to previously published ones, clearly discriminate between random and natural DNA sequences. Entropic profiles also show a different degree of variability within and between genomes. The results of these analyses are discussed in relation both to the genome compartmentalization in vertebrates and to the differential action of compositional and/or functional constraints on DNA sequences.

Introduction

The information theory, developed primarily by Shannon & Weaver (1949), was mostly devoted to communication problems. Yet, it was being applied, often successfully, to many other fields where the information concept was essential, such as the interpretation of DNA sequences. Several authors (Gatlin, 1972; Guiasu, 1977; Lipman & Wilbur, 1983; Wicken, 1987; Brooks & Wiley, 1988; Sibbald *et al.*, 1989; Hariri *et al.*, 1990) present results concerning measurements (information, redundancy, divergence) obtained from DNA chains. The general aim is to look for quantitative expressions of the complexity of such chains.

None of the attempts up to now have been as fruitful as might be expected (see Hariri *et al.*, 1990, for a recent review), although DNA sequences contain a complete

! Author to whom correspondence should be addressed.

0022-5193/93/040457+ 14 $08.00/O

set of information concerning living beings. Surprisingly, we are not able to construct mathematical relationships for all the information we have regarding genetic systems. One reason for this may be that only one entropy (Shannon's), one probability distribution (that of the relative frequencies of oligomers in the DNA chain), and one measurement of divergence from randomness (redundancy) were taken into account in previous studies. But other alternative entropies, probability distributions, or divergence measurements can be considered within the wide field of information theory. Our work avoids the above difficulties by considering a different probability distribution : the histogram of density points in chaos-game-derived images for DNA sequences, thus obtaining a different entropy estimate, the *histogram entropy*. This leads to entropy profiles that, as opposed to previously published ones, clearly discriminate between random and natural DNA sequences.

We used two different methods in approaching this subject. Jeffrey (1990) 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. Basically, the densities of points in subsquares 4^{-m} in size correspond to the frequencies of oligomers m in length. On the other hand, some concepts from the field of multiresolution-information for digital images recently developed (Quesada-Molina & Roman-Roldan, 1989; Roman-Roldan *et al.*, 1991) are here applied to the CGR images of DNA chains. First, the entropy of the histogram of grey levels is translated into the entropy of the histogram of point densities in CGRs; second, different resolutions in images are translated into different subsquare sizes of the CGRs, which in turn correspond to different oligomer lengths.

We begin by presenting the main properties of the CGR images; next, we describe the adaptation of the theory for entropic measurements of histograms in images, and the construction of the entropic profiles for DNA sequences. Finally, we present the application of the method to DNA sequences retrieved from the EMBL (Stoehr & Cameron, 1991) and GenBank (Burks *et al.*, 1991) nucleotide data banks.

Main Properties of CGRs

The chaos game is an algorithm, generally controlled by a series of random numbers, which allows one to produce pictures (attractors) of fractal structures. Jeffrey (1990) proposed the use of DNA sequences, rather than random numbers, to control the chaos game (see Jeffrey's paper for details about the construction rules of CGRs). In short, each of the four corners of a square is labelled "a", "*c*" "g" or "*t*". The first base of the sequence, "*c*" for example, is plotted half way between the centre of the square and the "*c*" corner; if the next base is "*t*", for example, then a point is plotted half way between the previous point and the "*t*" corner; and so on. Each base in the sequence determines a new point in the CGR diagram. Thus, there are

