Estimating the Entropy of DNA Sequences

Article *in* Journal of Theoretical Biology · November 1997

DOI: 10.1006/jtbi.1997.0493 · Source: PubMed

2 authors, including:

Hanspeter Herzel
Humboldt-Universität zu Berlin
446 PUBLICATIONS **14,629** CITATIONS

SEE PROFILE

J. theor. Biol. (1997) **1888,** 369–377

Estimating the Entropy of DNA Sequences

ARMIN O. SCHMITT*‡ AND HANSPETER HERZEL†

**MPI für molekulare Genetik, Ihnestraße* 73, *D*-14195 *Berlin, and †Institut für Theoretiche Biologie, Humboldt-Universität zu Berlin, Invalidenstraße* 43, *D*-10115 *Berlin, Germany*

(*Received on* 6 *March* 1997, *Accepted in revised form on* 12 *June* 1997)

The Shannon entropy is a standard measure for the order state of symbol sequences, such as, for example, DNA sequences. In order to incorporate correlations between symbols, the entropy of n-mers (consecutive strands of n symbols) has to be determined. Here, an assay is presented to estimate such higher order entropies (block entropies) for DNA sequences when the actual number of observations is small compared with the number of possible outcomes. The n-mer probability distribution underlying the dynamical process is reconstructed using elementary statistical principles: The theorem of asymptotic equi-distribution and the Maximum Entropy Principle. Constraints are set to force the constructed distributions to adopt features which are characteristic for the real probability distribution. From the many solutions compatible with these constraints the one with the highest entropy is the most likely one according to the Maximum Entropy Principle. An algorithm performing this procedure is expounded. It is tested by applying it to various DNA model sequences whose exact entropies are known. Finally, results for a real DNA sequence, the complete genome of the Epstein Barr virus, are presented and compared with those of other information carriers (texts, computer source code, music). It seems as if DNA sequences possess much more freedom in the combination of the symbols of their alphabet than written language or computer source codes.

Introduction: Order and Disorder of Sequences

Intuitively, a sequence S with symbols s_i $(i = 0. . .\lambda)$ from a given alphabet $\aleph$ is considered as ordered when it is periodical or when some symbols or sub-sequences occur repeatedly. On the other hand, it is considered disordered, when all of its symbols and combinations of symbols occur at equal frequencies. The measure for order or disorder in sequences is the Shannon-entropy (Shannon, 1948)

$$H = -\sum_i p_i \log p_i, \tag{1}$$

where i extends over all symbols of the alphabet, and p_i is the probability that symbol s_i occurs at any position. H is maximal when all symbols occur at equal probability $p_i = 1/\lambda$. (The maximum is unity if the basis of the logarithm is chosen to be λ.) The minimum is taken on if one symbol occurs at probability 1, the others being "forbidden"; then $H = 0$ holds.

Equation (1) gives information about the distribution of the symbols—do they occur at equal probability (equi-distribution, high entropy value) or do some symbols prevail (skew distribution, low entropy value)? It does, however, not reflect the correlations between the symbols. In order to incorporate them eqn (1) has to be generalized to the so-called block-entropies:

$$H_n = -\sum_i p_i^{(n)} \log p_i^{(n)}, \tag{2}$$

where $p_i^{(n)}$ are the probabilities of the combinations of n symbols. The index now extends over all λ^n possible combinations. Henceforth, combinations of n sym-

‡ Author to whom correspondence should be addressed.
E-mail: schmitt@mpimg-berlin-dahlem.mpg.de

0022–5193/97/190369 + 00 $25.00/0/jt970493

bols will be called n-words, n-blocks, or sub-strings of length n.

Correlations between symbols are reflected in a sub-linear growing of H_n with block-length n, for example: $H_2 < 2H_1$. Several scaling laws for block-entropies were suggested by Ebeling (1993).

Experience tells us that informational sequences carrying meaningful messages are neither completely ordered nor completely random. For example, in the English language the letter "e" is the most frequent, and the letter "z" is the rarest. This is reflected by H_1 being much smaller than 1. Similarly, the combinations "of", "th" or "un" are much more frequent than "fo", "ht" or "nu", which is reflected by H_2 being much smaller than the maximal value of 2.

Derived quantities which are sometimes used to highlight certain aspects of block entropies are the differential entropies

$$h_n = H_{n+1} - H_n \tag{3}$$

and the entropy of the source

$$h := \lim_{n\to\infty} h_n. \tag{4}$$

An interesting property of this quantity is that it allows the calculation of how "dense" real sequences of length l are in the λ^l-dimensional sequence space of all possible sequences. There are effectively $\lambda^{h\cdot l}$ sequences ("most probable" sequences), and their density is $\lambda^{l(h-1)}$. The density for 100 amino acid long polypeptides was estimated by Strait & Dewey (1996) to be 10^{-57}, i.e. only a tiny fraction of all possible amino acid combinations were realized by nature.

