Open Sys. & Information Dyn. **5**: 265–278, 1998

Printed in the Netherlands

Maximum Entropy Method in Analysis of Genetic Text and Measurement of its Information Content

N. N. Bugaenko[1], A. N. Gorban[1] and M. G. Sadovsky[1,2]

[1] *Computing Center, Siberian Division of Russian Academy of Sciences, Akademgorodok, Krasnoyarsk, 660036*

[2] *Institute of Biophysics, Siberian Division of Russian Academy of Sciences Akademgorodok, Krasnoyarsk, 660036*

(Received: September 10, 1997)

Abstract. The information capacity in frequency dictionaries of nucleotide sequences is estimated through the efficiency of reconstruction of a longer frequency dictionary from a short one. This reconstruction is performed by the maximum entropy method. Real nucleotide sequences are compared to random ones (with the same composition of nucleotides). Phages genes from NCBI bank were analyzed. The reliable difference of real genetic texts from random sequences is observed for the dictionary length $q = 2$, 5 and 6.

1. Introduction

Storage and processing of genetic information are the central problems of molecular genetics and molecular biology. Nucleotide sequences play the key role in these processes. Recent intense increase in the original genetic data for various genes challenges theoretical biologists: what terms and methods should be used to understand this vast experimental material [1 – 8]. Application of mathematical methods for studying nucleotide sequences is a long-standing story (for review see [1 – 12]).

At present, the variety of methods developed can be divided into two classes: methods involving context [3, 4, 7, 10, 11] and context-free [1, 2, 6, 12] analysis of symbol (nucleotide, in particular) sequences. The context-involving methods assume special biological knowledge and aim at the analysis of groups of nucleotide sequences to obtain statistical characteristics as well as consideration of single sequences to recognize functional sites, introns, exons etc. Being more abstract, the context-free methods are assumed to avoid any involvement of knowledge of this sort; mainly, they take origin from the methodology of statistical physics. The study of statistical properties of nucleotide sequences means a transition from the consideration of a specific nucleotide sequence to the consideration of ensembles of their fragments [7, 8, 10, 12]. Here we tried, within the framework of the second approach, to develop a method for the evaluation of information capacity of the dictionaries of nucleotide sequences. Basic issue of our methodology is the recon-

struction of a set of longer fragments of a sequence from a given set of shorter ones. This issue was used quite a time ago in linguistics by famous soviet mathematician A.N. Kolmogorov [13], who put a problem of prediction of the next symbol in a (linguistic) text. Similar problem arises in statistical physics [14, 15], where the behaviour of multi-particle system should be described by a distribution function of a small number of particles (some promising results in that field are presented in [16]). Statistical physicists have elaborated a number of approximate methods, some of which are recognized as very useful and powerful.

We consider the problem of the reconstruction of the set of longer fragments from the set of the fragments of the given length.

2. Reading Window and Frequency Dictionary

Here we present the methodology of investigating the DNA primary structure *as a text*. DNA or RNA nucleotide sequence is considered as a linear connected sequence of symbols and is called *genetic text* (GT); the number N of symbols in the text is called *the length of* GT. Recognition of structural units in GT is a natural way to introduce a logical order over the increasing amount of GTs. Some of these units are well known: codon, exon, intron, TATA box, signal sequences. One might expect new structure units to be discovered.

The principles of recognizing these structural units often depend on the subject of investigation. Probably, the most general principle is that *these units differ in function and/or history*. Consistent realization of this principle encounters significant difficulties. Beside purely technical problems, there are substantial ones: the meaning of specific GT regions which are presumed to be structural units depends on their relative disposition, to a great extent.

It should be stressed that there is no intracellular process which deals with the abundant genetic information stored in an entire nucleotide sequence. To the contrary, all known intracellular processes related to realization of genetic information involve different fragments of DNA (or RNA) not greatly varying in length. The information is read locally, by small portions and from small DNA regions, for most (if not all) cases. During the information processing, the reading "device" runs along the nucleotide sequence in small steps. Let us call such a device *the reading window*. The essential formalized features of the reading window are the size of the region being read and the space between the nearest regions. Let consider here the simplest case, when 1) the read fragment is of a permanent length, 2) the reading window moves permanently in the same direction, and 3) the step of the reading window motion is permanent and equals to one nucleotide. A site of DNA (or RNA) read by the reading window of the length q would be called *a word of the length* q.