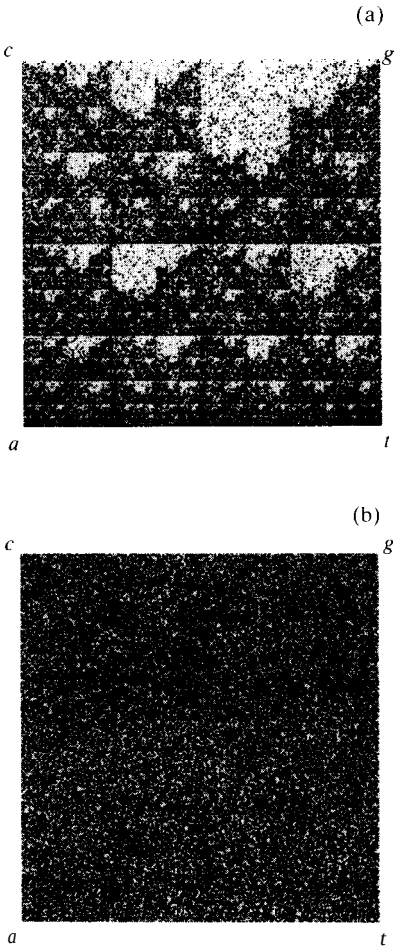

FIG. '1. (a) CGR obtained with the Jeffrey computer program for the human β-globin sequence **(HUMHBB)** retrieved the GenBank. (b) CGR image for a random sequence generated by means of the SDSE software package improved with the r-carry random-number generator. The lengths of both sequences were the same (73 326 nucleotides). Note in (a) the almost empty area in the upper right quadrant, as a result of the CpG dinucleotide avoidance shown by this sequence, and the repetitions of this "scoop" in the different quadrant and subquadrants, indicating repeated patterns in the gene, and giving a fractal-like structure to the CGR (Jeffrey, 1990).

as many points in the CGR as bases in the sequence. Figure 1 shows two CGR images generated by the Turbo Pascal program kindly provided by H. J. Jeffrey (Jeffrey, 1990). Figure I(a) shows the CGR for the human β-globin genomic region (HUMHBB). Figure l(b) shows the image corresponding to a random sequence of the same length obtained by means of the SDSE software package (Oliver *et al.*, 1989) improved with r-carry, a very long-period random-number generator (James,

1991). The uniformity of random CGR strongly contrasts with the apparent "structure" displayed by the CGR for the DNA sequence.

Four important features of the CGRs should be pointed out: (1) each point in the CGR can be allocated by knowing the corresponding base and the previous point only; (2) the allocation of each point allows one to retrace the whole sequence up to it and, therefore, the final point in the diagram reveals the total sequence; (3) if the CGR square is divided into small subsquares 4^{-m} in size, each m-square corresponds one to one with each of the different possible subsequences m in length. (4) The density of points (number of points in a given m-square) is the frequency of the corresponding m-subsequence in the chain; thus, relative densities of points correspond to probabilities of finding subsequences.

Histograms of Densities

The direct application of Shannon's formula to crude CGR subsquare counts (i.e. to subsequence frequencies), in order to obtain a direct estimate for the CGR entropy, would allow only a poor discrimination, if any, between random and DNA sequences, as has been repeatedly shown (see the review of Hariri et *al.,* 1990). Instead, we have extracted a density histogram from the CGR image; such histograms are graphics of the number of m-squares having a given density vs. the density of points. A histogram for each m can be drawn ($m = 0, 1, 2, \ldots, M$, there being 4^m squares in the CGR). Shannon's formula was then applied to the histogram probability distribution.

Histogram Entropy

Histograms were normalized by taking the relative number of occurrences (the actual numbers over the total number 4^m), resulting in the probabilities $P_{k,m}$ of finding squares with k points at resolution m.

Then, Shannon's entropy of the m-histogram for the CGR is

$$H_m = -\sum P_{k,m} \cdot \log_2 P_{k,m} \qquad k = 0, 1, 2, \ldots, K. \tag{1}$$

However, before this formula can be properly applied, some adjustments in the way in which histograms are constructed would be required in order for entropies to discriminate readily between random and DNA sequences.

Resolution Scale

There is a difference between the treatment of images and that of CGRs. The range of resolutions m (sizes of pixels) of interest in images is usually the closest to the binary pixels. But binary in CGR means such an $m = M$ for which each M-square would have either one or zero points. The M value would be such that the two closest points in the CGR would fall into different (and neighbouring) squares; it must be defined for each particular sequence. However, the range of resolutions (sizes of

squares) of interest for DNA sequences is in fact in the opposite extreme of the scale-the closest to the whole CGR image, which obviously has the maximum size with $m=O$. The resolution index series $m= 0, 1, 2, \ldots, m, \ldots, M$ corresponds to the square size series $1, 1/4, \ldots, 1/4^m, \ldots, 1/4^M$.