The exact determination of block-entropies for sequences in praxis, however, meets almost unsurmountable difficulties for large block-lengths. For the application of eqn (2) the knowledge of the exact probabilities is of crucial importance. But even for the longest available electronically stored sequences, the number of possible symbol combinations surpasses the number of extracted sub-strings. For example, there are about one billion DNA oligomers of length 15 ($\lambda = 4$), and the longest DNA sequences available in databanks are of the order of hundred thousand base pairs long. This means the average occurrence drops below 1, and the probability of the n-blocks can no longer be reliably approximated by relative frequencies

$$f_i^{(n)} = \frac{k_i}{N}, \tag{5}$$

* To be precise, $N - n + 1$ overlapping n-words can be extracted from a sequence of length N; for large N and small n this number can be approximated by N.

where k_i is the occurrence number of n-word number i and N is the length of the sequence.* Hence, the approximation of block-entropies by entropies from frequencies, which we call *observed* entropies,

$$H_n^{\text{obs}} = -\sum_i f_i^{(n)} \log f_i^{(n)}, \tag{6}$$

is no longer legitimate either. The block-entropies from frequencies underestimate the block-entropies systematically; this finite sample effect is discussed in, for example, Herzel (1988), Grassberger (1988), Schmitt *et al*. (1993), Herzel *et al*. (1994a) and Schmitt (1995).

It has been shown (Schmitt, 1995), however, that the block-entropies from frequencies observed in different realizations of the same length and with the same underlying probability distribution are highly concentrated around the mean $\langle H_n^{\text{obs}} \rangle$, the relative standard deviation being less than 0.1%. Thus, although observed entropies serve as an estimation for entropies only in the case of rich statistics (i.e. even the rarest events are realized a couple of times), they are nevertheless a characteristic signature of the underlying (unknown) probability distribution. This fact will be exploited for its reconstruction in the next chapter.

A Correction Method Based Upon the Maximum Entropy Principle

REVIEW: THE FLAT-DISTRIBUTION METHOD

Several methods have been developed to correct the finite sample effect (Herzel, 1988; Grassberger, 1988; Schmitt *et al*., 1993). The method sketched in the following is an improvement of the so-called flat-distribution *ansatz* (henceforth called FD-*ansatz*) presented by Schmitt *et al*. (1993).

The FD-*ansatz* is based on the McMillan theorem of equi-distribution. This theorem states that most of the combinations of symbols (the "most probable" combinations) occur at almost identical frequency; the relative number of those occurring at greater or smaller frequency than these combinations tend towards 0 with increasing block-length n (McMillan, 1953).

Thus, the approximation of the sought probability distribution by a staircase function

$$q_i = \begin{cases} \dfrac{1}{N^*} & : \quad i \leqslant N^* \\ 0 & : \quad i > N^* \end{cases} \tag{7}$$

is justified. One parameter, the number of effective words N^*, $(0 < N^* < \lambda^n)$ is free, and this parameter is determined in such a way that the entropy one would expect to observe from samples drawn from $\{q_i\}$ matches exactly the entropy from frequencies of the real sequence. Note, that the index i is now no longer assigned to individual n-words, as this was the case for the probabilities p_i.

Of course, there are clear limits of this correction technique when the underlying probability distributions deviate strongly from equi-distributions. Since earlier results (Herzel *et al*., 1994b) indicate that DNA sequences have entropies close to their maximal value, it seems reasonable to apply this approach for DNA sequences.

RANKED DISTRIBUTIONS OF n-WORDS

Experience shows, however, that information-carrying sequences are clearly not random. Exactly or approximately repeating segments (repeats) are a quite common phenomenon in DNA (Lewin, 1994), and to some extent this is also true for literary texts or computer source codes (Schmitt, 1995). Figure 1 presents a comparison of the rank-ordered distribution of words of length 8 from the yeast chromosome III (length 315338 bp) and from a random sequence of equi-distributed symbols of the same length.

The most frequently observed 8-words occur much more often in the yeast sequence than in a comparable random sequence, whereas the reverse is true for rare 8-words. The mean frequency of the most frequent

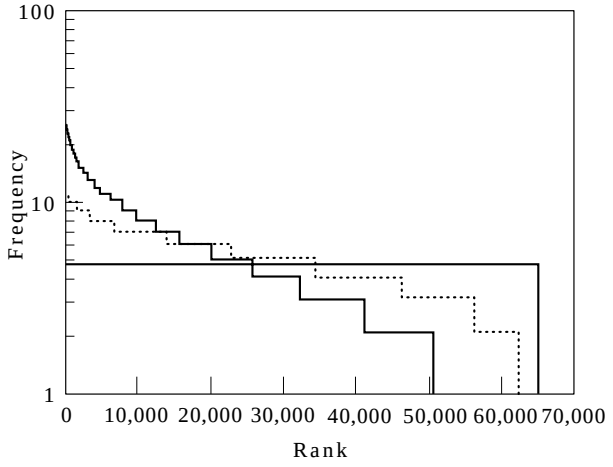

FIG. 1. Rank ordered distribution of 8-words from the yeast chromosome III and from a random sequence of the same length. Frequent identical fragments in the yeast sequence cause a dip at the highest ranks (—); statistical fluctuations cause a much less pronounced effect (. . .; from a random sequence of the same length). The expected occurrence is represented by the perpendicular full line. The frequency is in log-scale.