The concept of reading window focuses the study of nucleotide sequences (texts) on purely mathematical objects, namely Frequency Dictionaries (FD). Let us consider GT of the length N and a reading window of the length q located at an end of GT (conventional "start"). Let us move the reading window consecutively towards the opposite end, so that $N - q + 1$ words of the length q will be read. Identical words can occur among these. The complete set of words encountered in GT, accompanied by their frequencies is called the *Frequency Dictionary* (FD) [17 – 20]. Further, we consider the dictionaries of various lengths, from one to some specific length.

Obviously, a longer dictionary bears entire information about all shorter dictionaries of the same GT. One can easily obtain a dictionary of shorter length from a dictionary of a given length. One can unambiguously derive entire symbol sequence from the dictionary of some specific length $d^* + 1$, where d^* is defined as the minimal length for which all the words in the dictionary occur uniquely. This specific length can be easily calculated; d^* is a rather informative characteristics of real genes, it represents the redundancy of symbol sequences [17, 19]. Since the value of d^* depends strongly on both the structure of a symbol sequence and its length, one is to compare the values of d^* observed in real genes with the values of d^* obtained for random sequences rather than between themselves. It results from the following estimation of d^* for random sequences: $d^* \sim \log_2 N$, [17]. Thus, the value $d^*/\log_2 N$ should be compared for the real genes. It was found that human genes differ form the genes of human viruses with respect to this value; similarly, this value is less for exons than for introns, for genes of eukaryotic organisms, and decreases after the splicing of RNA [18].

We consider here the dictionaries of length shorter than d^*. This paper aims to introduce a strict definition of information capacity of FDs of various lengths.

3. Dictionary Reconstruction. Maximum Entropy Method

Shorter dictionary can always be obtained from a given one by summation of frequencies of the words; the inverse transitions, in general, impossible: a part of information is lost due to summation. Thus, the transition from a given dictionary to a longer one is ambiguous. This ambiguity can be decreased with the help of additional knowledge, e.g. from a consideration of biological functions of words, their interlocation in an original gene, and so on. The problem of longer dictionary derivation from a given one is called the *dictionary reconstruction*.

We solve this problem by the maximum entropy method, which implies using the information contained in the given dictionary only, and avoiding an involvement of any external knowledge, both explicit and implicit.

Let us reconstruct a longer dictionary from a given one so that the reconstructed dictionary shows maximal indeterminacy. It means that one must choose the dic-

tionary with maximal entropy among all longer dictionaries which can be derived from a given one. Entropy of a dictionary is defined in traditional way,

$$S_q = -\sum_{i_1 \dots i_q} f_{i_1 \dots i_q} \ln f_{i_1 \dots i_q}, \tag{1}$$

where i_j is a letter (symbol) from the text alphabet; $i_1 \dots i_q$ represents a word of length q. i_j runs over the letters {A, C, G, T} in our case. $f_{i_1 \dots i_q}$ is the word frequency. The summation is performed here over all words encountered in the text. In order to eliminate boundary effects, the text is closed into a circle. Maximal possible entropy is $\max\{S_q\} = -4^q(\frac{1}{4^q})\ln(\frac{1}{4^q}) = q\ln(4)$, for the dictionary of the length q.

Let us consider the reconstruction of the dictionary one symbol longer than the original one. This extremum problem is stated in the following way:

$$S_{q+1}[f_{q+1}] \to \max \tag{2}$$

with the constraints

$$\sum_{i_{q+1}} f_{i_1 \dots i_q i_{q+1}} = f_{i_1 \dots i_q}, \tag{2a}$$

$$\sum_{i_{q+1}} f_{i_{q+1} i_1 \dots i_q} = f_{i_1 \dots i_q}, \tag{2b}$$

following from the interrelation between these two dictionaries. Here S_{q+1} is the entropy of the reconstructed dictionary of length $(q+1)$ derived from the original dictionary. The summation in (2a) and (2b) is performed over all possible i_{q+1} (recall that i_q corresponds to {A, C, G, T}). The conditions (2a) and (2b) mean that the reconstructed dictionary must not be an arbitrary one, but it must yield the original shorter dictionary under summation.

