Journal of Biological Physics **29:** 23–38, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Comparison of Real Frequencies of Strings vs. the Expected Ones Reveals the Information Capacity of Macromoleculae

MICHAEL G. SADOVSKY
Institute of Biophysics of Siberian Division of Russian Academy of Sciences, Akademgorodok, Krasnoyarsk, 660036, e-mail: uvenal@ktk.ru

Abstract. The information capacity of nucleotide sequences is defined through the calculation of specific entropy of their frequency dictionary. The specific entropy of the frequency dictionary is calculated against the reconstructed dictionary; this latter bears the most probable continuations of the shorter strings. This developed measure allows to distinguish the sequences both from the randons ones, and from those with high level of (rather simple) order. Some implications of the developed methodology in the fields of genetics, bioinformatics, and molecular biology are discussed.

Key words: dictionary, entropy, information capacity, Markov model, ordered sequence, random sequence, reconstructed dictionary, specific entropy

1. Introduction

Biological macromoleculae play a key role in the processes of storage and transfer of inherited information. A study of statistical properties of nucleotide or amino acid sequences brings further attention to the details of biological processes involving the macromoleculae, and clarifies some of the non-evident and peculiar features of them. A study of statistical properties of nucleotide sequences has an extensive background [1–5]. A student may find two basic methodologies of such study: the former is based on the probabilistic approach and involves mainly the results and methods developed in the information theory [6–8]; the latter is based on the approach specific for the studies of Boltzmann equation [9, 10]. Apparently, the first method is more widely known than the second one. The popularity of this first approach stems from the high coherence between the objects of information theory and molecular biology. These two fields of science both operate with symbol sequences. Some interesting and significant results have been obtained in that area (see [1–8, 11–15], for example).

Statistical physics also contributes greatly the investigations of nucleotide sequences. The key idea here is to study the ensemble of short fragments of a biological macromolecule as a multi-particle distribution function. Symbol sequences are more simple and less dimensional objects, thus providing a student with more specific and detail results, as opposed to the real multi-particle systems. Thus, the famous Kirkwood approximation [16] becomes an exact solution of extremal problem, for symbol sequence, while it remains the approximation, for continuous systems. Further success of the investigation of statistical properties of nucleotide sequences follows from the implementation of the method of invarionat manifolds [9, 10, 17] developed for a study of Boltzmann equation [18] and some other problems in non-equilibrium statistical physics.

Here we concentrate on the study of nucleotide sequences, and we are investigating their statistical properties. We focus on the study of various frequency distributions of sets of strings observed within a nucleotide sequence. The study allows to answer the question about the information capacity of nucleotide sequence. In general, information capacity is a measure of the unexpectedness of an occurrence of a string of some peculiar length, provided that the frequences of shorter strings making up that string are known. Information capacity differs from the information content [19–25] and the entropy of a frequency dictionary [19, 20, 23–28]. Entropy of a frequency dictionary measures an indeterminacy of an occurrence of a string observed through a random choice, while the information content measures a difference between the indeterminacy observed for the real frequency dictionary and that one developed through the principle of the most probable continuation of a string.

A study of information capacity falls within the general problem of a transforming the entity into fragments and the inverse transformation of fragments into the entity. Nucleotide sequences enable quite low arbitrariness in what is a fragment; an isolation of a string is the only way to develop a fragment within a sequence. Contrary to symbol sequences, plane objects (such as lattices) yield significantly greater variety of methods to determine a fragment (see [29–31]). Let us now look into the exact definitions and precise statements.

2. Frequency Dictionaries

Thus, a set of strings occurring within a nucleotide sequence is considered to be all ensemble W of fragments. The length q of strings composing the ensemble, and the step of a reading window which isolates a string are the main parameters characterising the ensemble [19, 20, 32–37]. Here we consider the ensembles of strings of variable length, but the step of a reading window shift is always equal to 1. In other words, each nucleotide makes the start of a string.

We shall consider the continuous nucleotide sequences only. N denotes the length (that is the number of nucleotides) of the sequence. A continuous string (a continuous subsequence) of the length q, $1 \leq q \leq N$ is called a word. A consideration of unbound sequences (i.e. the sequences bearing the gaps inside) brings some substantial technical difficulties with no further comprehension of biological issues implied by this occurrence. A set of all the words (of the length q) occurred within the studied nucleotide sequence is called support (or q-support, when the indication of length is necessary). A q-support accomplished with the number n_ω of copies of each element (each word) makes the finite dictionary (of the thickness q). The index ω enlists the words in the support. A q-support accomplished with the frequency f_ω of each word makes the frequency dictionary W_q (of the thickness q):

$$f_\omega = \frac{n_\omega}{N}$$

Here the sequence must be connected into a ring. The motivation for such definition can be seen in [19, 20]. A sequence is transformed unambiguously into the dictionary.

The problem of transformation of an entity into fragments and back could be stated as the problem of a transformation of ensembles $W_1, W_2, \ldots, W_l$ into each other, both upward and downward:

$$W_1 \leftrightarrow W_2 \leftrightarrow W_3 \leftrightarrow \ldots \leftrightarrow W_l,$$

where the lower index indicates the length of strings occurred within the ensemble.

