Random Fields Approach to the Study of DNA Chains

JANUSZ SZCZEPAŃSKI¹ and TOMASZ MICHAŁEK²

¹*Polish Academy of Science, Institute of Fundamental Technological Research, Świętokrzyska 21, 00–049 Warsaw, Poland*

¹*e-mail: jszczepa@ippt.gov.pl*

²*e-mail: tmichal@ippt.gov.pl*

Abstract. We apply the random field theory to the study of DNA chains which we assume to be trajectories of a stochastic process. We construct statistical potential between nucleotides corresponding to the probabilities of those trajectories that can be obtained from the DNA data base containing millions of sequences. It turns out that this potential has an interpretation in terms of quantities naturally arrived at during the study of evolution of species i.e. probabilities of mutations of codons. Making use of recently performed statistical investigations of DNA we show that this potential has different qualitative properties in coding and noncoding parts of genes. We apply our model to data for various organisms and obtain a good agreement with the results just presented in the literature. We also argue that the coding/noncoding boundaries can correspond to jumps of the potential.

Key words: codons, DNA chain, entropy, exons, introns, mutation, random field, stochastic process

1. Introduction

The DNA chains contain the complete genetic information (Yockey, 1992) for construction of a living organism. In general, DNA sequences are strings of the nucleotides *A*, *T*, *C* and *G* containing as substrings genes which in turn consist of coding (exon) sequences and noncoding (intron) sequences. Since the early 1970s, scientists have attempted to discover some kind of order of hidden structures in DNA, to explore and understand the function of genes, to discriminate coding from noncoding regions, to find translation initiation positions. This was made possible due to the huge progress in experimental methods which allow us to recognize and store hundreds of millions of nucleotide sequences. Therefore many statistical approaches have been developed to study the way information is transmitted in living organisms. On the other hand, the DNA chains can be analyzed from the mechanical point of view and here the important fact is that typically they have a double helix structure. It would be interesting to study the relation between these two approaches.

In this paper we regard the DNA chains as trajectories of some stochastic process generated by a source that can be in fact governed by the biological process

in cells. In this approach a fundamental role is played by the description of the probability measure corresponding to this stochastic process. Then, making use of random field theory, we reconstruct the statistical potential of DNA chain pointing out its relations to concepts recently studied in the literature (mutations, mutual information, information, entropy, pair correlations). It is known that in one dimensional case Markov random fields are equivalent to Markov chains with positive cylinder probabilities (Kemeny et al., 1976). The idea to use in our modeling of DNA Markov random fields rather than Markov chains is based on the fact that in this approach we obtain in a simple way formulas relating potential between nucleotides and probabilities of mutations of codons i.e. quantities very important during the process of evolution of species.

2. Basic Concepts of Random Fields

In this section for completeness we provide basic concepts of the random field theory (Kemeny et al., 1976) making some modifications in order to adopt this theory for description of DNA chains.

Let T be a countable set and let S be a finite state space (for example $S = \{0, 1, \dots, m\}$). By Ω we denote the set of all configurations S^T , i.e. the set of all sequences $(s_t)_{t \in T}$. A *random field* can be understood as an association to each sequence $(s_t)_{t \in T}$ its probability $P((s_t)_{t \in T})$ such that $P(\{x_t = i_t; t \in A\}) > 0$ where A is a finite (nonempty) subset of T and $\{x_t = i_t; t \in A\}$ is the set of all sequences with fixed states at positions determined by the set A . Now defining the random variables $X_t : \Omega \rightarrow R$ such $X_t((s_t)_{t \in T}) := s_t$ that we obtain a stochastic process corresponding to this random field with trajectories being elements of Ω . When we try to establish the properties of a source generating the sequences the fundamental question in this framework is how to describe or to introduce this probability to retain the behaviour and properties of the sequences. It turns out that in this case an essential role is played by the concept of Gibbs field. To define the Gibbs field first we introduce the notion of potential.