Solution by the indeterminate Lagrange multiplier method gives

$$f_{i_1 \dots i_q i_{q+1}} = \exp\Big\{ \sum_{i_1 \dots i_q} \tilde{\alpha}_{i_1 \dots i_q} + \sum_{i_2 \dots i_{q+1}} \tilde{\beta}_{i_2 \dots i_{q+1}} - 1 \Big\}, \tag{3}$$

where $\tilde{\alpha}_{i_1 \dots i_q}$ and $\tilde{\beta}_{i_2 \dots i_{q+1}}$ are the indeterminate multipliers corresponding to linear restrictions (2a) and (2b). Denoting

$$\alpha_{i_2 \dots i_{q+1}} = \exp\Big(\sum_{i_2 \dots i_{q+1}} \tilde{\alpha}_{i_2 \dots i_{q+1}} - \frac{1}{2} \Big),$$

$$\beta_{i_1 \dots i_q} = \exp\Big(\sum_{i_1 \dots i_q} \tilde{\beta}_{i_1 \dots i_q} - \frac{1}{2} \Big), \tag{4}$$

and using the constraints (2a) and (2b), we obtain the solution of the problem (2) as (here $q > 1$)

$$f_{i_1 \dots i_q i_{q+1}} = \frac{f_{i_1 \dots i_q} f_{i_2 \dots i_{q+1}}}{f_{i_2 \dots i_q}} \tag{5}$$

where $f_{i_2 \dots i_q}$ are the frequencies of a dictionary one symbol shorter than the original one, yielded by its summation. For $q = 1$ the solution is obvious: $f_{i_1 i_2} = f_{i_1} f_{i_2}$.

A dictionary of length $(q + s)$ is reconstructed from the original dictionary of length q in a similar way:

$$S_{q+s}[f_{q+s}] \to \max \tag{6}$$

with the constraints

$$\begin{aligned}
\sum_{i_{q+1}, \dots, i_{q+s}} f_{i_1 \dots i_q i_{q+1} \dots i_{q+s}} &= f_{i_1 \dots i_q} \\
\sum_{i_{q+1}, \dots, i_{q+s}} f_{i_{q+1} i_1 \dots i_q i_{q+2} \dots i_{q+s}} &= f_{i_1 \dots i_q} \\
&\vdots \\
\sum_{i_{q+1}, \dots, i_{q+s}} f_{i_{q+1} \dots i_{q+s} i_1 \dots i_q} &= f_{i_1 \dots i_q}
\end{aligned} \tag{6a}$$

The solution is:

– for $q > 1$

$$f_{i_1 \dots i_a i_{a+1} \dots i_{a+s}} = \frac{f_{i_1 \dots i_q} f_{i_2 \dots i_{q+1}} \cdots f_{i_{q-s+1} \dots i_{q+s}}}{f_{i_2 \dots i_q} f_{i_3 \dots i_{q+1}} \cdots f_{i_{q-s+1} \dots i_{q+s-1}}}, \tag{7}$$

– and for $q = 1$

$$f_{i_1 \dots i_{q+s}} = f_{i_1} \cdots f_{i_{q+s}}. \tag{8}$$

The expressions (7) for the reconstructed frequencies are analogous to Kirkwood's approximation [14, 16], but unlike the latter, are exact solutions. The approach developed here to study the statistical properties of nucleotide sequence comes from statistical physics [15], where it is implemented to study multi-particle distribution functions. The problem of derivation of three-particle distribution functions from two-particle ones has been solved in [16]; we will not discuss the relation between Kirkwood's approximation in statistical physics and the solution of dictionary reconstruction any more here. The above approach to studying nucleotide sequence is a new step towards the idea of Schrödinger to consider life as an aperiodic crystal (ordered structure).

4. Entropy, Limit Entropy, Informativity and Dictionary Reconstruction Quality

The entropy S_q of a dictionary represents the indeterminacy of an occurrence a word of length q at an arbitrary location in a nucleotide sequence. If $q < d^*$, then the reconstructed dictionary is less determined than the real dictionary of this length. Let $S_i(j)$ be the entropy of a dictionary of the length i reconstructed from a given dictionary of the length j $(i > j)$. Formulae (1), (7) and (8) yield

$$S_i(j) = (i-j+1)S_j - (i-j)S_{j-1}, \qquad j > 1, \tag{9}$$

$$S_i(1) = iS_1, \qquad j = 1. \tag{10}$$

The entropies S_i steadily increase with the dictionary length i (obviously, up to d^*, when the entropy becomes constant), while the entropies $S_i(j)$ steadily decrease as j grows from 1 to i.