A transformation of a frequency dictionary of thickness q into the dictionary of thickness $q - 1$ is obvious and simple. To get the dictionary W_{q-1}, one must sum up the frequency of all the words differing in the first (or in the last) symbol. Further, the invariance of the dictionary W_{q-1} against the variation of the position of a summation symbol is required; that is why a sequence must be connected into a ring. It is evident, that the transformation of an ensemble of words of the given length q into the ensemble of longer words is the basic point here. While the downward transformation makes no problem (see [19, 20] for details), the upward transformation is, as a rule, ambiguous. A downward transformation here means a transformation of the given ensemble into one made of shorter fragments, and the upward transformation is a transformation of the given ensemble into that one of longer fragments. In general, an inverse transformation, from W_q to W_{q+1} is ambiguous [19, 20]. Note, that there exists the specific thickness d^* of the frequency dictionary, that yields an unambiguous transformation of the frequency dictionary W_{d^*} into the frequency dictionary W_q, as $q > d^*$, for any finite sequence. Some biological features of that specific thickness are discussed in [32–37]. The ratio

$$r = \frac{d^*}{\ln N}$$

could be considered as a redundancy measure. It was found that the value r differs for introns and exons, and the excision of these former results in a decrease of r. The patterns of the variation of r for eukaryotic organisms and viruses are different. Meanwhile, we focus on considering an ambiguous derivation of thicker dictionary from the given one.

3. Specific Entropy and Information Capacity

An information capacity is defined as a divergence between the real and expected frequency of a string. Hence, the definition strongly depends on the idea of the expected frequency. As soon, as a word ω, $\omega \in W_q$ has several continuations, then a derivation of dictionary W_{q+1} from W_q yields a family of dictionaries $\{W_{q+1}\}$, instead of a single one. Each frequency dictionary from that family generates the given frequency dictionary of the thickness q. Thus, a question arises what dictionary from the family should be considered as the reconstructed one. There may be several ways to solve that. The matter is to choose the dictionary bearing no external information and/or knowledge concerning the sequence under consideration. Such dictionary $\widetilde{W}_{q+1}$ exists among the dictionaries in the family $\{W_{q+1}\}$; it contains the words of the length $q + 1$ which are the most probable progression of the words of the length q from the given dictionary. This dictionary $\widetilde{W}_{q+1}$, satisfies the following extremal principle:

$$S\left[\tilde{W}_{q+1}\right] \to \max,$$

where S is the entropy of the dictionary [17, 18]. This extremal principle allows to determine the frequencies $\tilde{f}_{i_1 i_2 i_3 \dots i_q i_{q+1}}$ of the words ω in the dictionary $\widetilde{W}_{q+1}$, $\omega \in \widetilde{W}_{q+1}$ in explicit form:

$$\tilde{f}_{i_1 i_2 i_3 \dots i_q i_{q+1}} = \frac{f_{i_1 i_2 i_3 \dots i_{q-1} i_q} \cdot f_{i_2 i_3 i_4 \dots i_q i_{q+1}}}{f_{i_2 i_3 i_4 \dots i_{q-1} i_q}}. \tag{1}$$

Here $f_{i_1 i_2 i_3 \dots i_q}$ is the frequency of the word $i_1 i_2 i_3 \dots i_q$, i_j is the nucleotide occupying the j-th position. The deduction of (1), as well as the generalisation of (1) for the case of reconstructing frequency dictionary $\widetilde{W}_{q+w}$, $S > 1$ of the thickness $q + s$ from the dictionary of thickness q is provided in Appendix I (see also [19, 20]).

An information capacity is determined through a comparison of real frequency dictionary vs. the reconstructed one; another approach focusing on a measure of so called 'quality of reconstruction' is presented in [19, 20]. Here we develop another approach to determine the information capacity. Indeed, if an equilibrium distribution function ϕ^* exists, then one always can calculate the specific entropy of a distribution function ψ with respect to the equilibrium one. The specific entropy is defined as

$$\bar{S}(\psi \mid \phi^*) = \sum_{\omega} \psi_\omega \ln \frac{\psi_\omega}{\phi^*_\omega}. \tag{2}$$

Considering ψ and ϕ as frequency dictionaries, one can calculate the indeterminacy of an occurrence of a word ω in the distribution ψ, if the frequency of that latter in the distribution ϕ^* is known; bearing in mind (2), one hardly could get the indeterminacy of a frequency dictionary with respect to another one immediately. The point is that to make the expression (2) valid, the dictionary corresponding to the distribution ϕ^* must contain all the words which are found in the dictionary corresponding to distribution ψ. Such inclusion is not guaranteed beforehand, for real frequency dictionaries. Luckily, the principle of the most probable continuation of a word yielding the frequencies of (1) type guarantees all the words observed in the real dictionary (and, maybe, some others), exist in the reconstructed dictionary. Thus, the specific entropy

$$\bar{S} = \sum_{\omega} f_{\omega} \cdot \ln \frac{f_{\omega}}{\tilde{f}_{\omega}} \tag{3}$$

is always defined. Here f_{ω} is the frequency of the word ω (at the real frequency dictionary W_q), and $\tilde{f}_{\omega}$ is the frequency of the word ω at the reconstructed frequency dictionary $\widetilde{W}_g$. Everywhere below we shall consider a reconstruction of the dictionary $\widetilde{W}_q$ over the one symbol thinner real dictionary W_{q-1}. This makes the expression (1) for the dictionary $\widetilde{W}_q$ as follows:

$$\tilde{f}_{i_1 i_2 i_3 \dots i_q} = \frac{f_{i_1 i_2 i_3 \dots i_{q-1}} \cdot f_{i_2 i_3 \dots i_q}}{f_{i_2 i_3 \dots i_{q-1}}}. \tag{4}$$

Hence, the specific entropy (3) with respect to (4) is defined as

$$\bar{S} = \sum_{\omega} f_{\omega} \cdot \ln \frac{f_{\omega} \cdot f_{\bar{\omega}}^{(q-2)}}{f_{\omega'}^{(q-1)} \cdot f_{\omega''}^{(q-1)}} =$$
$$= \sum_{\omega} f_{\omega} \cdot \ln f_{\omega} + \sum_{\omega} f_{\omega} \cdot \ln f_{\bar{\omega}}^{(q-2)} - \sum_{\omega} f_{\omega} \cdot \ln f_{\omega'}^{(q-1)} - \sum_{\omega} f_{\omega} \cdot \ln f_{\omega''}^{(q-1)} \tag{5}$$