A *potential* U on a finite set T is a family $\{U_A; A \subset T\}$ of functions from Ω to the real numbers R with the property that $U_A(s) = U_A(s')$ whenever $s_t = s'_t$ for all $t \in A$, and such that $U_\emptyset = 0$. The energy H_U of the potential U is given by

$$H_U(s) = \sum_{A \subset T} U_A(s) \quad (1)$$

A finite random field (i.e. T is finite) is a *Gibbs field* with potential U if

$$P(s) = z^{-1} e^{H_U(s)} \text{ for all } s \in \Omega \quad (2)$$

where $z = \sum_{s \in \Omega} \exp \{H_U(s)\}$ is called the partition function.

The idea to introduce Gibbs field takes its source from statistical mechanics where so called canonical ensembles (the system of particles in thermal equilibrium

with a large system) and grand canonical ensembles (additionally the number of particles can be changed) were described by the formula of the form given in (2) and this formula was derived in a natural way from physical laws (Huang, 1963).

It turned out that in the more general case there is also the possibility to reconstruct the potential having the probabilities P of given sequences generated by the source. The main Theorem (below) states (Kemeny et al., 1976) that to each random field P there corresponds a canonical potential V such that P can be interpreted as a Gibbs field constructed by means of this potential.

2.1. THEOREM

Let (P, Ω) be a finite random field. Then (P, Ω) is a Gibbs field with canonical potential V defined by $V_\emptyset = 0$ and for $A \neq \emptyset$,

$$V_A(s) = \sum_{B \subset A} (-1)^{|A \setminus B|} \ln P(s^B) \quad (3)$$

where s^B is a configuration such that $s_t^B = s_t$ for $t \in B$ and $s_t^B = 0$ otherwise (here $|A|$ means the cardinality of the set A). The potential V is *normalized* which means that $V_A(s) = 0$ if for some $i \in A$ $s_i = 0$.

Now we adopt this formula to the case when we assume that only some close neighbours have influence on a nucleotide in a sequence. To formalize this we first assume that to each element t of T there corresponds some *neighbourhood* $\partial t \subset T$ such that $t \notin \partial t$, and $t \notin \partial t'$ if and only if $t' \notin \partial t$. Denote $\bar{t} = \{t\} \cup \partial t$. The random field P is called a *Markov field* (with respect to ∂) if the following expression for the conditional probabilities holds:

$$\begin{aligned} P_t^A(s) &:= P(\{s' : s'_t = s_t | s'_k = s_k \text{ for all } k \in A - \{t\}\}) = \\ p_t^{\bar{t}}(s) &:= P(\{s' : s'_t = s_t | s'_k = s_k \text{ for all } k \in \partial t\}) \end{aligned} \quad (4)$$

whenever $\bar{t} \subseteq A \subseteq T$.

This formula means that the probability P that at the position denoted by t in the sequence there is some state s_t (element of the sequence s), depends only on states at the positions determined by ∂t . It is natural to call a potential V a *neighbour potential* when $V_A = 0$ for every A such that there exist different elements $a, b \in A$ and $b \notin \partial a$. The family $\wp$ of all subsets C of T such that each $b \in \partial a$ whenever $a, b \in C$, $a \neq b$ is called a *clique family*.

It turned out that the Markov fields are precisely those for which the corresponding potential is a neighbour potential. In this case the formula (3) is of the following form

$$\begin{aligned} V_A(s) &= \sum_{B \subset A} (-1)^{|A \setminus B|} \ln P_t^T(s^B) \quad \text{for } t \in A \in \wp \\ &= 0 \quad \text{for } A \notin \wp \end{aligned} \quad (5)$$

where

$$P_t^T(s) = z^{-1} \exp \left\{ \sum_{B: t \in B \subset \bar{T}} V_B(s) \right\} \quad (6)$$

and $s^B = s_t$ for $t \in B$ and $s_t^B = 0$ otherwise. Consequently we have

$$V_A(s) = \sum_{B \subset A} (-1)^{|A \setminus B|} \ln P_t^T(s^B) \quad \text{for } t \in A \in \wp$$

$$(= 0 \quad \text{for } A \notin \wp) \quad (7)$$

Now we apply this formalism to the study of DNA chains.