It should be stressed that the largest jump of the value $S_i(j)$ is observed at the dictionary lengths $j = 6$, 7, and 8. We have examined all the nucleotide sequences in NCBI-bank, and maximal $S_i(j)$ jump was always observed at these lengths. To make this effect clearer, let us introduce the notion of informativity. The maximal possible entropy for a dictionary of length i is

$$\max\{S_i\} = -4^i\left(\frac{1}{4^i}\right)\ln\left(\frac{1}{4^i}\right) = i\ln(4);$$

4^i is the maximal possible number of words in a dictionary of length i. Then the informativity is characterized as

$$I_i = \max\{S_i\} - S_i = i\ln(4) - S_i. \tag{11}$$

Zero informativity corresponds to a dictionary with complete set of words of equal frequencies. As opposed to the entropy, this characteristic is a measure of deviation from disorder. Zero informativity corresponds to complete indeterminacy, "chaos", rather than to the absence of information. The value I_i can be considered as the "distance from chaos".

Using I_i, one can compare dictionaries of equal lengths of different sequences with respect to their indeterminacy. In order to compare dictionaries of different lengths, we introduce the notion of limit specific entropy. Reconstruct a dictionary of length n from a given one of smaller length j and consider the limit $S_n(j)/n$ as $n \to \infty$. It corresponds to the well-known thermodynamic limit in statistical physics.

$$\lim_{n\to\infty}\frac{S_n(j)}{n} = \lim_{n\to\infty}\frac{(n-j+1)S_j - (n-j)S_{j-1}}{n} = S_j - S_{j-1}. \tag{12}$$

For the dictionary of length $j = 1$ we have

$$S_n(1) = nS_1, \qquad \lim_{n\to\infty} \frac{S_n(1)}{n} = S_1. \tag{13}$$

This is the specific entropy (entropy per symbol) in a dictionary of infinitely long words reconstructed from a given dictionary.

Further on we omit the word "specific", since we do not mean the true limit (infinite) entropy. It is convenient to measure the limit entropy in portions of the unit dictionary with equal letter frequencies because, firstly, it is the greatest possible limit entropy value and, secondly, it eliminates the dependence on entropy units (bits, dits, etc.). Since $\max(S_1) = \ln 4$, let us define the limit entropy as

$$S_\infty(j) = \frac{S_j - S_{j-1}}{\ln 4}, \qquad j > 1, \tag{14}$$

$$S_\infty(1) = \frac{S_1}{\ln 4}, \qquad j = 1. \tag{15}$$

It varies from zero (complete determinacy) to one (complete indeterminacy).

Limit entropy is an essential characteristic of dictionaries. It shows that the information in a dictionary increases as the length of the latter grows. Besides, the entropy difference between two dictionaries of two consecutive lengths yields the information gain of the longer dictionary. In most cases, the smallest limit entropy difference is observed between the dictionaries of lengths 2 and 3, which shows little difference in their information capacity. This means that although the information in nucleotide sequences is encoded with triplets (codons), a significant part thereof is stored in doublets, which brings into connection the degeneracy of genetic code. This fact seems to be of general nature.

Similarly, we introduce limit specific informativity for a dictionary of length i (we omit the word "specific" below) by

$$I_\infty(i) = 1 - S_\infty(i), \tag{16}$$

and for a dictionary of length i reconstructed from a dictionary of length j by

$$I_i(j) = \max(S_i) - S_i(j) = i\ln(4) - S_i(j). \tag{17}$$

The comparison of dictionaries in terms of informativity differs from that in terms of entropy. Two dictionaries with the same entropy but of different lengths have different informativities (the informativity is higher for the longer dictionary). Equal entropies indicate similar indeterminacy of choosing a word of a given length in an arbitrary text site, while unequal informativities indicate that the longer dictionary is more exotic. Indeed, similar experimental outcomes for the longer dictionary should be less frequent, since the number of possible word sets is higher.