Here ω' and ω'' are the words of the length $q-1$, so that $i\omega' = \omega'' j = \omega$, and $\bar{\omega} = \omega' \cap \omega''$ is the word of the length $q-2$. A summation over the surplus indices in the terms of (5) where the words of different length are present, converts each term in this expansion into the entropy of the real frequency dictionary of the thickness determined by the set of the shortest included words. Finally, it makes the expression for the specific entropy as following:

$$\bar{S} = 2S_{q-1} - S_q - S_{q-2}. \tag{6}$$

The equation (6) turns into

$$\bar{s} = 2S_1 - S_2 \tag{7}$$

for the case of the reconstruction of the dictionary of the thickness 2 over the dictionary of the thickness 1. Obviously, a reconstruction of the dictionary of the thickness q could be performed over the dictionary of the thickness t, so that $q - t > 1$. The frequencies of the words of the reconstructed dictionary are then determined as

$$\tilde{f}_{i_1 i_2 i_3 \dots i_q} = \frac{f_{i_1 i_2 i_3 \dots i_t} \cdot f_{i_2 i_3 \dots i_{t+1}} \cdot f_{i_3 i_4 \dots i_{t+2}} \cdot \dots \cdot f_{i_{q-t} i_{q-t+1} \dots i_{q-1}} \cdot f_{i_{q-t+1} i_{q-t+2} \dots i_q}}{f_{i_2 i_3 \dots i_{t-1}} \cdot f_{i_3 i_4 \dots i_t} \cdot \dots \cdot f_{i_{q-t+1} i_{q-t+2} \dots i_{q-2}} \cdot f_{i_{q-t+2} i_{q-t+3} \dots i_{q-1}}}$$

Table I. Information capacity of four genes and three complete genomes; the length of entity is indicated below the entry

q	U18601	L11265	X69833	X76108	AF086833	X79547	Y19184
	1771	1815	1620	3307	18959	16660	16652
2	0.0586	0.0061	0.0700	0.0414	0.0111	0.0073	0.0080
3	0.0176	0.0094	0.0164	0.0178	0.0028	0.0054	0.0050
4	0.0462	0.0383	0.0447	0.0360	0.0062	0.0116	0.0098
5	0.1789	0.2054	0.1872	0.1099	0.0190	0.0280	0.0226
6	0.3734	0.4976	0.4056	0.2962	0.0735	0.0829	0.0805
7	0.4061	0.4124	0.3869	0.4171	0.2632	0.2805	0.2698
8	0.1996	0.1520	0.1881	0.2917	0.6403	0.3715	0.4189
9	0.0649	0.0512	0.0593	0.1146	0.3109	0.3227	0.3268
10	0.0238	0.0083	0.0169	0.0394	0.0408	0.1555	0.1419

where $i_1 i_2 i_3 \dots i_q$ is the word of the length q (see [19, 20] for more details). The specific entropy is then determined as

$$\bar{S} = (q - t + 1)S_t - S_q - (q - t)S_{t-1}$$

and

$$\bar{S} = qS_1 - S_q$$

for the case of $t = 1$. Everywhere below we shall use the formulae (6, 7) in the study of the information capacity of nucleotide sequences.

4. The Information Capacity of Some Nucleotide Sequences

To illustrate the approach of nucleotide sequences study through their information capacity, we have examined several entities. Table I shows the data for information capacity variation from thickness 2 to 10 for sequences of U18601 (azarae interphotoreceptor retinoid-binding protein gene, intron 1, of Azara's night monkey), L11265 (Argopecten irradians 18S ribosomal RNA), X69833 (European woodmouse mRNA for serine protease inhibitor) and X76108 (European eel mRNA for sodium/potassiuin ATPase, α-subunit). This Table also bears the data on variation of information capacity of six genomes: complete genome of Ebola virus, strain Mayinga (entry AF086833), mitochondrial DNA complete sequence of horse (entry X79547) and *Lama pacos* complete mitochondrial genome (entry Y19184). These genomes have been chosen just to illustrate the efficiency of information analysis of nucleotide sequences through their information capacity.

A comparative study of the information capacity value observed for various segments of a genome provides a student with a new tool for investigating the

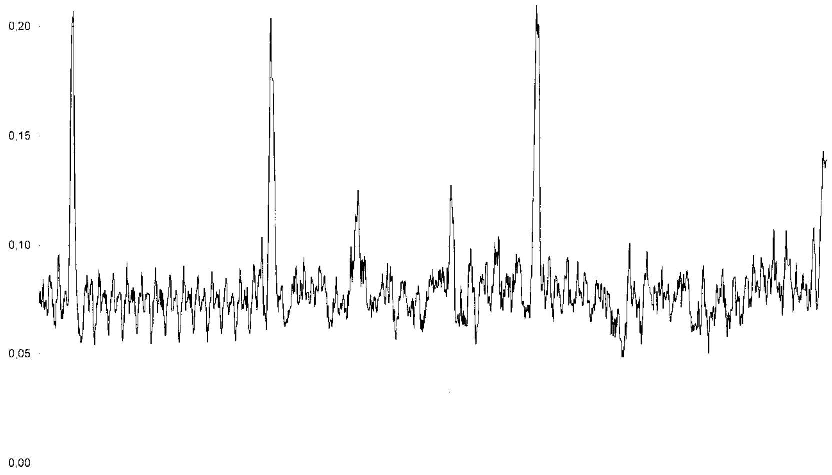

Figure 1. Scanning of genome of Epstein-Barr virus for $q = 2$ (black line), $q = 3$ (dark grey line) and $q = 4$ (grey line). The entire genome is represented at the horizontal axis; vertical axis shows the specific entropy values.