Random Versus Natural DNA Sequences

The critical point in looking for meaning in DNA chains (as in other fields of natural languages) is to find some measurements of how the given message differs from those that could have been generated in a purely random way. We are looking for the divergence between DNA and randomness, and therefore it is very important to see which histogram entropy corresponds to a random sequence of nucleotides. The CGR would be uniform for a fullly random sequence. In fact, this means that, for all the resolutions from $m = 0$ to co, the CGRs would be strictly uniform (the same density in all squares), the histograms would be punctual (only one density value with a non-zero probability), and the histogram entropy would be zero, *if the length of the sequence is infinite.* DNA sequences are finite in length, so the CGRs are not uniform, the histograms are not punctual, and the entropies are no longer zero. What would they be for random finite sequences? The question may be posed in terms of the probabilities of allocating one point after the other in particular squares of the CGR, and then asking for the probability of having a given number of points in any m-square. The result is given by the binomial distribution, 4^{-m} being the probability of having a point inside a given m-square, and $(1-4^{-m})$ outside of it:

$$P(N_{j,m}) = \binom{N}{N_j} \cdot (4^{-m})^{N_j} \cdot (1-4^{-m})^{(N-N_j)} \qquad (2)$$

where $P(N_{j,m})$ is the probability of having N_j points in an m-square (from the total number N).

As is well known, the above distribution can be approximated either by the Poisson one (for small probability 4^{-m}, m large) or by the Gaussian one (for a large mean $\mu = N \times 4^{-m}$). We adopt the latter approximation for all m for which $\mu \geq 50$. Current values for N in nucleotide data bank entries are up to 10^5. Gaussian approximation of random sequences with this length are valid up to a resolution of $m = 7$. Figure 2 presents the histograms corresponding to the CGRs shown in Fig. 1. While the histogram of the random sequence shows an apparent Gaussian structure, that corresponding to HUMHBB significantly departs from a normal distribution.

Adjusting the Density Scale

Histograms have a very fine density scale, as might be expected from the large N: there are many squares with density values which differ from one another. Thus, the entropy should always be great, despite histograms being sharp *or* spread. If so, we should not be able to distinguish between random and DNA sequences. In order to

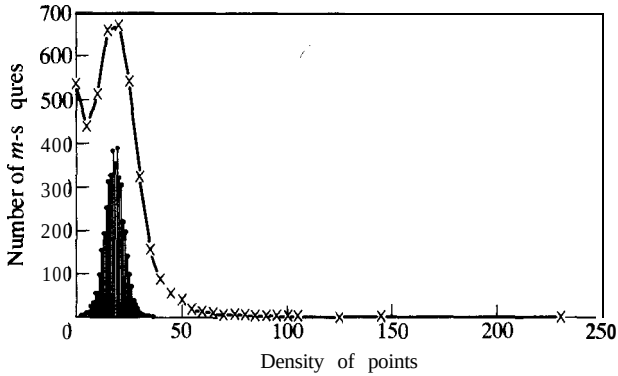

FIG. 2. Histograms of the number of *m*-squares having a given density vs. the density of points. Sequences were the same as in Fig. 1. The resolution chosen to elaborate this figure was $m = 6$. When a chi-square test was applied, the differences between the two distributions were statistically significant at the 0·01 level. (•), random; (x), HUMHBB.

have discriminant entropy values with respect to the relative sharpness of the histogram, we must adjust appropriate intervals for the density scale, thereby reducing the range of the probability distribution. The amount of reduction is a question of convenience.

We propose a density scale in which the histogram entropy for a random sequence becomes $H_m = 0{\cdot}25$ bits for all resolutions *m*. This option is justified by the following arguments: (1) all deviations of the DNA sequences apart from randomness would make the histograms more spread, and the entropies higher; (2) entropic profiles are intended for distinguishing order from randomness; (3) therefore, entropies of random sequences must be held low. The chosen value (0·25 bits) fulfils the proposed conditions, taking into account that the upper bound for H_m is *2m (see* below).

The Gaussian approximation allows us to determine the appropriate density scale. After easy computations (see Appendix), we get the following results:

$$\mu = N \cdot 4^{-m}; \qquad \sigma = \sqrt{\mu} = \sqrt{N} \cdot 2^{-m}$$
$$\Delta\rho = 4\sigma; \qquad n = 2\mu/4\sigma = \sqrt{N} \cdot 2^{-(m+1)} \tag{3}$$

μ being the mean value, ρ the density of points, σ the standard deviation, and n the number of intervals with a density scale in the 2μ range. This range must be broadened for high values of *m*, as long as the binomial distribution becomes better approximated by the Poisson than by the normal distribution. Histograms so constructed are shown in Fig. 3. The histogram for the coding sequence is clearly different from that for the random one.

Properties of Histogram Entropy

When eqn (I) was applied to the histograms constructed as described above, random and DNA sequences showed clearly distinguishable entropy values, and so the normalization of entropies was considered unnecessary.