3. Adoption of Random Field Theory to DNA Chains

Since at the current level of experimental investigations we have access to the large data base of DNA chains we are able (after making some assumptions concerning stochastic process describing the creation of this data base) to estimate probabilities of the events needed for the reconstruction of the potential between nucleotides in the DNA chain. On the other hand, we have a lot of statistical investigations of these data (next Section) that we can use to characterize properties of the potential just obtained.

In the case of DNA chains for a given species we shall assume that the length of each chain is finite, i.e. $T < +\infty$. The state space consists of four elements called nucleotides (formally we should define an appropriate mapping of the symbols into numbers what is extensively elaborated in the case of proteins (Herzel et al., 1998)). The problem is to find an appropriate family of neighbours of a given nucleotide of DNA, i.e. appropriate definition of the relation ∂ . During the construction of the Gibbs field in the general model it was assumed that the generic role is played by the sequence $\mathbf{0} = (0, 0, \dots, 0)$. In the case of DNA we assume that the *generic sequence* $(g_n)_{n=1}^T$ consists of nucleotides such that for every position $n \in T$ the frequency of the nucleotide g_n is the largest in our data base.

Let us assume that the operation ∂ is such that $\partial 0 = \{1\}$, $\partial T = \{T - 1\}$, and

$$\partial n = \{n - 1, n + 1\} \quad (8)$$

for every $n \geq 1$. Thus any clique set is of the form $A = \{n - 1, n\}$ or $\{n\}$, $n \geq 1$. Let $g = (g_n)_{n=1}^T$ be the generic sequence for a given species.

In this notation we have the following formula for the potential

$$V_{\{n, n+1\}}(s) = \sum_{B \subset \{n, n+1\}} (-1)^{|[n, n+1] \setminus B|} \ln P(s^B) =$$

$$\ln P(s^{n, n+1}) - \ln P(s^n) - \ln P(s^{n+1}) + \ln P(g) =$$

$$\ln \frac{P(s^{n, n+1}) P(g)}{P(s^n) P(s^{n+1})} \quad (9)$$

From this we find that the potential of the n -th nucleotide denoted by $V^{(n)}(s)$ in the configuration s is

$$\begin{aligned} V^{(n)}(s) &= V_{\{n-1,n\}}(s) + V_{\{n,n+1\}}(s) + V_{\{n\}}(s) + \ln P(g) = \\ &\ln \frac{P(s^{n-1,n})P(g)}{P(s^{n-1})P(s^n)} + \ln \frac{P(s^{n,n+1})P(g)}{P(s^n)P(s^{n+1})} + \ln \frac{P(s^n)}{P(g)} + \ln P(g) \end{aligned} \quad (10)$$

where P is the probability function corresponding to the stochastic process describing the creation of DNA chains. Observe that here $s^B = (s'_n)_{n=1}^T$ means the sequence of DNA such that $s'_n := g_n$ for $n \in T \setminus B$ and $s'_n := s_n$ for $n \in B$.

In the next Section we present some concepts that are actually used to describe DNA chains and then we make use of the relations between them and the terms that are used to reconstruct statistical potential.

4. Statistical Tools to Study DNA Chains

Before considering the properties of the potential just obtained we recall briefly some recent results that are related to the problem of reconstruction of this potential. These results concern statistical investigations of DNA with particular attention devoted to discovering differences between exons and introns. Applying these techniques to individual sequences one must make some general assumptions concerning stochastic processes describing the generation of DNA such as stationarity, ergodicity to ensure an appropriate rate of convergence of estimators used. The main issue is finding some area of DNA where the chain exhibits randomness or finding deterministic rules (or more generally correlations) governing the generation of DNA (or its introns and exons areas). In general, randomness means that for any n the generator produces all sequences with length n with the same probability. There are many statistical tests which describe ‘typical properties’ of random systems where ‘typical’ in the language of statistics means that the probability of occurrence of the property within a species is large and should increase when we consider longer and longer parts of the sample (chain). So the concept of typical property is a probabilistic counterpart of the concept called in the literature as “trait”. On the other hand, it is known that the reliability of statistical tests increases if the length of the sample is large and such situation occurs in the case of DNA.