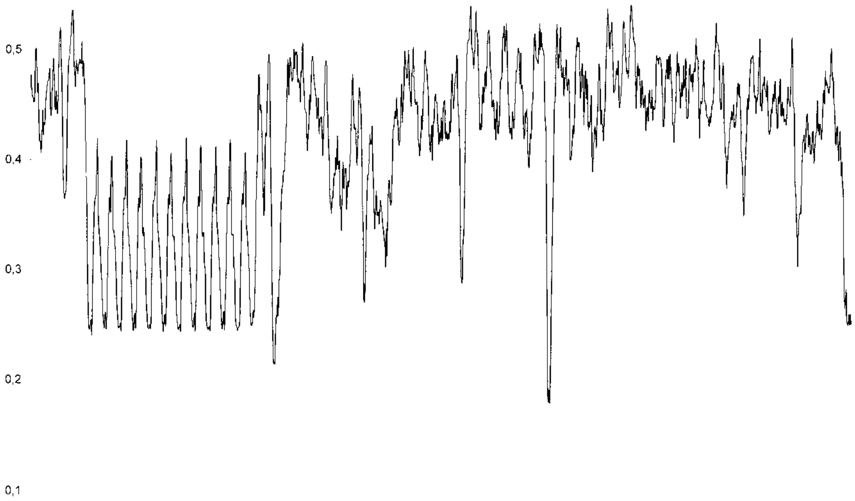

Figure 2. Scanning of genome of Epstein-Barr virus for $q = 5$ (black line) and $q = 6$ (grey line). The entire genome is represented at the horizontal axis; vertical axis shows the specific entropy values.

inner structure of a genome. Consider some nucleotide sequence; let us isolate comparatively long segment in it. Let, then, the position of the isolated segment run consequently, step by step, alongside the sequence with the given step T while the length L of that latter is kept permanent. Determining the information capacity for each segment position provides a researcher with the scanning procedure of a sequence. Figure 1 shows the scan of the genome of Epstein-Barr virus obtained for $T = 60$ and $L = 1500$ nucleotides; this figure presents the scans for di- and trinucleotide composition. Figure 2 shows similar scans obtained for the dictionary thickness $q = 5$ and 6. Horizontal axis represents the genome, from the first nucleotide till the last one; there are no marks at the axis, since we display no custom structure in it (such as gene-junk structure). Vertical axis represents the specific entropy value; the curves observed for dictionaries of various thicknesses are shown in a different colour.

5. Discussion

In the majority of cases, the processing and storage of inherited information takes place step by step, in small portions. Such pattern seems to be rather universal: it takes place both for natural (consider, e.g., a reading process of texts in a natural language) and artificial devices of an information processing (as it runs, for example, in computers). Symbol sequence seems to be the most general form of the information-carrying medium. The content of a message is determined by the occurrence of the greatest possible variety of combinations of symbols within the sequence. Hence, the information capacity of a message is determined by the number of strings of symbols, that are highly unexpected: the greater is the number of these strings, the higher is the information capacity of a message.

Apparently, the concept of an information capacity of a message depends strongly on the anticipation that a string will differ in that point from an information content [5, 11, 14, 19–21, 23, 24, 38]. A researcher has no way to foresee the occurrence of a string of length q, except to consider all continuations of it. Hence, one must admit the hypothesis concerning the procedure of a continuation of the given strings. In general case, this hypothesis may be based on the most common assumption and should be restricted by the weakest possible constraints. For instance, one must avoid an implementation of a symbol which is *á priori* known to fall beyond the alphabet. Another problem is the formulation of a hypothesis itself. There are many ways to estimate the probability of an occurrence of the symbol which is considered to be a continuation of a string. Quite often, a hypothesis may be developed based on the ideas and concepts falling beyond the statistics of symbols itself; such hypothesis may be useful in numerous applications. The key question here is how one should estimate an added information value. There is now clear and unambiguous answer, even if one has nothing to do with ‘semantics’ of various strings. The only way to eliminate any additional or external information or knowledge is to choose the most probable continuations; an extreme principle described above allows to do that explicitly.

To clarify the sense and meaning of the introduced measure of information capacity, consider several ultimate cases. Consider, first, the case of random, non-correlated symbol sequence. Its information capacity is obviously equal (or very much close) to zero. Indeed, the randomness and absence of correlations means that the (real) frequency of any string $i_1 i_2 i_3 \ldots i_q$ is equal to

$$f_{i_1} \cdot f_{i_2} \cdot f_{i_3} \cdot \ldots \cdot f_{i_{q-1}} \cdot f_{i_q},$$

where f_{i_j} is the frequency of j-th nucleotide. The expected frequency of this word is absolutely the same, as the real one, thus making each term in the sum (3) equal to zero, since the ratio of real frequency to the expected one is equal to 1, and ln $1 = 0$.

Highly ordered sequences exhibit the information capacity very close to zero, as well. Indeed, as soon, as the dictionary thickness reaches the period length, the continuation of a string becomes unambiguous, which makes the real and expected frequencies of longer string coincide. Any deviations from zero level of information capacity observed for finite surrogate symbol sequences obtained due to the random non-correlated process result from the finiteness of a sequence, or from the hidden correlations occurred in the imitator of a random process.

The expression (1) is sometimes considered to be an evidence of the Markov property of the original sequence, although that is not true (see, e.g., [39]). The original sequence could be considered as completely corresponding to Markov process if, and only if, the expression (1) holds true for infinitely growing sequence and for the strings of any length (i.e., for infinitely thick dictionary). The main idea of the coincidence of the expression (1) for expected frequency and Markovian property of a chain is that the Markov chain realises with necessity the hypothesis of the most probable continuation of a string; see [19, 20] for more detail.