Hence, informativity pertains to the relationship between the whole dictionary and the set of possible dictionaries of the same length rather than to a particular experiment.

Let compare informativities of the real and reconstructed dictionaries. We call $Q_i(j)$ *the quality of reconstruction* of a dictionary of length i from dictionary of length $j < i$,

$$Q_i(j) = \frac{I_i(j)}{I_i} = \frac{\max(S_i) - S_i(j)}{\max(S_i) - S_i}. \tag{18}$$

Since $S_i(j) \geq S_i$, the quality of reconstruction varies within $0 \leq Q_i(j) \leq 1$, where the value 1 corresponds to the case $S_i(j) = S_i$. The procedure of dictionary reconstruction takes into account all possible extensions of a word, so that the resulting dictionary contains all the words from the real dictionary and, probably, some extra ones, while the frequencies may differ. That is why $Q_i(j) = 1$ for the case of exact reconstruction, while deviation from unity reveals the difference between the real and reconstructed dictionaries.

5. Comparison of Real vs. Random Texts

Random noncorrelated sequence is the first and most convenient model of a real genetic text. Let us call *a random text* corresponding to a given real text a text of the same length with the same proportions of nucleotide composition. This random text is constructed by the method of random choice of elements ("urn model").

We have considered a number of texts of nucleotide sequences from NCBI server. The dictionaries of lengths from 1 to 10 were constructed for each text and the limit entropies and the quality of reconstruction were calculated. 100 random texts have been implemented each real sequence, and the average values and standard deviations of those entities were calculated. In general, the patterns seem to be rather similar for sequences of different organisms. Typical values for the quality of reconstruction are shown in Fig. 1. The differences between limit entropies of the dictionaries of random and real sequences are shown in Fig. 2.

It is peculiar that the dictionary of length 2 is badly reconstructed from the unit dictionary for real nucleotide sequences, while $Q_3(2)$ is similar for real and random texts. It means that the dictionary of real texts of length 2 contains an essential part of information about the dictionary of length 3. The dictionaries of lengths 5 and 6 reconstructed from dictionaries shorter by one symbol are better for real sequences in comparison to the random ones. To demonstrate the generality of these effects and to observe entire picture, we have analyzed over 1000 phages genes. The results of calculation of the quality of reconstruction for this group of genes are shown in Fig. 3.

One can see that clearly distinguishable peculiarities are observed only for the texts long enough, as a rule of more than 500 nucleotides. The effect of finiteness

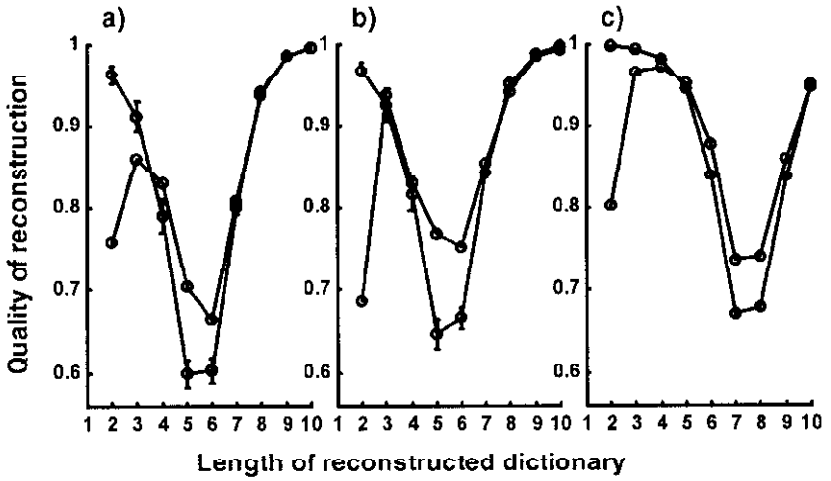

Fig. 1. Examples of the dependence of reconstruction quality for a dictionary of length $q+1$ derived from a dictionary of length q; the length of the reconstructed dictionary is represented on the horizontal axis. a) a sequence from chicken genome, $N = 2136$; b) a sequence from human genome, $N = 1639$; c) a sequence from nematoda genome, $N = 26139$. Dashed line connects the quality of reconstruction for real sequences, while solid line connects that for random noncorrelated sequences of the same nucleotide composition. Standard deviations are shown.

shows up for shorter texts. In Fig. 4, the differences of random and real limit entropies are shown only for the phages genes with $N \geq 500$.