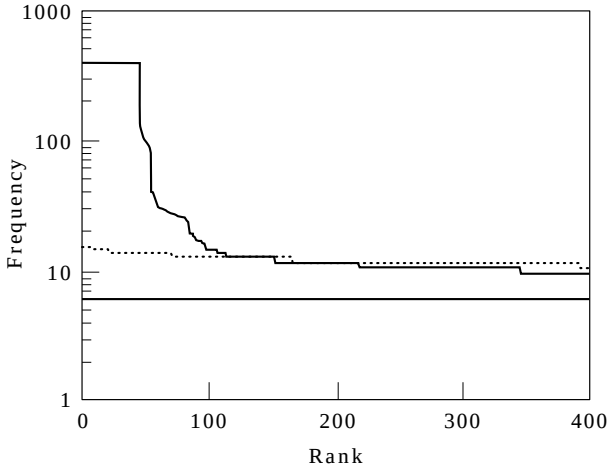

FIG. 2. The rank ordered distribution of the most frequent 7-words taken from a model string of length 100000 with 400 identical fragments of length 50 interspersed at random positions (—) and from a random sequence of the same length (. . .). The horizontal line marks the expectation value for the random case. The 44 different 7-words from the repeated fragments form a distinct plateau.

8-word determined from 1000 comparable random sequences was 16.79, with a standard deviation of 0.06. This is significantly less frequent than the top ranking 8-words from the yeast sequence which occur 30 to 50 times.

A second example of rank distributions with marked deviations from the average occurrence at the top ranks is the distribution of 7-words from a model string as described in Herzel *et al*. (1994b). In that work a model of DNA is presented where a number of identical fragments is embedded in a random sequence. The model string of length 100000 symbols analysed here was a random equi-distributed sequence with 400 identical fragments of length 50 interspersed at random positions. Figure 2 shows that the repetitive fragments induce a very pronounced plateau at the top ranks.

These two examples demonstrate that for many interesting sequences the rank ordered occurrence distribution of n-words is quite far from that of a random (or nearly random) process, and, thus, the equi-distribution predicted by the McMillan theorem will be reached only for very long n-blocks. There is a large number of blocks that will occur much more frequently than standard words, and neglecting them would be an oversimplification.

Consequently, the FD-*ansatz* (7) can be improved by taking into account the peak observed at the top ranks. The probability of words occurring in the order of 10 or 100 times—the three most frequent 8-words of the yeast sequence, for example, occur 307, 217,

and 145 times, resp.—is excellently estimated by their relative frequencies f_i [eqn (5)].

RECONSTRUCTION OF THE UNDERLYING PROBABILITY DISTRIBUTION

Hence, a combination of two approaches should be a better *ansatz* to fit the sought-for probability distribution q_i: for the i_0 most frequent words (top ranks) the relative frequencies are accepted as probabilities, and the rare words (low ranks) follow a staircase distribution (the index n indicating the block-length will be dropped henceforth):

$$q_i = \begin{cases} f_i & : \quad 1 \leqslant i \leqslant i_0 \\ \dfrac{Q}{L} & : \quad i_0 < i \leqslant L + i_0, \end{cases} \tag{8}$$

where

$$P = \sum_{i=1}^{i_0} f_i \tag{9}$$

is the mass of probability distributed among the i_0 most frequent words, and

$$Q = 1 - P \tag{10}$$

is the amount of probability shared equally by the L standard words. The flat distribution (7) can, of course, be seen as a special case of eqn (8) for $i_0 = 0$.

Two parameters, i_0 and L, can be varied freely in eqn (8), under the condition that the normalization

$$\sum_{i=1}^{M} q_i = 1 \tag{11}$$

be respected. $M = i_0 + L$ is the total number of occupied (non-vanishing probability) ranks. Note, that fixing i_0 is equivalent to fixing P and Q, since P is a monotonically increasing function of i_0.

The constraint

$$H^{\exp}[\mathbf{q}] = H^{\text{obs}}[\mathbf{f}], \tag{12}$$

says that the observed entropy of a given sample is a characteristic number of the underlying probability distribution $\mathbf{p}$ that also the test-distribution $\mathbf{q}$ is demanded to have. $H^{\exp}[\mathbf{q}]$ is the expected entropy of the test-distribution.

It does not have to be determined by Monte–Carlo simulations, i.e. simulated drawing from the distribution [eqn (8)], but can be calculated in the following way.

The probability of the ith n-block with probability q_i to occur k times for a sample of length N is:

$$\rho_i(k) = \binom{N}{k} q_i^k (1 - q_i)^{N-k}, \tag{13}$$

which is the well-known binomial distribution.