Information capacity is a new measure for statistical properties of nucleotide sequences. Redundancy is another one, well known measure of statistical properties of these sequences [1, 2, 5, 8, 40]. Usually, the redundancy R is defined as the difference between the real two-particle distribution and the one obtained for the supposition of an independent random distribution of single particles; in our case nucleotides play the role of particles. This deviation is measured through the difference of the entropy values of these two distributions [1, 2, 5, 8, 40]:

$$R = 1 - \frac{S_{ij}}{S_i S_j}. \tag{8}$$

Here S_{ij} is the entropy of dinucleotide distribution, and S_l is the entropy of mononucleotide distribution. There are some other definitions of redundancy, one of them is based on the non-ambiguity of a continuation of a string [32–37]. Here the redundancy is defined as a ratio of the specific thickness d^* of dictionary and the logarithm of the length of a sequence. Redundancy measured according to (8) coincides almost entirely to the information capacity determined for dictionary of the thickness 2. The difference is not essential: $R = 0$, when the information capacity is maximal, and vice versa.

Now let us consider the data presented above in more detail. First of all, one can observe the significant maximum in specific entropy data presented at the table. The information capacity declines, as the dictionary thickness q exceeds 6 for sequences U18601, L11265 and X69833. The capacity goes down, as q exceeds 7 for other sequences. Such decay for longer strings is obvious, and results from the finiteness effect. As the length of a string increases, the number of possible continuations declines. At the end, each string has a single continuation. The biological meaning of such specific length is discussed in [32–37]. The shift in the maximum of information capacity for longer sequences is, therefore, obvious, and will be eliminated for infinitely long entities. Careful examination of the table shows that the information capacity of the dinucleotide composition is greater, than that one of triplets; biologically, this is the most interesting observation. Such inversion of information capacity values means that the greater part of the inherited information is stored in dinucleotides. This fact corresponds to the wobbling hypothesis of the origin of genetic code [14, 41].

Scanning the sequence through the comparison of information capacity of similar segments laying alongside of the primary sequence provides a researcher with the ability and means to identify new structures in genomes. Figures 1 and 2 show the scan of the complete genome of Epstein-Barr virus; this genome bears a long part made of a series of periodic repeats [42]. This pattern is also revealed by the information capacity measurements (see Figures 1 and 2). Similar structures could be revealed through a redundancy measure of a segment [32, 38, 40, 43, 44].

Two questions arise when studying a structure revealed by the scanning with information capacity. The first one is the meaning of the structure; the second one is the problem of correlation between the structure and some other pattern, say, gene junk, or exon-intron structure. The answer to the first question is rather clear. The scan identifies the regions within a sequence, which are either highly ordered, or, on contrary, very close to a random one. A discrimination of these two types of regions could be carried out due to an examination of scanning pattern observed for different lengths of words. A region seems to be rather close to a random sequence if information capacity of that sequence is quite low for any length of words. A region suspected to be highly ordered contrasts in the behavior of information capacity observed for various lengths of words. An examination of Figures 1 and 2 allows to expect that the regions of the Epstein-Barr virus genome of the approximate length of 10^3 nucleotides and located around the nucleotide nos. 7650, 50650, 67900, 89700, 107200 and 170000 are the highly ordered regions. The first region is located quite closely to the promoter, the second region contains a series of 12 very short (125 nucleotides long) repeats, the third region bears 5 51 nucleotide long repeats and 9 repeats of the length of 15 nucleotides, the fourth and fifth regions contain a group of various promoters, and the sixth one bears four terminal repeats of the length of 523 nucleotides.

A question of whether the scan observed due to information capacity calculation correlates to other well known patterns (such as gene junk, or intron-exon structures) should be studied additionally. The Epstein-Barr virus genome yields no distinct, clear and unambiguous correlation between gene junk, or exon-intron structure and the scan plotted with information capacity. Meanwhile, out studies of other genomes, such as genornes of several chloroplasts, and the *B. subtilis* genome show that the scan obtained through the calculation of information capacity distinctly identifies the clusters of ribosornal RNA genes and some other conservative regions within a genome.

The approach developed here has obvious numerous applications in the areas beyond biophysics, bioinformatics or molecular genetics. A study of information capacity of nucleotide sequences provides a researcher with various specific tools. Those who investigate the problem of symbol classification may capitalise a lot upon the approach based on the information capacity studies. It could be the problem of the symbol classification, for molecular biology and biochemistry. Unlike the problem of classification of sequences on the basis of their statistical properties [35], one may pose the problem of the classification of the symbols. It means a reduction of the alphabet cardinality, in the terms of statistical properties of symbol sequences, that is close to a classification of ainino acids, in the terms of biochemistry [20].

6. Conclusion

The main goal of this paper is to introduce a new method of studying statistical and information properties of nucleotide sequences, rather than to carry out a substantial investigation of these properties of some specific entities. An implementation of various stochastic models for an evaluation of the most unexpected strings means bringing in some additional, external information which completely falls beyond the symbol sequence itself.

The main purpose of the paper is to present a new methodology to study the information capacity of nucleotide sequences. The study of information capacity is based on the principle of the most probable continuation of a word. This principle yields the explicit formulae for the frequency of the strings of a reconstructed dictionary. These formulae have nothing to do with any assumption on statistical properties of the sequence under consideration. An explicit implementation of the frequency of the reconstructed strings yields the explicit formulae for the specific entropy of the real dictionary with respect to the reconstructed one.