Different techniques, such as spectral analysis (Coward, 1997; Voss, 1992), detrended fluctuation analysis (Buldyrev et al., 1995; Buldyrev et al., 1998) fractal structures and random walk analysis (Roman-Roland et al., 1998; Peng et al., 1992; Dreismann and Larhammer, 1993), have been used to determine the information content and the properties of DNA sequences. Recently, the wavelet transform has been proposed as a very powerful technique for analysis of range correlations for DNA sequences (Arneodo et al., 1995; Arneodo et al., 1995; Audit et al., 2002). In most cases long-range correlations were reported to exist in introns, while exons were close to random sequences.

One of the natural measures of randomness is entropy. Recently, to study DNA chains some authors (Farach et al., 1995) applied new algorithms (entropy estimators) with fast convergence rates based on the pattern's recurrence time (Wyner et al., 1998). These investigations imply that the coding areas are closed to random while the non-coding areas are much more correlated. The results were recently confirmed in the paper by Barral et al., 2000, where the so called nonlinear modeling technique was applied. This approach explores, quantitatively, similarity along the sequence at subchain vicinities of equal subchains of size d (the embedding dimension). This could also suggest that in introns there is a rule coded to define the splice junctions, i.e. the intron/exon boundaries (Farach et al., 1995).

Another measure of statistical dependence recently intensively explored is so called mutual information (Cover and Thomas, 1991; Grosse et al., 2000) which is defined as follows:

$$I(k) = \sum_{i,j=1}^4 p_{ij}(k) \log_2 \frac{p_{ij}(k)}{p_i p_j} \quad (11)$$

where p_i is the probability of finding the nucleotide A_i , $p_{ij}(k)$ is the joint probability of finding the symbol A_i and k letters downstream the symbol A_j . $I(k)$ gives the information on the letter A_j knowing the letter A_i . The study of this quantity (Grosse et al., 2000; Luo et al., 1998) shows that for most species mutual informations which correspond to interactions within codons, i.e. $I(1)$ and $I(2)$ are much larger in all noncoding regions than in the coding regions what support further arguments concerning stronger statistical independence of generation nucleotides in exon areas as in intron areas.

In the next Sections we show the consequence of these results on the properties of the potential just reconstructed. We start by adoption of the formula (10) to the form corresponding to the probabilities of 'codons'.

5. Codons Interpretation of the Statistical Potential

Since in DNA chains there is complete information concerning construction of proteins of living organisms an important point in the study of the evolution process should be an investigation of influence of mutations of nucleotides in the chains and possibly generation of longer chains with more complex or new structures (Kreitman and Cameron, 1999). Formula (10) shows that the potential V_A can be reconstructed from the probabilities P of configurations s^B which are mutations of the generic chain (for a given species) at positions determined by B . On the other hand, it is an interesting fact that these probabilities can be understood as the Shannon information contained in the event that such mutations occur.

It is known that in the process of transferring information from DNA to the creation of proteins the crucial role is played by so called 'codons', i.e. triples of consecutive nucleotides of DNA. Now we adopt the formula (10) to the form expressed in terms of probabilities of mutations of codons.