When a word is drawn k times its entropy contribution amounts to

$$\Delta H^{\exp} = -\frac{k}{N} \log \frac{k}{N}. \tag{14}$$

Since the ith n-block can occur $k = 1, 2, \ldots, N$ times—occurring 0 times does not contribute to the entropy—with the probability $\rho_i(k)$ we expect the total entropy produced by this word to be

$$H^{\exp,i} = \sum_{k=1}^{N} \rho_i(k) \cdot \Delta H^{\exp}. \tag{15}$$

We obtain finally the expected entropy by summing up all the contributions from each individual element:

$$H^{\exp} = \sum_{i=1} H^{\exp,i} = \sum_{i=1}^{M} \sum_{k=1}^{N} \rho_i(k) \cdot \Delta H^{\exp}. \tag{16}$$

The constraint eqn (12) eliminates one degree of freedom: the choice of i_0 determines L; the crossover index i_0 can still be varied freely.

The question of how to fix this parameter is answered by the Maximum Entropy Principle (hereafter denoted MEP).

This principle states that, out of a number of possible probability distributions consistent with one (or several) constraints, the one disposing the greatest entropy value is the most likely one. The observations, mostly average values, are incorporated in the constraints, and the most random distribution (that with the highest entropy) explains these observations most naturally.

Thermodynamic functions can be calculated by means of this principle introducing fixed variables of state (e.g. pressure, temperature) as constraints. In the field of image reconstruction, the "best" picture can be inferred from a blurry picture according to this principle. A number of applications of the maximum entropy principle is described in Justice (1986) and Kapur (1989).

The MEP is intimately connected with Bayesian statistics (Wickmann, 1990). Its relationship to this approach and to classical statistics is elucidated, e.g. by Jaynes (1983). Generally speaking, the MEP constitutes a method to describe systems mathematically whose properties are only vaguely known, or that can only be characterized by average values.

SKETCH OF THE ALGORITHM USING THE MEP AND ITS TEST

An algorithm to determine the distribution $\mathbf{q}_{\mathrm{ME}}$ of maximum entropy is sketched below.

- Determine L for each of the possible values of $i_0 = 0, 1, \ldots$ so that the expected entropy matches the observed entropy according to eqn (12).
- Calculate the entropies of the thus obtained distributions $\mathbf{q}_{i_0}$: $H_{i_0} = H[\mathbf{q}_{i_0}]$ according to eqn (16) and store them in an array.
- Determine the maximum value out of all H_{i_0}, $i_0 = 0, 1 \ldots$. The index maximizing this series of entropy values, i_{ME} characterizes the distribution $\mathbf{q}_{\mathrm{ME}}$ of maximal entropy compatible with the constraint eqn (12).

Now we test the described method using stochastic model sequences with analytically known block-entropies.

The test-sequences consisted of the aforementioned random sequence with 400 or 800 randomly interspersed repeats of length 50; thus, 20 or 40% of the total sequence consisted of repeats. Figure 3 shows the convincing result of this hybrid-approach. While the FD-approach fails—much the same as the "uncorrected" values do—from $n = 7$ on, the MEP-corrected values follow the theoretical curve exactly up to $n = 14$, thus doubling the range of correct entropy calculation. It is worthwhile noting that the expectation value of a block of length 14 symbols in a sequence of length 100000 is as low as $3.7 \cdot 10^{-4}$. The theoretical values are given (Herzel *et al.*, 1994b) approximately by

$$h_n = 2[1 - \rho(l - k_c)\Theta(n - k_c)], \tag{17}$$

where

$$k_c = \frac{1}{2}\log\frac{1}{\rho} \tag{18}$$

is the number of symbols which are necessary to identify a repeat, l is its length, and ρ is the probability to find its first symbol at an arbitrary site in the whole sequence. $\Theta(x)$ is the well-known staircase-function:

$$\Theta(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x \leqslant 0 \end{cases} \tag{19}$$