The information capacity fails to distinguish the random non-correlated sequences from the highly ordered ones; these two types of sequences exhibit the same level of information capacity (close to zero). Unlike the complexity measure of a sequence, its entropy, or redundancy, the information capacity reveals the sequences which contain the greatest possible number of the short strings which seem to be rather unexpected. Further steps in the development of the methodology presented here may be devoted to creating the methods to implement the classifications of symbols in various specific problems, from the point of view of the orderliness (or, unlike, the stochasticity) of the sequences transformed under the implemented classification.

Acknowledgements

I am grateful to Prof. Alexander N. Gorban for his kind and encouraging interest to the studies presented here. Also, I want to extend my thanks to Dr Tatyana G. Popova and Dr Nikolas N. Bugaenko, who helped with programme implementation. My special thanks to Dr J. Priezzhikh for her support in manuscript preparation.

References

1. Waterman, M.S. (ed.): *Mathematical Methods for DNA Sequences,* CRC Press, Boca Raton, 1998, 389 p.
2. Alexandrov, A.A., Alexandrov, N.N., Borodovsky, M.Yu., Kalambet, Yu.A., Kister, A.Z., Mironov, A.A., Pevzner, P.A. and Shepelev, V.A.: *Computer Analysis of Genetic Texts*, Nauka, Moscow, 1990, 264 p.
3. Claverie, J.-M., Sauvaget, I. and Bougueleret, L.: *k*-Tuple Frequency Analysis: From Intron/Exon Discrimination to T-Cell Epitope Mapping, In: R.F. Doolittle (ed.), *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences* Meth. Enzymol. vol. **183**, 1990, pp. 252–281.
4. Karlin, S. and Cardon, L.R.: Computational DNA Sequence Analysis, *Ann. Rev. Microbiol.* **48** (1994), 619–654.
5. Yockey, H.P.: *Information Theory and Molecular Biology*, Cambridge Univ. Press, N.Y., 1992.
6. Kolmogorov, A.N.: Three Approaches to the Quantitive Definition of Information, *Problems of Information Transmission* **1** (1965), 1–17.
7. von Neumann, J.: *Theory of Self-Reproducing Automata*, University of Illinois Press, Urbana, Illinois, 1966.
8. Vitushkin, A.G.: *Theory of Transmission and Processing of Information*, Pergamon Press, New York, 1961.
9. Bugaenko, N.N., Gorban, A.N. and Karlin, I.V.: Universal Expansion of the Triplet Distribution Function, *Teoreticheskaya i Matematicheskaya Fisica* **88** (1991), 430–441. (in Russian)
10. Gorban, A.N. and Karlin, I.V.: Structure and Approximations of the Chapman-Enskog Expansion for Linearized Grad Equations, *Transport Theory & Stat. Phys.* **21** (1992), 101–117.
11. Hariri, A., Weber, B., and Olmsted, J.-3$^{\text{rd}}$: On the Validity of Shannon-Information Calculations for Molecular Biological Sequences, *J. Theor. Biol.* **21** (1990), 235–254.
12. Churchill, G.A.: Stochastic Models for Heterogeneous DNA Seqeuces, *Bull. Math. Biol.* **51** (1989), 70–94.
13. Finkelstein, A.V. and Roytberg, M.A.: Computation of Biopolymers: A General Approach to Different Problems, *BioSystems* **30** (1993), 161–185.
14. Schneider, T.D.: Evolution of Biological Information, *Nucl. Acids Res.* **28** (2000), 2794–2799.
15. Smith, T.F.: Genetic Sequence Semantic and Syntactic Patterns Location, In: G.I. Bell and T.G. Marr (eds.), *Computers and DNA*, Addison-Wesley, 1990, pp. 259–270.