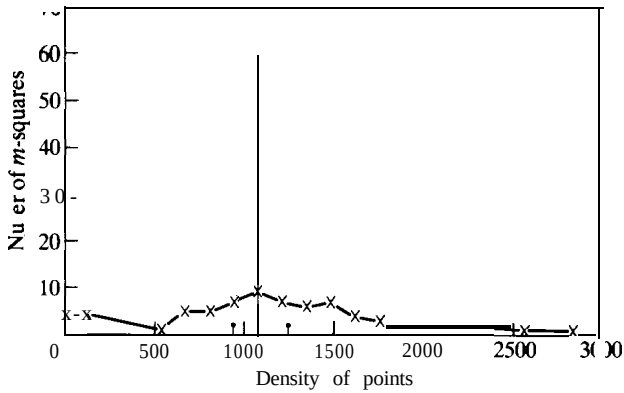

FIG. 3. Entropy histograms 4σ-scaled. Sequences were the same as in Figs 1 and 2; the chosen resolution was now $m = 3$. Differences between the two distributions were also significant at the 0·01 level. (•), random; (x), HUMHBB.

Let us see now why H_m is a measure of the deviation from a pure random sequence. First, the CGR for such a random sequence must have a uniform density of points, except for statistical fluctuations. So, the probability distribution for the different locations of points (which corresponds to the different subsequences in the chain) must also be uniform (except for statistical fluctuations again), and the corresponding CGR entropy must be near maximum. But we are dealing with another entropy, that of the histogram; this should now be punctual, that is to say, all squares should have the same density (or near, again) ; so, $P_{i,m} = 1$, $P_{j,m} = 0\ \forall j \neq i$, and therefore histogram entropy should be near zero. Thus, no constraints in the chain lead to zero histogram entropy. Conversely, as more constraints are imposed on the sequence (not random now), more variety occurs in the densities of the squares, and H_m rises (while the direct entropy of the CGR falls). Thus, we expect to have a measurement of the departure from randomness by computing histogram entropies. However, as opposed to the entropy which could be directly computed on the CGR image, and owing to the fact that we applied the Shannon's formula to the histogram probability distribution, H_m *goes* along with the amount of "structure" or non-randomness displayed by the DNA sequence.

Entropic Profiles

There is one histogram and also one entropy for each m. Therefore, we can make a representation of H_m *vs.* m, which will be called the *entropic profile* for the DNA sequence (Fig. 4).

Entropic profiles simultaneously display the constraints acting on the DNA chain at different oligomer lengths. In fact, the profiles include the constraints owing to deviations from equiprobability of individual nucleotide occurrences ($m = 1$), as well as higher-order constraints arising from unequal values of conditional probabilities,

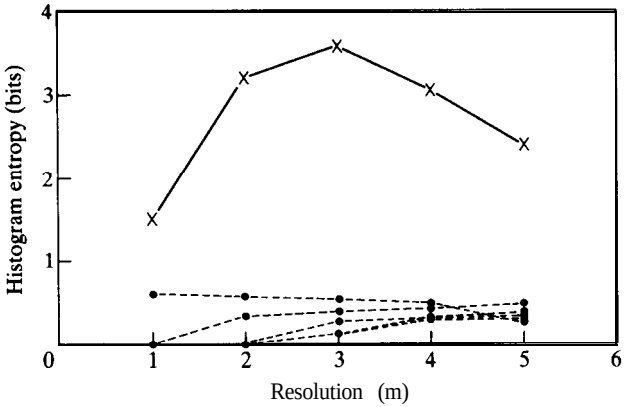

FIG. 4. The entropic profiles of HUMHBB as compared with those from several random sequences. (•), random; (×), HUMHBB.

and corresponding to the different resolution levels $(m = 2, .., M)$. Profiles shown in Figs 4-6 summarize the constraints from $m = 1$ to $m = 5$.

Properties of the Entropic Profiles

UPPER BOUND

For each resolution m, the maximum histogram entropy is $\log_2 4^m = 2m$, since 4^m is the number of squares, assuming that all of them have different densities. However, N may not be large enough to provide such a large number of squares of different values. In this case, the maximum entropy decreases to log L, L being the maximum number of squares with different densities from each other, $\rho = 1, 2, 3, \ldots, \rho_i, \ldots, L$, $N = \sum \rho_i$. Since the density series is arithmetic, it can be seen that $L \simeq \sqrt{2N}$.