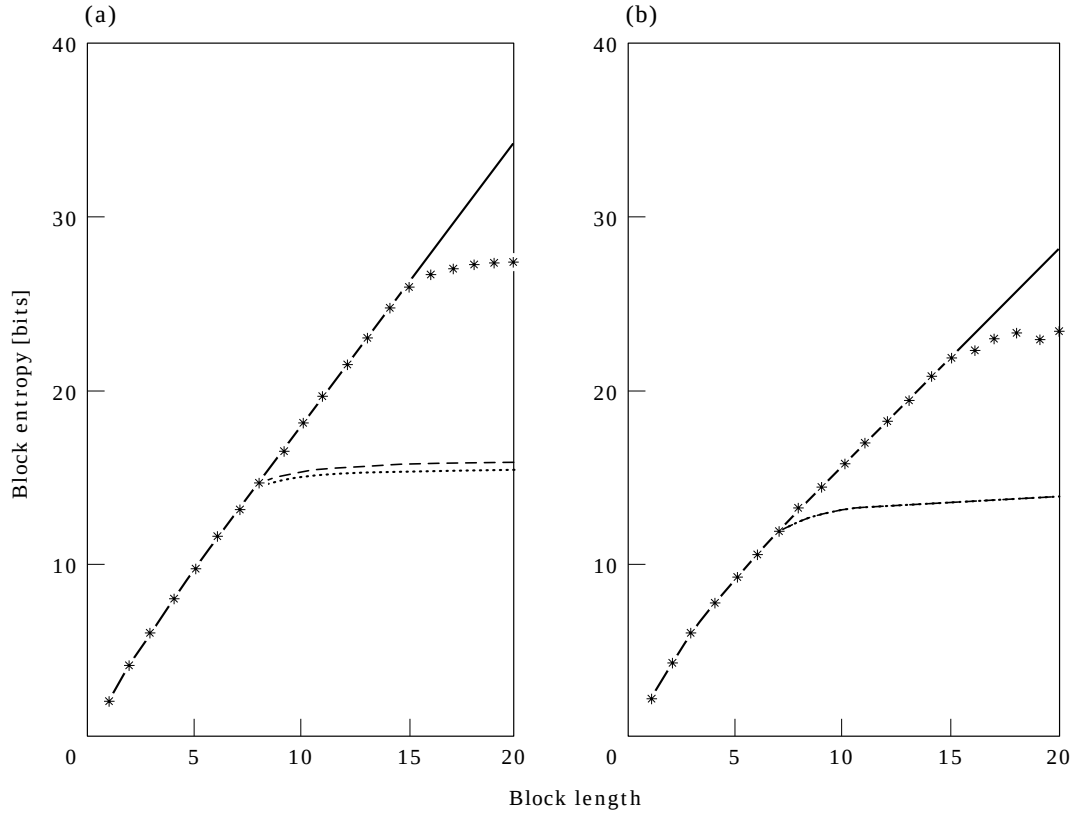

FIG. 3. While the correction method based upon the flat distribution approach (– – –) proves uneffective for a random sequence containing identical repeats, the method based on the Maximum Entropy Principle (* * *) yields excellent results up to block lengths of $n = 15$. The random sequence (total length 100000 symbols) contained 400 (a) or 800 (b) identical repeats of length 50. The uncorrected values are given as dotted lines (. . .) and the theoretical values are denoted by a full line (—).

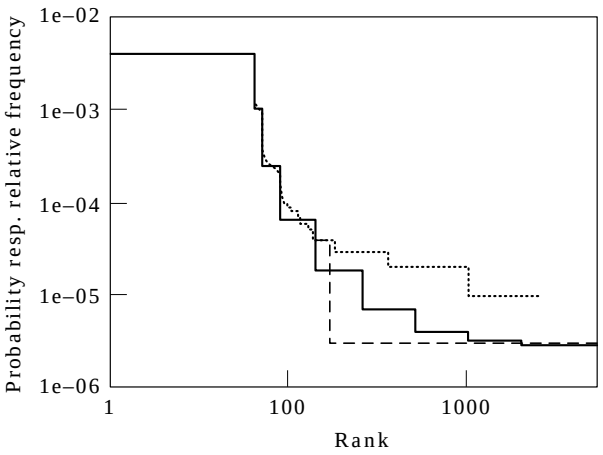

FIG. 4. Log-log plot of the probability distribution $\mathbf{q}_{\mathrm{ME}}$ for 9-words constructed by means of the MEP algorithm (– – –). It is in very good accordance with the theoretical distribution (—); ranks 1 to 309 are identical with relative frequencies. While relative frequencies (. . .) coincide practically with the theoretical probabilities for the top ranks, they fail to approximate them for the low ranks. Same model string as specified in Fig. 3 (a).

Tests have shown that this correction method provides also correct results when it is applied to a random sequence with no interspersed repeats. Then i_0 is determined to zero (in the ideal case) or to a small number close to zero.

Let us also study the distribution of maximal entropy $\mathbf{q}_{\mathrm{ME}}$ itself which was generated by this algorithm. We are interested in words of length 9 of our model sequence. The relative frequencies of the 309 highest ranks were accepted by the algorithm as probabilities; the probabilities of 261682 other substrings (the standard words) were determined to be uniformly $3.07 \cdot 10^{-6}$.

Figure 4 shows that the constructed probability distribution resembles much more the theoretical distribution (Herzel *et al*., 1994b) than the relative frequencies do. The calculated entropy is $H_9^{\mathrm{MEP}} = 16.40$ bits, as compared with the theoretical value $H_9^{\mathrm{theo}} = 16.39$ bits according to eqn (17). The uncorrected value is 14.80 bits, and that corrected according to the equi-distribution approach 15.07 bits. (The base of the logarithm is 2 in this example.)

Applications

BLOCK ENTROPIES OF LITERARY TEXTS

The MEP correction method introduced in the previous section is first applied to literary texts because their block-entropies have been estimated in guessing experiments (Shannon, 1951). To imitate these experiments the text of the novel "Alice in Wonderland" was coded over a 27-symbol alphabet (26 letters and the space), producing a sequence of 134454 symbols. The block entropies corrected according to the MEP correction method are given in Fig. 5, together with the raw values, the values corrected according to the FD-correction, and the arithmetic means of Shannon's upper and lower bounds. Furthermore, the experimental block entropies of the German language according to Völz (1990) and values from Pöschel *et al*. (1995) are presented.

While the FD-approach and the approach suggested in Pöschel *et al*. (1995) prove uneffective for this type of sequence, the MEP corrected values are compatible with the experimental results. Uncorrected entropies from relative frequencies are a good approximation to the experimental values only up to block length $n = 5$.