We make use of the following general formula connecting conditional probabilities with probabilities of mutations of the generic chain for a finite random field

$$\frac{P_n^T(s^A)}{P_n^T(s^{A \cup \{a\}})} = \frac{P(s^A)}{P(s^{A \cup \{a\}})} \quad (12)$$

where $a \in T$ and $A \subseteq T$ and $a \notin A$. Applying (12) and then making use of the Markov field property with the neighbour sets defined by (8) and using the definition of conditional probability we have

$$\begin{aligned} V_{\{n-1, n\}}(s) &= \ln \frac{P(s^{n-1, n})P(g)}{P(s^{n-1})P(s^n)} = \ln \frac{P_n^N(s^{n-1, n})P_n^N(g)}{P_n^N(s^{n-1})P_n^N(s^n)} = \\ &= \ln \frac{P(s_n | s_{n-1} \text{ and } g_{n+1})P(g_n | g_{n-1} \text{ and } g_{n+1})}{P(g_n | s_{n-1} \text{ and } g_{n+1})P(s_n | g_{n-1} \text{ and } g_{n+1})} = \\ &= \ln \frac{P(s_{n-1} \text{ and } s_n \text{ and } g_{n+1})P(g_{n-1} \text{ and } g_n \text{ and } g_{n+1})}{P(s_{n-1} \text{ and } g_n \text{ and } g_{n+1})P(g_{n-1} \text{ and } s_n \text{ and } g_{n+1})} \end{aligned} \quad (13)$$

Applying similar transformations to terms in (10) we find that the potential energy of the n -th nucleotide in the chain s is

$$\begin{aligned} V^n(s) &= \ln \frac{P(s_{n-1} \text{ and } s_n \text{ and } g_{n+1})P(g_{n-1} \text{ and } g_n \text{ and } g_{n+1})}{P(s_{n-1} \text{ and } g_n \text{ and } g_{n+1})P(g_{n-1} \text{ and } s_n \text{ and } g_{n+1})} + \\ &+ \ln \frac{P(s_n \text{ and } s_{n+1} \text{ and } g_{n+2})P(g_n \text{ and } g_{n+1} \text{ and } g_{n+2})}{P(s_n \text{ and } g_{n+1} \text{ and } g_{n+2})P(g_n \text{ and } s_{n+1} \text{ and } g_{n+2})} + \\ &+ \ln \frac{P(g_{n-1} \text{ and } s_n \text{ and } g_{n+1})}{P(g_{n-1} \text{ and } g_n \text{ and } g_{n+1})} + \ln P(g) = \\ &= \ln \left(\frac{P(s_{n-1} \text{ and } s_n \text{ and } g_{n+1})P(s_n \text{ and } s_{n+1} \text{ and } g_{n+2})}{P(s_{n-1} \text{ and } g_n \text{ and } g_{n+1})P(s_n \text{ and } g_{n+1} \text{ and } g_{n+2})} \right. \\ &\quad \left. \frac{P(g_n \text{ and } g_{n+1} \text{ and } g_{n+2})}{P(g_n \text{ and } s_{n+1} \text{ and } g_{n+2})} \right) + \ln P(g) \end{aligned} \quad (14)$$

Finally, we obtained the formula of the potential of the n -th nucleotide in terms of the probabilities of mutations of codons. These expressions can be calculated with high accuracy from DNA data bases.

Observe that when we assume in terms of formula (14) that there is no correlations between nucleotides within codons (or correlations are weak) the probabilities under logarithms factorize (after this terms in numerators and denominators cancel) and we obtain under logarithm quantities that are 1 (or close to one so the logarithms are close to zero) thus the support of these terms to the potential is zero (or is small).

Table I. List of organisms to which we apply the model

	Regnum	Description/organisms	Sequence length	Number of total sequences	Fano factor
1	Eukaryota	Mammalia			
		Bos taurus (cow) gen for MHC class I heavy chain	451	48	12.687
2	Eukaryota	Mammalia			
		Homo sapiens, (exon region)	300	214	14.135
3	Eukaryota	Aves			
		Gallus gallus (chicken)	352	99	10.484
4	Eukaryota	Mammalia			
		Canis familiaris (dog) dla-88 gen	270	88	7.829
5		Virus			
		Human immunodeficiency virus (HIV)	550	181	23.104
6	Prokaryota	Bacteria			
		Uncultured pig faeces bacterium	558	219	39.121

6. Results of Numerical Calculations

We apply the model to real data for various organisms (Table I) making use of nucleotide sequences coming from the EMBL Nucleotide Sequence database maintained by European Bioinformatics Institute. We perform the calculations for Eukaryotes sequences as well as Prokaryotes. For comparison we also present results of application of the model to ‘random’ and ‘bias’ sequences.