LOWER BOUND

Minimum entropy is, for all resolutions, zero. This value corresponds to punctual histograms, having all-square densities in the same interval. A practical lower bound should be the entropy for a random sequence (0.25 bits in our so-defined intervals).

GENERAL PATTERN OF ENTROPIC PROFILES

Histogram entropy profiles, therefore, must run below the straight $H_m = 2m$ and above $H_m = 0{\cdot}25$. As m increases, a higher Markov order is considered. For any m-step further in the profile, dependence may exist or not; if it exists, the entropy must rise (increasing profile); if there is no dependence, the entropy must fall slowly; slow decrease of entropy is a result of the fractal nature of the CGRs, which causes any

feature in one m-square to extend (with attenuation) to successively smaller m-squares throughout the square image. The entropy should approximate 0·25 bits for a given m for which there is no longer any dependence.

Application to DNA Sequences

Several FORTRAN programs (running under DOS) were written to compute histogram entropies in both random and DNA sequences for the necessary range of resolutions. Figure 4 shows the entropic profile for a DNA sequence (HUMHBB) as compared to several simulated random ones of the same length. Statistical fluctuations can be appreciated between the profiles from random sequences. DNA and random sequences look very different, since the first departs markedly from randomness. This result strongly contrasts with previous results obtained by other authors. Thus, the plots of the fractional conditional divergence vs. the level of conditionality were unable to reveal differences between coding and random sequences (see fig. 3 of Hariri et al., 1990). The CGR approach we used, the extraction of a histogram of densities from the CGR image, and particularly the adjustment of the density scale, were perhaps the means which allowed us to obtain such discriminant entropic profiles between both types of sequences. In this way, our application of information theory to DNA sequences, as with other approaches finding order in biological sequences (Nussinov, 1980, 1981, 1990; Brendel *et al.*, 1986; Pietrokovski *et al.*, 1990; Peng et al., 1992), also proves capable of distinguishing between random and DNA sequences.

We have also computed the entropic profiles in several eukaryotic (Fig. 5) and prokaryotic (Fig. 6) DNA sequences retrieved from nucleotide data bases. Because of the restrictions imposed by the Gaussian approximation which we have used (see above), only those resolutions holding $\mu \geq 50$ were considered, resulting in $m \leq 5$. One problem of such restriction is that it somewhat increases the lower bound of DNA lengths to which our method is applicable-only relatively large DNA sequences can be properly analysed. However, this restriction also has the advantage of securing a higher statistical reliability to the entropy estimates obtained.

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 ECUNC, and the *Xenopus* mitochondrial genome, all showing a maximum for m = 3. This rise of the entropic profiles up to $m = 2, 3$ might be interpreted in several ways. First, an explanation could be that it is an artefact resulting from the computational method used: however, such an entropy maximum does not arise when it is applied to random sequences of the same length (Fig. 4). Second, a maximum at m = 2 could also be attributed to the well-known avoidance of CpG dinucleotides in some genomes (Bird, 1980), but DNA sequences without such avoidance, as the mitochondrial genomes [Fig. 6(a)] or *Escherichiu coli* sequences [Fig. 6(b)] also show peaks at m = 2. Lastly, the possible influence of the triplet code in provoking maximum entropy measurements at m = 3 was also discarded upon the finding that "pure" coding sequences as the large mature mRNA for human dystrophin (HSDMDR) show an entropy maximum at $m = 2$