Unfortunately, no further specification about the technique applied to obtain the data is given by Völz (1990) so that the almost parallel, but shifted, graph for his experimental values can be either attributed to the different language (German instead of English) or to a different method.

BLOCK ENTROPIES OF DNA SEQUENCES

Under the assumption that real DNA sequences are not too different from random strings with interspersed repeats it is justified to apply the algorithm described in the previous section to DNA sequences. To obtain good statistics very long contiguous strings have to be chosen: the yeast chromosome III sequence

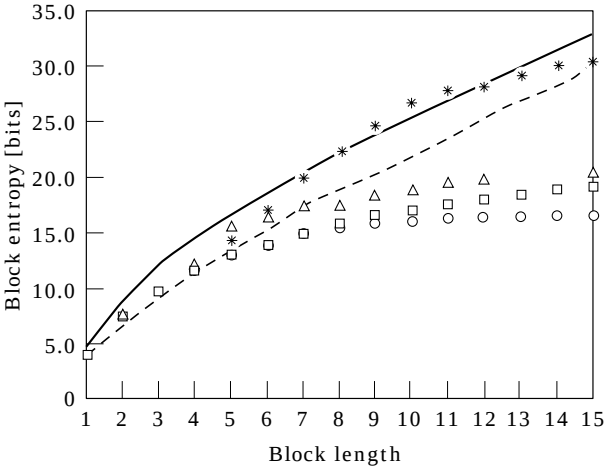

FIG. 5. The artifact induced by the finite size of samples in entropies for literary texts is best compensated for using the MEP correction method (*, see text). The values thus corrected are consistent with experimental values given in (Shannon, 1951) (– – –) and (Völz, 1990) (—; in German). The flat distribution correction (□) and the correction according to (Pöschel *et al*., 1995) (△) do not sufficiently compensate for the finite sample effect. Observed values are denoted by circles (○).

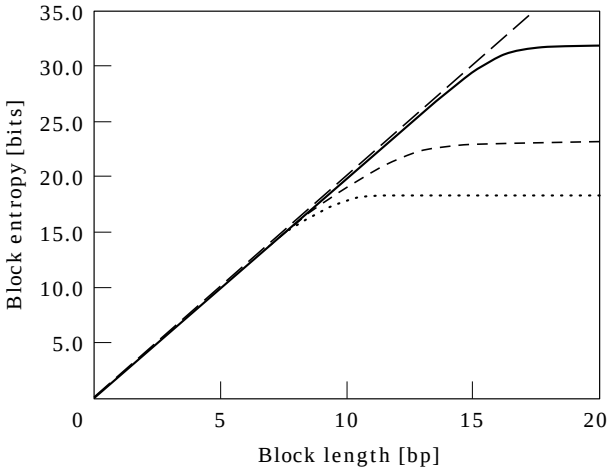

FIG. 6. Almost no internal correlations are exhibited by the symbols of the DNA sequence of yeast chromosome III (—): the block entropies lie only slightly below those of a random equi-distributed sequence (– –). Note, that the raw values (. . .) and those corrected according to the flat distribution correction method (– – –) imply the appearance of strong correlations at about $n = 8$; the bending off, however, is an artifact due to the finite size of the sample.

of length 315338 base pairs (Oliver *et al*., 1992) and the Epstein-Barr virus genome of length 172281 bp (Baer *et al*., 1984).

Figure 6 shows the block entropies for the yeast sequence corrected according to the FD-correction method and the MEP correction method along with the raw values and the graph for a random sequence $H_n^{\text{rand}} = 2 \cdot n$. While the raw values are trustworthy until $n = 7$ and the values corrected according to the FD-method up to $n = 9$, the bending off typical for the finite sample effect appears only at $n = 15$ with the MEP corrected values, thus doubling the range of estimation. The MEP corrected values are definitely to be rejected for very large n, since $h_{19} = 0.002$, i.e. prediction of the 20th letter would be possible at almost certainty given the 19 preceding ones; this is in contradiction to experience.

The most striking feature of Fig. 6 is that the block entropies determined for the yeast chromosome sequence are almost maximal, which is indicative of a very disordered sequence. This is most surprising as one would ascribe to life a high complexity and a high degree of structure, and these criteria are certainly not fulfilled by a completely random string. Moreover, safe information transmission relies inevitably upon the redundancy of a message, which is obviously lacking here.*

An explanation of this phenomenon, which is, besides, consistent with the so-called Random Origin Hypothesis† shall be given in the following.

Swanson (1989) discovered that the genetic code, —the dictionary, as it were, describing the translation from base triplets (codons) to amino acids—is approximately a Gray code (Hamming, 1980). A Gray code can be compared with a continuous function in mathematics: small changes in the domain of a continuous function entail small changes in its range. This means for the genetic code that small changes in the codon—for example, the exchange of one purine (A and G) against the other purine—result in an amino acid which is similar to the original one. Such deliberations demand the definition of a property space for amino acids which is endowed with a metric.