To analyze the properties of the potential $V^n(s)$ we calculate, within each set of DNA sequences for a given organism, the value of the first term (we denote this term by $V[n](s)$) on the right hand side of the formula (14). This is due to the difficulty of the calculation of the value of the second term $\ln P(g)$. The term $\ln P(g)$ plays similar role as Grand Canonical Partition Function in statistical physics. It is well known that for most physical statistical models it is not easy to find their partition function. However one can see that in our model the value of the term $\ln P(g)$ is smaller for the random case than for any other one. This follows from the simple observation that the probability of the generic sequence g for a non-random case is larger than for a random one. This implies that the potential $V^n(s)$ of bases in the random area of the DNA sequences is smaller than the potential of bases in non-random area.

Typical graphs of the potential $V[n](s)$ are presented in Figures 1, 2, 3, 4. We consider several interesting cases.

Figure 1. The graph of potential $V[n](s)$ for a typical nucleotide sequence s (from base 1 to base 256 is the exon area while from base 257 to base 451 is the intron area) in the case of *Bos taurus* (cow). We can observe qualitative difference of the potential for the exon and the intron areas.

In Figure 1 we present a graph of $V[n](s)$ for the DNA sequence of Eukaryota, *Bos Taurus* (cow) that consists of two areas: exon and intron. The exon area is between the bases 1-256 while the intron area is between the bases 257-452. One can observe the qualitative difference of the potential within these areas. The variability of the potential in the exon area is essentially smaller. This suggests that the mutations of bases in exon area for Eukaryotes are almost equiprobable whereas in intron parts the probabilities of mutations are more diverse.

We consider the next case of the data in order to compare the exon area of *Homo Sapiens* with the calculations for random sequences. In Figure 2 one can see similarity of these two examples. Moreover they have close Fano factors, which are for these organisms relatively smaller than for other examples (see Table I). In comparison with the previous Figure the potential in this case is more variable since it was built only on the exon area.

Other examples of the potential for the exon area are presented in Figure 3. Here we show the results of calculations for *Gallus gallus* (chicken) and *Canis familiaris* (dog). We can observe that there is a large number of areas for which the potential is constant what is in good agreement with the results presented in previous Figures. Calculations of the Fano factors for these examples confirm small variability of the potential within the exons areas.

Figure 2. The graph of potential $V[n](s)$ calculated for a nucleotide sequence s : a) containing only the exon area in the case of Homo sapiens, b) for a random sequence. One can see similarity.

Figure 3. The graph of potential $V[n](s)$ for a typical nucleotide sequence s (exon area) in the case of: a) *Gallus gallus* (chicken), b) *Canis familiaris* (dog). We can see that there are large areas where the potential is flat.

Figure 4. The graph of potential $V[n](s)$ for a typical sequence s of bases in the case of: a) Virus (HIV), b) Bacteria (uncultured pig faces bacterium), c) 'bias' sequence. Areas where the potential is flat are very short.

Figure 4. Continued.

We apply our model also to the base sequences coming from Virus and Bacteria. Typical graphs of the potentials for these organisms are presented in Figure 4. For comparison we also include the graph of the potential for 'bias' sequences. One can see that the variability of the potential is larger for Virus, whereas for Bacteria the potential $V[n](s)$ has flatter areas. Moreover one can spot similarity between potential for the Bacteria and 'bias' case (Figure 4c). These suggest that for the Bacteria the mutations of adjacent bases are more likely than single mutations.