16. Kirkwood, J. and Boggs, E.: The Radial Distribution Function in Liquids, *J. Chem. Phys.* **10** (1942), 394.
17. Gorban, A.N.: *The Bypass of the Balance*, Novosibirsk, Nauka, 1984, 364 p.
18. Balescu, R.: *Equilibrium and Nonequilibrium Statistical Mechanics*, Wiley, New York, 1975.
19. Bugaenko, N.N., Gorban, A.N. and Sadovsky, M.G.: Towards the Determination of Information Content of Nucleotide Sequences, *Mol. Biology* (Russian) **30** (1996), 529–546.
20. Bugaenko, N.N., Gorban, A.N. and Sadovsky, M.G.: Maximum Entropy Method in Analysis of Genetic Text and Measurement of its Information Content, *Open Sys. Inf. Dyn.* **5** (1998), 265–278.
21. Pavesi, A., De Iaco, B., Granero, M.I. and Porati, A.: On the Informational Content of Overlapping Genes in Prokaryotic and Eukaryotic Viruses, *J. Mol. Evol.* **44** (1997), 625.
22. Liaofu Luo, Weijiang Lee, Lijun Jia, Fengmin Ji and Lu Tsai: Statistical Correlation of Nucleotides in a DNA Sequence, *Phys. Rev. E* **58** (1998), 861–871.
23. Shenkin, P.S., Erman, B. and Mastrandrea, L.D.: Information-Theoretical Entropy as a Measure of Sequence Variability, *Proteins* **11** (1991), 297–313.
24. Weiss, O., Jimenez-Montano, M.A. and Herzel, H.: Information Content of Protein Sequences, *J. Theor. Biol.* **206** (2000), 379–386.
25. Yu, Z.G., Anh, V.V. and Wang, B.: Correlation Property of Length Sequences Based on Global Structure of the Complete Genome, *Phys. Rev. E* **63** (2001), 903–910.
26. Crochemore, M. and Verin, R.: Zones of Low Entropy in Genomic Sequences, *Comput. Chem.* **23** (1999), 275–282.
27. Loewenstern, D. and Yianilos, P.N.: Significantly Lower Entropy Estimates for Natural DNA Sequences, *J. Comput. Biol.* **6** (1999), 125–142.
28. Ragosta, M., Cosmi, C., Cuomo, V. and Macchiato, M.: An Application of Maximum Entropy Techniques to Determine Homogeneous Sets of Nucleotidic Sequences, *J. Theor. Biol.* **155** (1992), 129–136.
29. Kirsanova, E.N. and Sadovsky, M.G.: Entropy Approach to a Comparison of Images, *Open Sys. Inf. Dyn.* **8** (2001), 183–199.
30. Kirsanova, E.N. and Sadovsky, M.G.: Method of Statistical Comparison of Objects, *Radioelectr. Inform. Contr.* **2** (2000), 71–82.
31. Kirsanova, E.N. and Sadovsky, M.G.: *On the Anisotropy of Digital Images*, 9th Natl. Conf. Neuroinformatics and its applications, Krasnoyarsk, Oct. 5–7, 2001.
32. Gorban, A.N., Popova, T.G. and Sadovsky, M.G.: A Redundancy of Genetic Sequences and Mosaic Structure of a Genome, *Mol. Biology* (Russian) **28** (1994), 313–322.
33. Gorban, A.N., Mirkes, E.M., Popova, T.G. and Sadovsky, M.G.: Comparative Redundancy of Genes of Various Organisms and Their Viruses, *Rus. J. Genet.* **29** (1993), 1413–1419.
34. Gorban, A.N., Popova, T.G. and Sadovsky, M.G.: Human Genes are more Redundant than Genes of Human Viruses, *Rus. J. Genet.* **32** (1996), 281–294.
35. Gorban, A.N., Popova, T.G. and Sadovsky, M.G.: Classification of Symbol Sequences over their Frequency Dictionaries: Towards the Connection Between Structure and Natural Taxonomy, *Open Sys. Inform. Dyn.* **7** (2000), 1–17.
36. Popova, T.G. and Sadovsky, M.G.: Splicing Results in Decrease of Genes Redundancy, *Mol. Biology* (Russian) **29** (1995), 500–506.
37. Popova, T.G. and Sadovsky, M.G.: Introns Differ from Exons from the Point of View of Their Redundancy, *Rus. J. Genet.* **31** (1995), 1365–1379.
38. Jimenez-Montano, M.A., Ebeling, W., Pohl, T. and Rapp, P.E.: Entropy and Complexity of Finite Sequences as Fluctuating Quantities, *Biosystems* **64** (2002), 23–32.
39. Churchill, G.A.: Hidden Markov Chains and the Analysis of Genome Structure, *Comput. Chem.* **16** (1992), 107–115.
40. Kisliuk, O.S., Boronina, T.A. and Nazipova, N.N.: Evaluation of Genetic Text Redundancy using a High-Frequency Component of the l-gram Graph, *Biofizika* **44** (1999), 639–648.

41. Tao Jiang, Ying Xu and Zhang, M.Q. (eds.): *Current Topics in Computational Molecular Biology*, MIT Press, Cambridge, 2002, 540 p.
42. Cheung, A. and Kieff, E.: Long Internal Direct Repeat in Epstein-Barr Virus DNA, *J. Virol.* **44** (1982), 286–294.
43. Popova, T.G. and Sadovsky, M.G.: Investigating Statistical Properties of Genetic Texts: Local Redundancy Displays a New Structure of Genes, *Adv. Model. & Anal. C, AMSE Press.* **48** (1995), 17–22.
44. Bugaenko, N.N., Popova, T.G. and Sadovsky, M.G.: Information Structure of Genetic Sequences, In: A.N. Gorban (ed.), *Proc. 8th Natl. Conf. Neuroinformatica and its applications*, Krasnoyarsk, KSTU Press, (2000), pp. 26–27.
45. Bugaenko, N.N., Sadovsky, M.G. and Sapozhnikov, A.N.: Classification of Symbols and Alphabet Development Optimal for a Revealing the Statistical Regularities, *Proc. 5th Natl. Conf. Neuroinformatics and its Applications*, Krasnoyarsk, Sept. 22–25, 1997, pp. 27–30.

Appendix I

RECONSTRUCTED DICTIONARY AND PRINCIPLE OF MAXIMUM ENTROPY

The method of the development of reconstructed dictionary takes its origin at the method of invariant manifolds developed for the investigations of Boltzmann equation [9, 10]. The main idea of the method is to derive the frequency of strings at frequency dictionary of the thickness $(q + 1)$ from the data stored in frequency dictionary of the thickness q, only. A prohibition to use an additional information means that the reconstructed dictionary (bearing the strings of length $(q + 1)$), should be as undetermined as possible for a given W_q, i.e., such reconstructed dictionary must have the highest possible entropy among all frequency dictionaries of the thickness $(q + 1)$ [19, 20, 35]:

$$S_{q+1}\left[f_{q+1}\right] \to \max, \tag{I.1}$$

where

$$S_q = -\sum_{i_1 i_2 \ldots i_q} f_{i_1 i_2 \ldots i_q} \mathrm{ln f}_{\mathrm{i}_1 \mathrm{i}_2 \ldots \mathrm{i}_q} \tag{I.2}$$

is the entropy of the frequency dictionary of the thickness q; $f_{i_1 i_2 \ldots i_q}$ is the frequency of a string $i_1 i_2 \ldots i_q$. i_j denotes a symbol (nucleotide, in particular) occupying the j-th position.