6. Discussion

This paper presents the results of the exploration of the problem of determination of information content in nucleotide sequences. These results are of two kinds: firstly, methodological, and secondly, concerning the properties of some particular groups of sequences studied.

The results of methodological value are the following. An explicit formula for the reconstruction of longer dictionaries from the shorter ones is obtained; this formula has maximal generality and does not require (explicitly or implicitly) any special assumptions on the properties of original nucleotide sequences, or their models. Kolmogorov used these special formulae in his linguistic studies, without any detailed discussion of their validity. Our formula implies all known approximate methods of statistical investigation of nucleotide sequences based on their modelling by Markov chains of various order.

The information content per single symbol (letter, nucleotide) is estimated; these estimations are obtained for very long (in the limit, infinitely long) symbolic

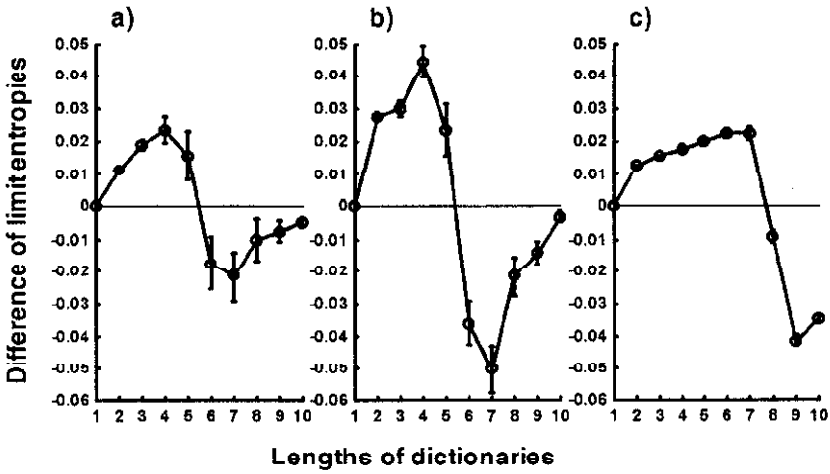

Fig. 2. The difference between limit specific entropies of a random noncorrelated sequence and a real genetic text. To visualise the pattern, the points are connected. The length of dictionary is plotted on the horizontal axis. a) a sequence from chicken genome, $N = 2136$; b) a sequence from human genome, $N = 1639$; c) a sequence from nematoda genome, $N = 26139$. Standard deviations are shown.

sequences which were reconstructed from a given dictionary. This entity is related the redundancy estimations introduced in Kolmogorov's works. The primitive variants of the redundancy estimations for modelling sequences dealing with one-letter dictionaries were introduced and discussed in detail in [18]. These estimations were based on the modelling of GT by random non-correlated chains.

The approach presented above is illustrated with calculations performed on a number various real genes. The results of analysis of all the phage sequences obtained from NCBI databank (release 94) are described below. The comparison of real and random genetic texts shows that:

- the dictionaries of length two of real and random sequences possess significantly different information capacity;
- the dictionary of length two of a real genetic text bears significantly more information about the whole text in comparison to the dictionaries of length two of random texts;
- the increase of information content in dictionaries of length three when compared to the dictionaries of length two is visibly smaller for real texts than for random ones;
- the dictionary of length eight bears over 90% of total information about genetic text.