Our conclusions following from the model concerning properties of coding and non-coding areas are in a good agreement with the recent results of other authors. Similarly to the results of these papers our calculations of the potentials (14) show that coding regions are more random. Moreover we also argue that the tolerance of non-coding regions to various mutations is higher than in the coding ones (Buldyrev et al., 1998; Li et al., 1994).

7. Final Remarks

The basic form of DNA is a double helix, consisting of two backbones and a chain of nucleotide pairs inside. Therefore one of the natural way to investigate DNA or peptide chains is a mechanical approach the aim of which is studying (after some assumptions concerning mainly the choice of a model and potential of interactions) typical configurations, solitary-wave solutions (Calladine and Drew, 1992; Lipni-

acki, 1999), base opening (Chen et al., 2000) and generally its dynamics. These considerations allow a better understanding of important processes appearing during the transfer of information from DNA to the creation of proteins since this is connected with a mechanical process of splitting of pairs of nucleotides. On the other hand, at present we have access to large data bases of DNA sequences. Therefore it could be useful to develop a statistical approach and to find a relation between these two approaches; an especially challenging problem is the description of potential of interactions.

In the paper we investigate the model in which the DNA chains are trajectories of some stochastic process. After making some general assumptions concerning the process we reconstruct the potential of nucleotides corresponding to the interactions along the helix. It turns out that this potential depends on the probabilities of mutations in the codons constituting the DNA chain. Now observe that the terms appearing in the sums (10), (14) are small when the correlations between the corresponding nucleotides are small and in general should increase when the correlations between nucleotides increase. This is due to the factorization of probabilities within codons in the case of statistical independence (see remark at the end of the previous section). Thus, following the results of Arneodo et al., 1998; Barral et al., 2000; Buldyrev et al., 1995; Farach et al., 1995; Grosse et al., 2000; Luo et al., 1998; Voss, 1992 we see that the potential in the coding area (being close to random) is smaller than the potential in the non-coding area. This difference of potentials should play an essential role in the process of transcription (particularly when exons are separated from introns to get mRNA). Moreover from the form of the potential obtained it follows that the jump of the potential should be between random and non-random part which coincide with the coding/noncoding boundary of a DNA chain (e. g. Figure 1).

In the paper we present a random field model of DNA with a definition of neighbouring sets leading in a natural way to probabilities of codons. However, it is worth noticing that this approach can be applied to other systems like proteins or more generally polymers. The central point in such constructions is an appropriate definition of clique sets to take into account correlations between units of the system. This seems to be possible due to the intensive statistical investigations of correlations between the units for many systems of interest.

Acknowledgements

The author is grateful to Polish Committee for Scientific Research (KBN) for financial support, grant no. 7T 07A 01913. All nucleotide sequences come from the EMBL (European Molecular Biology Laboratory) Nucleotide Sequence database maintained by European Bioinformatics Institute (EBI) and are accessible through WWW (<http://srs6.ebi.ac.uk>) or FTP (<ftp://ftp.ebi.ac.uk/pub/databases/embl/>).