Entropy extremum problem (I.1) should be resolved with respect to the obvious constraints

$$\sum_{i_{q+1}} \tilde{f}_{i_1 i_2 \ldots i_q i_{q+1}} = f_{i_1 i_2 \ldots i_q} \tag{I.3.1}$$

$$\sum_{i_{q+1}} \tilde{f}_{i_{q+1} i_1 i_2 \ldots i_q} = f_{i_1 i_2 \ldots i_q} \tag{I.3.2}$$

where q is the dictionary thickness, $i_1 i_2 \ldots i_q$ is a string from a q-thick dictionary, $f_{i_1 i_2 \ldots i_q}$ is the frequency of this string and $\tilde{f}_{i_1 i_2 \ldots i_q i_{q+1}}$ is the frequency from the reconstructed dictionary. The constraints (I.3) distinguish the ensemble of the possible dictionaries of the thickness $(q+1)$: not any frequency set consistent with one of them is consistent with another.

The solution of (I.1–I.3) by the indeterminate Lagrange multiplier method yields

$$\tilde{f}_{i_1 \ldots i_q i_{q+1}} = \exp\left\{ \sum_{\mathrm{i_1 \ldots i_q}} \alpha_{\mathrm{i_1 \ldots i_q}} + \sum_{\mathrm{i_2 \ldots i_{q+1}}} \beta_{\mathrm{i_2 \ldots i_{q+1}}} - 1 \right\} \qquad \text{(I.4)}$$

where $\alpha_{i_1 \ldots i_q}$ and $\beta_{i_2 \ldots i_{q+1}}$ are the indeterminate multipliers satisfying the linear constraints (3). Denoting

$$\alpha'_{i_2 \ldots i_{q+1}} = \exp\left(\sum_{\mathrm{i_2 \ldots i_{q+1}}} \alpha_{\mathrm{i_2 \ldots i_{q+1}}} - \frac{1}{2} \right), \qquad \text{(I.5.1)}$$

$$\beta'_{i_1 \ldots i_q} = \exp\left(\sum_{\mathrm{i_1 \ldots i_q}} \beta_{\mathrm{i_1 \ldots i_q}} - \frac{1}{2} \right) \qquad \text{(I.5.2)}$$

and using the eonstraints (I.3), one gets the solution of the problem (I.1)

$$\tilde{f}_{i_1 i_2 \ldots i_q i_{q+1}} = \frac{f_{i_1 i_2 \ldots i_q} \cdot f_{i_2 i_3 \ldots i_q i_{q+1}}}{f_{i_2 \ldots i_q}}. \qquad \text{(I.6)}$$

where $\tilde{f}_{i_1 i_2 \ldots i_q i_{q+1}}$ is the frequency of a string from the dictionaiy one symbol thicker than the current dictionary; here $q > 1$.

A dictionary of the thickness $(q+s)$ is reconstructed from a dictionary of the thickness q in a similar manner:

$$S_{q+s}\left[f_{q+s}\right] \to \max \qquad \text{(I.7)}$$

with $s+1$ linear constraints

$$\sum_{i_{q+1}, \ldots, i_{q+s}} \tilde{f}_{i_1 \ldots i_q i_{q+1} \ldots i_{q+s}} = f_{i_1 \ldots q} \qquad \text{(I.8.1)}$$

$$\sum_{i_{q+1}, \ldots, i_{q+s}} \tilde{f}_{i_{q+1} i_1 \ldots i_q i_{q+2} \ldots i_{q+s}} = f_{i_1 \ldots q} \qquad \text{(I.8.2)}$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot$$

$$\sum_{i_{q+1}, \ldots, i_{q+s}} \tilde{f}_{i_{q+1} \ldots i_{q+s} i_1 \ldots i_q} = f_{i_1 \ldots q} \qquad \text{(I.8.}s+1)$$

The solution of (I.7, 1.8) is (for $q > 1$)

$$\tilde{f}_{i_1 \ldots i_q i_{q+1} \ldots i_{q+s}} = \frac{f_{i_1 \ldots i_q} \cdot f_{i_2 \ldots i_{q+1}} \cdot \ldots \cdot f_{i_{q-s+1} \ldots i_{q+s}}}{f_{i_2 \ldots i_q} \cdot f_{i_3 \ldots i_{q+1}} \cdot \ldots \cdot f_{i_{q-s+1} \ldots i_{q+s-1}}} \tag{I.9}$$

Equations (I.6) and (I.9) turns into

$$\tilde{f}_{i_1 \ldots i_{q+s}} = f_{i_1} \cdot f_{i_2} \cdot \ldots \cdot f_{i_{q+s}} \tag{I.10}$$

for $q = 1$.

The expressions (I.6, I.9, I.10) for the reconstructed frequencies are entirely similar to the Kirkwood's approximation [16] while they are the exact solutions in our case. The above approach to studying the statistical properties of nucleotide sequence originates in statistical physics [16], where the distribution functions of various states of one or several particles are changed for the strings in our case, and the sets of all such distributions are changed for dictionaries. The problem of reconstructing of three-partial distribution functions from two-partial one is solved in [9, 10]. Contrary to the distribution functions common in statistical physics, the nucleotide sequences are the unidimensional (linear) objects; this fact greatly simplifies the problem.

The formulae (I.9, I.10) coincide with the well-known expressions for the transition probabilities in a symbol sequence obtained as a realisation of the Markov random process, for $s = 1$ (there are some peculiarities for $s > 1$). It should be stressed, that these formulae for the reconstructed dictionary are yielded with no respect to a hypothesis on a peculiar structure of a sequence. These formulae present the hypothesis of the most probable continuations in the frequency dictionary of the thickness $q+s$, that could be implemented from a consideration of the dictionary of the thickness q. One should consider the original symbol sequence to be Markovian one if and only if the expressions for the real (but not the reconstructed) frequencies would hold true in the limit case of infinitely long original sequence.