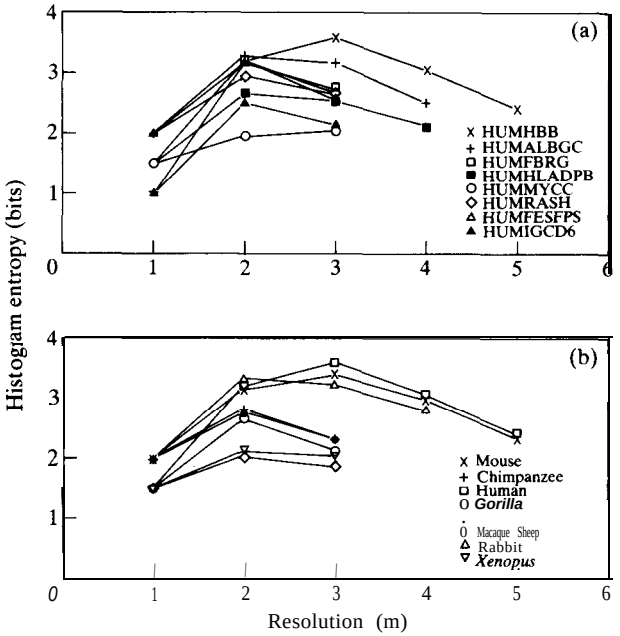

FIG. 5. Variability in the entropic profiles of different DNA sequence sets from vertebrate nuclear genomes. (a) Several human sequences. (b) Sequences of the β-globin genomic region in different vertebrate species (human, HUMHBB; chimpanzee, PTGLB1; gorilla, GGHBPSBD; macaque, MCBGLOG; mouse, MMBGCXD; sheep, OABBGLOB; rabbit, OCBGLOO1; *Xenopus*, XTGLB).

(results not shown). Therefore, our results seem to suggest the actual existence of a maximum of dependence at 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 $m = 3$. Regularities in DNA sequences with periods of three bases are well known (Shepherd, 1981). More recent investigations (Li, 1991; Tsonis *et al.*, 1991) also showed evidence for periods two and three in natural DNA sequences when they were considered in a two class, purine and pyrimidine, representation of the four bases. Our results were also consistent with the reported second to fourth Markov orders (Blaisdell, 1984; Phillips *et al.*, 1987; Arnold *et al.*, 1988; Hong, 1990) for DNA chains. However, they strongly contrast with the prediction derived from the hierarchical information theory that "... the shortest subunits of DNA ought to show very little inhomogeneity (randomization at lowest levels of the hierarchy) while longer substrings should show pronounced inhomogeneity (constraints at higher levels of the hierarchy)" (Brooks & Wiley, 1988: 117).

For a given resolution, histogram-entropy profiles show different degrees of variability depending on the genome considered. The sample of vertebrate sequences analysed here (Fig. 5) was more variable than the prokaryotic one (Fig. 6). Furthermore, the higher variability of eukaryotic entropic profiles is apparent at two levels: among different sequences from a same genome [several human entries, Fig. 5(a)]

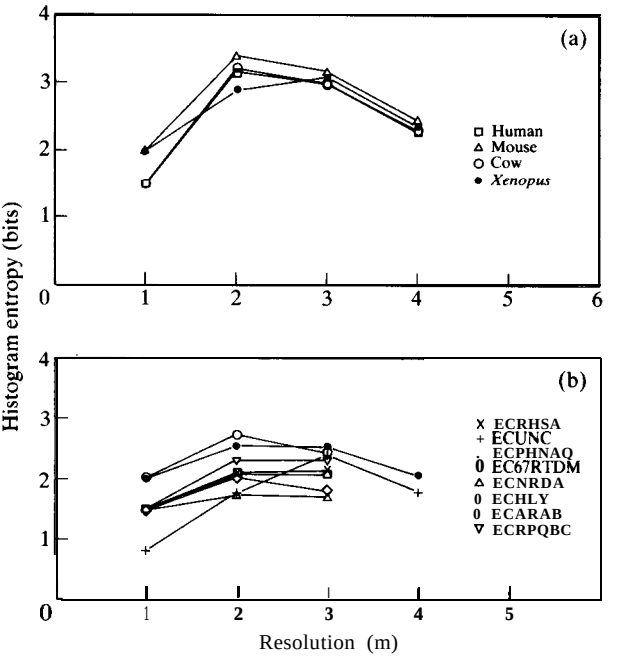

FIG. 6. (a) Variability in the entropic profiles of complete mitochondrial genomes from vertebrates (human, HUMMT; mouse, MIMM; cow, BOVMT; *Xenopus*, XELMTCG). (b) Entropic profiles corresponding to different ***E. coli*** genes.

and among homologous genes from different genomes [β-globin regions from several vertebrate species, Fig. 5(b)]. The genomes from vertebrate mitochondria show the lowest variability in entropic profiles [Fig. 6(a)], while *E. coli* sequences show an intermediate one [Fig. 6(b)].