Mistakes during the transcription procedure therefore do not result in a faulty protein most of the time. Thus, the redundancy is shifted to the translation mechanism; as is well known, 20 amino acids can be represented by 61 codons (three stop codons do not code for amino acids). Figure 6 also allows an estimation of about 0.95 ($H_{15}/15$) for the entropy of the source of the examined DNA sequence. Such a value would result in a density of 10^{-9} in the

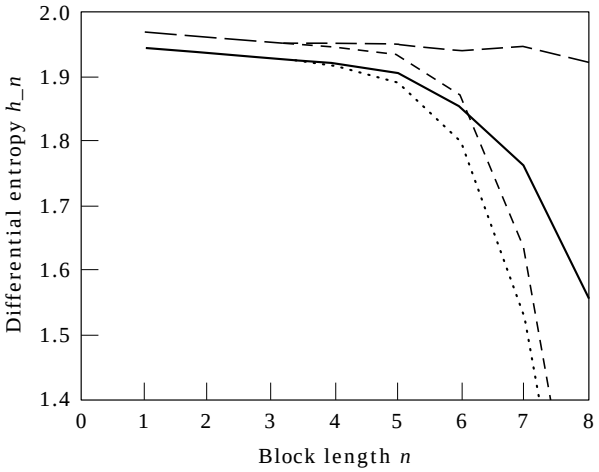

FIG. 7. The representation as differential entropies (here: those of the Epstein-Barr virus genome) shows clearly the influence that repetitive sequences make on block entropies. When all of the documented repeats were removed, the entropy values rose by about 0.2 bits. The artificial lowering of the differential entropies can be located at $n = 4$ in the sequences both with and without repeats. Key: . . . complete sequence, no correction; — complete sequence, corrected; – – – repeats removed, no correction; – – repeats removed, corrected.

* One could compare the sequence of yeast DNA with the sequence of the digits of π or e: both are certainly meaningful numbers, nevertheless the sequence of their digits is perfectly random.

† According to this hypothesis polypeptide sequences were created as random sequences at an early stage of evolution. Only those sequences whose three-dimensional conformation had an enzymatic or structural function survived in the course of evolution. For a review discussing this subject, see White (1994).

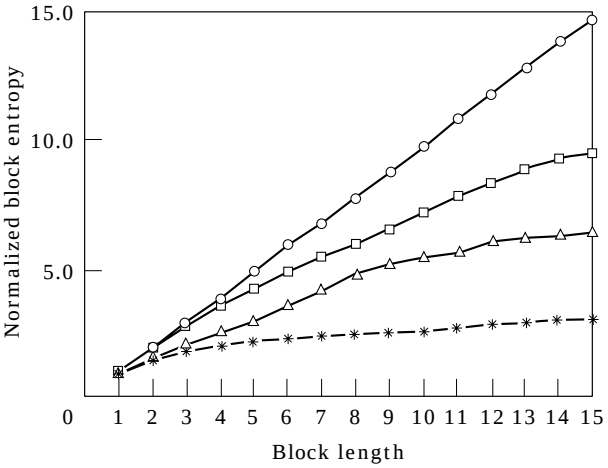

FIG. 8. In order to compare sequences coded over alphabets of different sizes λ the logarithm in the entropy calculation has to be taken to base λ. While the yeast DNA sequence is very close to a random sequence, the restrictive rules of human languages and computer languages are reflected in low entropies (high redundancy). Music tends to be more irregular than language. The alphabet size is given in brackets. Key: ○, yeast chromosome III (4); □, Beeethoven Sonata no. 32 (3); △, Alice in Wonderland (27); *, FORTRAN source code (90).

space of 300 bp long DNA sequences ($4^{300 \cdot 0.95}/4^{300}$), i.e. the space of DNA sequences would be populated much more densely than the space of the corresponding polypeptides. The degeneracy of the genetic code, thus, seems to sort out many of the possible combinations in DNA space.

The effect that repetitive sequences have on block entropies is demonstrated in Fig. 7. Upon removal of all repetitive sequences of the Epstein-Barr virus genome described in the data bank annotation,—they constitute about 25% of the whole genome—the block entropy H_8 rose by about 0.4 bits. The presentation of differential entropies [eqn (3)] instead of block entropies allows one to assess the range to which the absolute values of the corrected block entropies can be considered as reliable. Theoretically, the differential entropies decrease monotonously with n. The slight increase between $n = 6$ and $n = 7$ indicates that even corrected values should be taken as *estimations* rather than *calculations* of block entropies from this block length on.

Figure 7 illustrates clearly the relevance of the MEP method. The uncorrected curves resemble each other since their decay is dominated by the finite sample effect. The corrected values display, however, the essential difference: the full sequence exhibits redundancy due to repeats whereas the sequence without repeats is close to a random one.

Figure 8 demonstrates that the block entropies for various processes can scale in quite different ways. As seen above, block entropies of the yeast sequence are close to those of a random process. On the other pole of the spectrum lies the computer source code. Very restrictive rules limit the number of syntactically correct combinations, which is reflected in low entropies.

Language and music* lie between these poles.