References

- Arneodo, A., d'Aubenton-Carafa, Y., Audit, B., Bacry, E., Muzy, J.F. and Thermes, C.: What Can We Learn With Wavelets about DNA Sequences, *Physica A* **249**(Feb.1) (1998), 439–448.
- Arneodo, A., Bacry, E., Graves, P.V. and Muzy, J.F.: Characterizing Long-Range Correlations in DNA Sequences from Wavelet Analysis, *Phys. Rev. Lett.* **74**(16) (1995), 3293–3296.
- Audit, B., Vaillant, C., Arneodo, A., d'Aubenton-Carafa, Y. and Thermes, C.: Long-Range Correlations between DNA Bending Sites: Relation to the Structure and Dynamics of Nucleosomes, *J. Mol. Biol.* **316**(4, Mar. 1) (2002), 903–918.
- Barral, J.P., Hasmy, A., Jimenez, J. and Marcano, A.: Nonlinear Modeling Technique for the Analysis of DNA Chains, *Phys. Rev. E* **61**(2, Feb.) (2000), 1812–1815.
- Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.K., Simons, M. and Stanley, H.E.: Long-Range Correlation Properties of Coding and Noncoding DNA Sequences: GenBank Analysis, *Phys. Rev. E* **51** (1995), 5084–5091.
- Buldyrev, S.V., Dokholyan, N.V., Goldberger, A.L., Havlin, S., Peng, C.K., Stanley, H.E. and Viswanathan, G.M.: Analysis of DNA Sequences Using Methods of Statistical Physics, *Physica A* **249** (1998), 430–438.
- Calladine, C.R. and Drew, H.R.: *Understanding DNA*, Academic Press, New York, 1992.
- Chen, Y.Z., Mohan, V. and Griffey, R.H.: Base Opening in RNA and DNA Duplexes: Implication for RNA Stability, *Phys. Rev. E* **61**(5, May) (2000), 5640–5645.
- Cover, T.M. and Thomas, J.A.: *Elements of Information Theory*, J. Wiley & Sons, New York, 1991.
- Coward, E.: Equivalence of two Fourier Methods for Biological Sequences, *J. Math. Biol.* **36** (1997), 64–70.
- Dreisammann, A.C. and Larhammer, D.: Long-Range Correlations in DNA, *Nature* **361** (1993), 212–213.
- Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A. and Ziv, J.: *On the Entropy of DNA: Algorithms and Measurements based on Memory and Rapid Convergence*, Symposium on Discrete Algorithms (SODA) 1995.
- Grosse, I., Herzel, H., Buldyrev, S.V. and Stanley, H.E.: Species Independence of Mutual Information in Coding and Noncoding DNA, *Physical Review E* **61**(5, May) (2000), 5624–5629.
- Herzel, H., Trifonov, E.N., Weiss, O. and Grosse, I.: Interpreting Correlations in Biosequences, *Physica A* **249** (1998), 449–459.
- Huang, K.: *Statistical Mechanics*, John Wiley and Sons, Inc., New York, 1963.
- Kemeny, J., Snell, J.L. and Kanpp, A.W.: *Denumerable Markov Chains*, New York-Heidelberg-Berlin, Springer Verlag, 1976.
- Kreitman, M. and Cameron, J.M.: Coding Sequence Evolution, *Curr. Opinion Genet. Devel.* **9** (1999), 637–641.
- Li, W.-H., Marr, T.G. and Kaneko, K.: Understanding Long-Range Correlations in DNA Sequences, *Physica D* **75** (1994), 392–416.
- Lipniacki, T.: Chemically Driven Travelling Waves in DNA, *Phys. Rev. E* **60**(6, December) (1999), 7253–7261.
- Luo, L., Lee, W., Jia, L., Ji, F. and Tsai, L.: Statistical Correlation of Nucleotides in a DNA Sequence, *Phys. Rev. E* **58**(1, July) (1998), 861–871.
- Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H.E.: Long-Range Correlations in Nucleotide Sequences, *Nature* **356** (1992), 168–170.
- Roman-Roldan, R., Bernaola-Galvan, P. and Oliver, J.L.: Sequence Complexity of DNA through an Entropic Segmentation Method, *Phys. Rev. Lett.* **80** (1998), 1344–1347.
- Voss, R.F.: Evolution of Long-Range Fractal Correlations and $1/f$ Noise in DNA Base Sequences, *Phys. Rev. Lett.* **68** (1992), 3805–3808.
- Wyner, A.D., Ziv, J. and Wyner, A.J.: On the Role of Pattern Matching in Information Theory, *IEEE Trans. on Inform. Theory* **44**(6, October) (1998), 2045–2056.

Yockey, H.P.: *Information Theory and Molecular Biology*, Cambridge University Press, 1992.