We think that the variability in entropic profiles revealed by our analysis might be related to the differential action of compositional and/or functional constraints. In particular, the differential degree of compositional heterogeneity exhibited by a genome seems to be a critical factor. In fact, the genomes of higher eukaryotes, but not those from bacteria or mitochondria, are compositionally heterogeneous (Bernardi et al., 1985). It is known that the compositional compartmentalization of a genome plays an important role in genome structure and function, leading to differential constraints on G + C content and codon usage, and, therefore, on nucleotide frequencies (Bernardi *et al.*, 1985). This variability of constraints acting on nucleotide composition could explain the wider variability in entropic profiles we found among the different genomic regions from vertebrate compartmentalized genomes.

Conclusions

Both the CGR of gene structure and some techniques from the field of multiresolution-information for digital images allows us to develop a new estimate for entropy

in DNA sequences. Histogram entropy goes along with the level of constraints, or negentropic effects, on the DNA chain: no constraints lead to zero histogram entropy, and the higher the constraints, the higher the value obtained. The method derived to draw the entropic profile of a DNA sequence is computed easily enough (through the CGR image), and it unifies the mathematical treatment given to the different sequence divergence levels. When it is applied to DNA sequences, the following conclusions emerge: (1) entropy profiles markedly differ among (a) random and natural sequences, and (b), coding sequences from both the same and different genomes; (2) vertebrate nuclear genomes show more variable entropic profiles than bacterial and mitochondrial ones.

Since there is no theoretical limit to the number of points which can be plotted on CGRs, and since the reliability of histogram entropy estimates increases with the sequence length, the method described here is particularly useful for analysing very large sequences. Such an approach would be of particular interest as larger sequences become available in DNA data banks, providing a tool capable of summarizing the complexity of large genome regions (i.e. isochores) or even entire genomes at different resolution levels.

Hariri et al. (1990) have recently expressed a reluctance to accept the conclusion that Shannon/Gatlin type calculations are not productive. The method reported here perhaps provides a way of escaping such a negative conclusion.

We thank Dr H. J. Jeffrey for kindly providing us with the computer program to trace CGRs. We also thank David Nesbitt for technical assistance in compiling this paper. This work was partially supported by grants PB90-0847 to J.L.O. and TIC91-646 to R.R.R. from the DGICYT of the Spanish Government.

REFERENCES

ARNOLD, J., CUTICCHIA, A. J., NEWSOME, D. A., JENNINGS, W. W. & IVARIE, R. (1988). Mono- through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis. *Nucl. Acids Res.* **16, 7145-7158.**

BERNARDI, G., OLOFFSON, B., FILIPSKI, J., ZERIAL, M., SALINAS, J., CUNY, G., MEUNIER-ROTIVAL, M. & RODIER, F. (1985). The mosaic genome of warm-blood vertebrates. *Science* **228, 953-958.**

BIRD, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucl. Acids Res. 8,* 1499–1504.

BLAISDELL, B. E. (1984). Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eukaryotic nuclear DNA sequences both protein-coding and non-coding. *J. molec. Evol.* **21, 278-288.**

BRENDEL, V., BECKMANN, J. S. & TRIFONOV, E. N. (1986). Linguistic of nucleotide sequences: morphology and comparison of vocabularies. *J. Biomol. Struc. Dyn. 4,* 11–21.

BROOKS, D. R. & WILEY, E. 0. (1988). *Evolution as Entropy: Toward a Unified Theory of Biology,* **2nd edn. Chicago, IL: University of Chicago Press.**

BURKS, C., CASSIDY, M., CINKOSKY, M. J., CUMELLA, K. E., GILNA, P., HAYDEN, J. E.-D., KEEN, G. M., KELLY, T. A., KELLY, M., KRISTOFFERSON, D. & RYALS, J. (1991). GenBank. *Nucl. Acids Res.* **19,** 2221–2225.

GATLIN, L. (1972). *Information Theory and the Living System. New* **York: Columbia University Press.**

GUIASU, S. (1977). *Information Theory with Applications. New* **York: McGraw-Hill.**

HARIRI, A., WEBER, B. & OLMSTED, J. (1990). On the validity of Shannon-information calculations for molecular biological sequences. *J. theor. Biol.* **147, 235-254.**

The mean value is, obviously, the total number of bases N over the total number of squares 4^m: $\mu = N \times 4^{-m}$. For the binomial distribution, the standard deviation is

$$\sigma = \sqrt{N \,.\, P \,.\, (1-P)}, \qquad P = 4^{-m}. \tag{A.2}$$

For $m \geq 2$, the above expression can be approximated by

$$\sigma \simeq \sqrt{N \,.\, P} = \sqrt{\mu}. \tag{A.3}$$