This order—DNA, music, human language, computer language—when ordered by decreasing entropy, is confirmed by the calculation of the Lempel–Ziv complexity (Lempel & Ziv, 1976) which also serves as an estimation of the entropy of the source, i.e. the difference between two consecutive block-entropies for block-lengths n tending towards infinity [see eqn (4)].

Summary

We presented a new method to estimate block entropies from small samples of symbol sequences. Making use of two fundamental principles,—the McMillan theorem of asymptotic equi-distribution and the Maximum Entropy Principle—the most probable underlying probability is determined from a series of parameterized test-distributions. Tests carried out with strings containing repeated sections suggest that the range of applicability for this method exceeds that of previous methods by far. Fortunately, real sequences of interest like DNA sequences or literary texts dispose of a sub-word rank-statistics similar to that of our chosen test sequences. Although secure absolute values for the entropies of long blocks cannot be achieved it allows interesting statements about the inherent order of symbol sequences and the processes generating them. Surprisingly, DNA sequences behave closer to completely random sequences than to written text. The very strict syntax of computer languages on the other hand is reflected by a very low average information content of its sub-strings.

We are grateful to W. Ebeling for stimulating discussions and to R. Herwig for commenting on the manuscript. We acknowledge financial support from the Stiftung Volkswagenwerk and the Deutsche Forschungsgemeinschaft.

* The piece of music was encoded by Ebeling & Nicolis (1992) using a dynamic partitioning: the symbols were attributed to the change in pitch (lower or higher than the previous note or constant).

REFERENCES

BAER, R., BANKIER, A. T., BIGGIN, M. D., DEININGER, P. L., FARRELL, P. J., GIBSON, T. J. *et al*. (1984). DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature* **310,** 207–211.

EBELING, W. & NICOLIS, G. (1992). Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos*, *Solitons & Fractals* **2,** 635–650.

EBELING, W. (1993). Entropy, predictability and historicity of nonlinear processes. In: *Statistical Physics and Thermodynamics of Nonlinear Nonequilibrium Systems* (Ebeling, W. & Muschik W., eds). Singapore: World Scientific.

GRASSBERGER, P. (1988). Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A* **128,** 369–373.

HAMMING, R. W. (1980). *Coding and Information Theory*. Englewood Cliffs, NJ: Prentice-Hall.

HERZEL, H. (1988). Complexity of symbol sequences. *Sys. Anal. Mod. Sim.* **5,** 435–444.

HERZEL, H., SCHMITT, A. O. & EBELING, W. (1994a). Finite sample effects in sequence analysis. *Chaos*, *Solitons & Fractals* **4,** 97–113.

HERZEL, H., EBELING, W. & SCHMITT, A. O. (1994b). Entropies of biosequences—the role of repeats. *Phys. Rev. E* **50,** 5061–5071.

JAYNES, E. T. (1983). Where do we stand on maximum entropy? In: *The Maximum Entropy Formalism* (Levine, R. & Tribus, M., eds). Cambridge, MA: MIT Press.

JUSTICE, J. H. (1986). *Maximum Entropy and Bayesian Methods in Applied Statistics*. Cambridge: Cambridge University Press.

KAPUR, J. N. (1989). *Maximum-Entropy Models in Science and Engineering*. New York: John Wiley & Sons.

LEMPEL, A. & ZIV, J. (1976). On the complexity of finite sequences. *IEEE Trans. Inf. Theory* **IT-22,** 75–81.

LEWIN, B. (1994). *Genes V*. New York: Oxford University Press.

MCMILLAN, B. (1953). The basic theorems of information theory. *Ann. Math. Statist.* **24,** 196–210.

OLIVER, S. G., VAN DER AART, Q. J., AGOSTONI-CARBONE, M. L., AIGLE, M., ALBERGHINA, L., ALEXANDRAKI, D., *et al*. (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357,** 38–46.

PÖSCHEL, T., EBELING, W. & ROSE, H. (1995). Guessing probability distributions from small samples. *J. Stat. Phys.* **80,** 1443–1452.

SCHMITT, A. O., HERZEL, H. & EBELING, W. (1993). A new method to calculate higher-order entropies from finite samples. *Europhys. Lett.* **23,** 303–309.

SCHMITT, A. O. (1995). Structural analysis of DNA sequences. Ph.D. Thesis, Berlin: Verlag Dr. Köster.

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

SHANNON, C. E. (1951). Prediction and entropy of printed English. *Bell Syst Tech. J.* **30,** 50–64.

STRAIT, J. B. & DEWEY, T. G. (1996). The Shannon information entropy of protein sequences. *Biophys. J.* **71**, 148–155.

SWANSON, R. (1989). A unifying concept for the amino acid code. *Bull. Math. Biol.* **51,** 417–432.

VÖLZ, H. (1990). *Computer und Kunst*. Leipzig: Urania-Verlag.

WHITE, S. H. (1994). Global statistics of protein sequences: implications for the origin, evolution, and prediction of structure. *Annu. Rev. Biophys. Biomol. Struct.* **23,** 407–439.

WICKMAN, D. (1990). *Bayes-Statistik*. Mannheim: BI Wissenschaftsverlag.