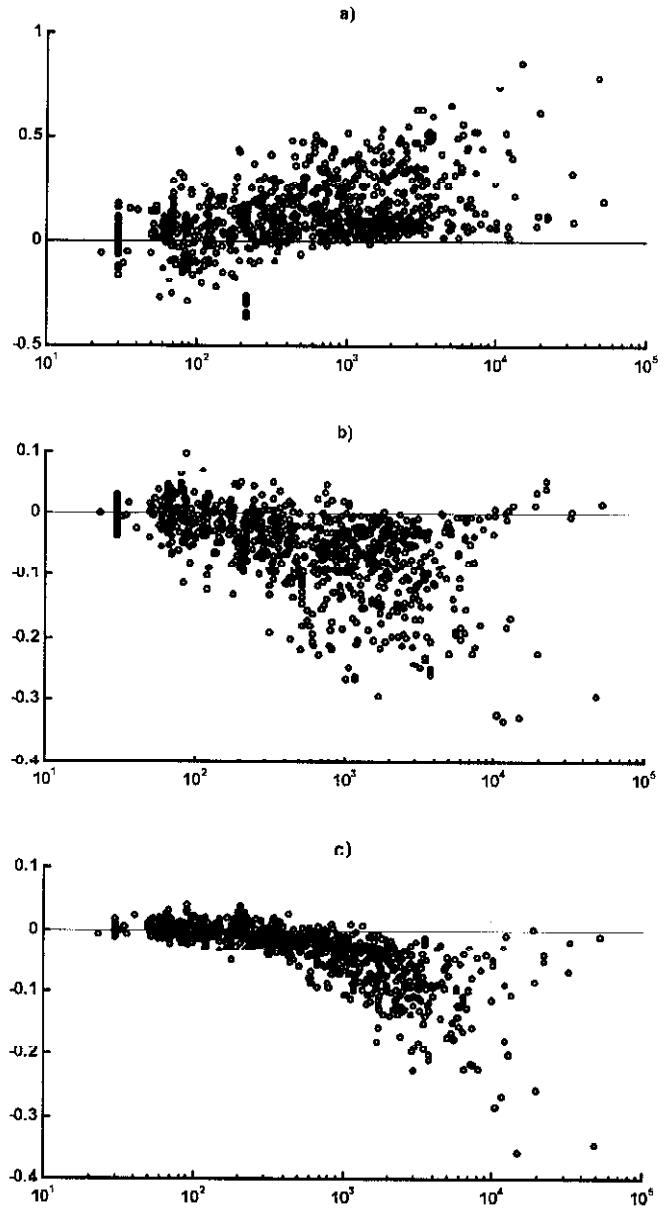

Fig. 3. The difference between reconstruction qualities of random sequences and real genetic texts of various lengths. Entire set of phage genes from NCBI server (release 94) is presented (over 1000 genetic texts). For convenience, standard deviations are not shown. a) Q_2 (1); b) Q_5 (4); c) Q_6 (5).

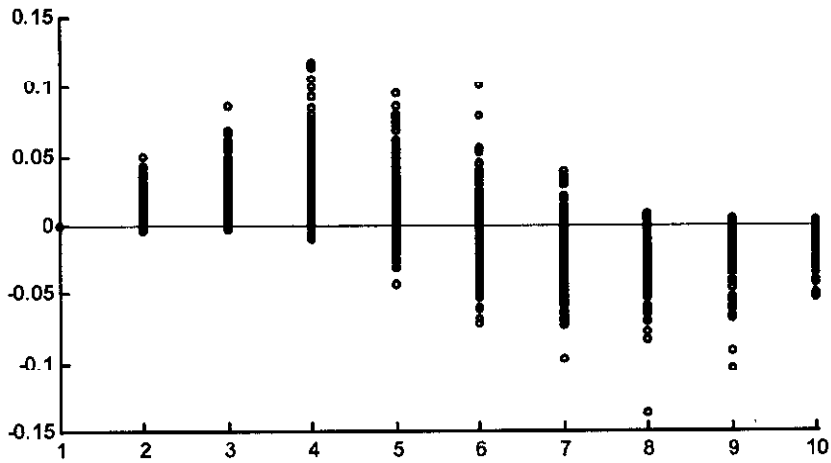

Fig. 4. The difference between limit specific entropies of random noncorrelated sequences and real genetic texts of phage genes longer than 500 nucleotides. The length of dictionary is plotted on the horizontal axis. Standard deviations are not shown.

7. Conclusions

The basic idea is to determine the quality of reconstruction of a longer FD from a given one. Generally, such reconstruction is ambiguous. Similar situations occur in statistical physics, and the general approach is to extend the consideration of a single object to the consideration of sets of such objects (ensembles). The main idea of the method proposed is as following: when unambiguous reconstruction of longer FD is impossible, one should reconstruct a set of all possible FDs (i.e., the ensemble) from the original one. The entropy of the ensemble reveals the reconstruction accuracy (and hence the information value).

The problem of GT reconstruction could be promoted by the reconstruction of FDs in general, it is always a matter of probability. One can estimate precisely the indeterminacy for such a reconstruction. This indeterminacy decreases as the length q of fragments (words) used for the reconstruction increases. The variation in indeterminacy of this reconstruction is an exact measure of information content of a given FD. Besides the probabilistic origin of the approach developed to define information capacity of genes, one should bear in mind that the information capacity measured through the quality of reconstruction can only be determined relatively. It means that one can calculate the capacity of two (or several) genes comparatively, while no absolute scale of information capacity can be introduced.

It should be stressed that the estimation of information capacity is neutral to the context of the sequence studied. This implies in the notion of information, and any deviation from it must be carefully justified.

In conclusion, we outline an important problem which may be treated by the above method. It is the problem of the determination of microstructure of a gene. (i.e. a distinguishing homogeneous regions different in their information characteristics). Exons and introns are, probably, the first candidates for such regions. Introns are recognized with the help of syntax and semantic features of GT. One may assume that coding and noncoding DNA regions can be reliably distinguished exclusively due to their statistical properties. Moreover, one can expect to discover new structures of genes determined by the statistical properties of their nucleotide sequences only; that is the microstructure of genes.

Acknowledgements

We thank Dr. Eugene M. Mirkes for valuable discussion and Dr. Tatyana G. Popova for useful remarks. This work was partly supported by the Krasnoyarsk Regional Science Foundation (grants 4F0153, 5F0108 and 7F0012) and Russian Foundation for Basic Research (grant 95-02-03836a).

Bibliography

1. H.P. Yockey, *Information Theory and Molecular Biology*, Cambridge Univ. Press, N.Y., 1992.
2. A.A. Alexandrov, et al.: *Computer Analysis of Genetic Texts* (in Russian), Nauka, Moscow, 1990.
3. S. Karlin and L.R. Cardon, Ann. Rev. Microbiol. **48**, 619 (1994).
4. A.K. Konopka, in: D. Smith, ed., *Biocomputings: Informatics and Genome Projects*, Acad. Press, San Diego, p. 119, 1995.
5. A.K. Konopka, in: R.A. Meyers, ed., *Molecular Biology and Biotechnology*, VCH Publishers, Weinheim, p. 888, 1995.
6. P.W. Garden, J. Theor. Biol. **82**, 679 (1980).
7. V. Brendel, J.S. Beckmann, and E.N. Trifonov, J. Biomol. Struct. Dyn. **4**, 11 (1986).
8. P.A. Pevzner, M. Yu. Borodovski, and A.A. Mironov, J. Biomol. Struct. Dyn. **6**, 1013 (1989).
9. M.A. Roytberg, in: S. Gindikin, ed., *Biosystems*, AMS, Providence, p. 103, 1992.
10. C. Martingale and A.K. Konopka, Computers and Chemistry **20**, 45 (1996).
11. A.K. Konopka and C. Martingale, Science **268**, 1789 (1992).
12. E.M. Mirkes, T.G. Popova, and M.G. Sadovsky, Adv. in Modelling and Analysis, ser. B **27**, 1 (1993).
13. A.N. Kolmogorov, Dokl. AN SSSR **65**, 793 (1949).
14. J. Kirkwood and E. Boggs, J. Chem. Phys. **10**, 394 (1942).
15. R. Balescu, *Equilibrium and Nonequilibrium Statistical Mechanics*, John Wiley & Sons, New York-London-Sidney-Toronto, 1975.
16. N.N. Bugaenko, A.N. Gorban, and I.V. Karlin, Teoret. i mat. fizika **88**, 430 (1991); (english translation: Theoretical and Mathematical Physics, Plenum Publ. Corp., p. 977, 1992).
17. T.G. Popova and M.G. Sadovsky, Advances in Modelling & Analysis, ser. A **22**, 13 (1994).
18. E.M. Mirkes, T.G. Popova, and M.G. Sadovsky, Advances in Modelling & Analysis, ser. B **27**, 11 (1993).

19. T. G. Popova and M. G. Sadovsky, Modelling, Measurement & Control, ser. C **45**, 27 (1994).
20. N. N. Bugaenko, A. N. Gorban, and M. G. Sadovsky, Molecular Biology **30**, 313 (1996).