STATISTICAL PATTERNS IN PRIMARY STRUCTURES OF THE FUNCTIONAL REGIONS OF THE GENOME IN Escherichia coli. I. FREQUENCY CHARACTERISTICS

M. Yu. Borodovskii, Yu. A. Sprizhitskii, E. I. Golovanov, and A. A. Aleksandrov

UDC 576.315.42

The frequency of mono- and dinucleotides is analyzed in sequenced fragments of the E. coli genome of 135,000 nucleotides total length. It is shown that DNA regions differing in functional properties also differ in the correlation parameters of neighboring nucleotides. Furthermore, it is noted that correlation characteristics in the coding regions of E. coli DNA are periodically dependent upon the position occupied by neighboring nucleotides with respect to the initiating codon. We discuss the evolutionary significance of the patterns found, as well as their potential use in the special statistical models of nucleotide sequences necessary for developing algorithms for the computer recognition of functional units in the genome.

The development of methods for the rapid sequencing of nucleic acids was soon followed by the active study of the statistical properties of their primary structures [1-11]. This research direction is closely associated with the use of computer data bases for the analysis of a common problem – determining the relationship between the physical structure and biological functions of DNA. This already quite familiar approach involves the search among elements of the data base for the nucleotide sequence completely or partially homologous to an earlier-specified DNA fragment. Information available on this sequence would permit progress (sometimes quite substantial [12]) in the study of the functional properties of the primary structure of interest. As the volume of data bases expands this methods ever more effectively reduces the expenditures for biochemical experiments. However, on the other hand, computational problems arise associated with the search for homologies throughout the bank.

It would doubtless be useful to turn to a relatively small number of "integral" characteristics of nucleotide sequences, the comparison of which would be sufficient for solving the problem of the functional classification ("recognition of function") of a particular fragment. It is natural to assume that these characteristics would be of substantial physical and biological significance; therefore their search, study, and interpretation is of independent interest.

A large number of such "integral" parameters can be obtained from a statistical analysis of DNA primary structures [1, 2, 10]. The frequencies of mono- and dinucleotides, examined in the present paper, are also characteristics of this type.

It should be noted that the opinion is encountered in the literature that the statistical rules of alternation of nucleotides in DNA primary structures (which are reflected in the nucleotide frequencies) are stable throughout the genome of each organism [4, 5, 8, 9] and even partially explain the phenomenon of the selection of synonymous codons [6, 7, 13]. The results obtained here as well as in the work of Smith et al. [10] enable arguments to be made in support of the opposite viewpoint.

METHODS AND RESULTS

The sample of nonoverlapping E. coli DNA fragments of a total length of 135,000 nucleotides was taken from the third edition of the EMBL nucleotide sequence bank. The coding and noncoding DNA fragments formed two subsamples of 79,900 and 42,600 nucleotides length, respectively. All results presented below were ob-

Institute of Molecular Genetics, Academy of Sciences of the USSR, Moscow. Translated from Molekulyarnaya Biologiya, Vol. 20, No. 4, pp. 1014-1023, July-August, 1986. Original article submitted September 18, 1985.

0026-8933/86/2004-0826 $12.50

TABLE 1. Frequencies of Nucleotides (f_a) and Mean Squared Errors of Determination of These Quantities (σ) for Various Samples of Nucleotide Sequences

Sample	f_T	f_C	f_A	f_G	σ
E. coli as whole	0,243	0,243	0,252	0,262	0,0012
Coding regions	0,231	0,251	0,246	0,272	0,0015
Noncoding regions	0,259	0,231	0,261	0,248	0,0020

tained by a computer analysis of the foregoing masses of information.

We introduce essential definitions.

By the frequency f_a of a nucleotide of type a is understood, as is usual, the ratio of the number $N(a)$ of nucleotides of type a present in a given sequence to the total length of the sequence. The f_a quantities (which are more logically indexed by letters rather than numerals, i.e., a = T, C, A, and G) are presented in Table 1. Their values confirm the known fact that the content of C and G is higher in coding regions than the content of A and T, and vice versa in noncoding regions.

The number of dinucleotides of type ab found in some group of M nucleotide sequences of a total length of N nucleotides is designated by $N(ab)$. We shall examine a very simple statistical model according to which the appearance of the nucleotides a and b in neighboring positions is due to independent causes. On the basis of this model, the expected number of dinucleotides of type ab is $(N-M-1)\cdot f_a \cdot f_b$; this number is designated $\overline{N}(ab)$. We note that the mean squared deviation of the $\overline{N}(ab)$ is $(N(ab))^{1/2}$. Figure 1 presents the quantities $D(ab) = (N(ab) - \overline{N}(ab))/(\overline{N}(ab))^{1/2}$ for all 16 possible ab dinucleotides. Figure 1a refers to the entire set of E. coli DNA fragments; 1b, to the sample of noncoding regions; and 1c, to the sample of coding regions.

If the samples under examination were described by the simplest statistical model, the corresponding quantities $t = \sum_{ab} [D(ab)]^2$ would have a χ^2 distribution with nine degrees of freedom (χ^2_9). In fact, the t quantities for cases 1a, b, and c are 2,986, 2,461, and 616, respectively. The values obtained are "great" in the sense that they appear to reject the null hypothesis with a probability very close to one. In such situations both here and later we restrict ourselves to a level of significance of "no less than 0.999." Thus, the obtained result indicates the existence of a relationship between neighboring nucleotides in E. coli DNA both in coding and in noncoding regions.

As correlation parameters we shall use the magnitudes of the probabilities of appearance of a nucleotide of type b after a nucleotide of type a, i.e., the conditional probabilities $P(b|a)$. Their approximate values $\hat{P}(b|a)$ are calculated from the formulas $\hat{P}(b|a) = N(ab)/N(a)$, a, b = T, C, A, and G. These quantities are presented for the three samples of E. coli nucleotide sequences in Table 2.

It is easily verified that the f_a frequencies satisfy the system of equations

$$\sum_a \hat{P}(b|a)\cdot f_a = f_b. \qquad (1)$$

This means that the data presented in Table 2 can be considered as three matrices of transitional probabilities $P(b|a)$ for uniform first order Markov chains, which are the next, more complicated means for statistically describing nucleotide sequences compared with the simplest model. In a uniform Markov model, f_a represents the approximate values of the so-called final probabilities P_a of the chain states, which in the given case are T, C, A, and G.

Clearly, the $\hat{P}(b|a)$ quantities, calculated for the overall sample (Table 2, A), should occupy some intermediate position with respect to the corresponding quantities $\hat{P}^N(b|a)$ and $\hat{P}^C(b|a)$ from Table 2, B and C, which, as is easily shown, is the case. The following considerations can be used to answer the question of the possibility of a statistically significant agreement between correlation parameters of neighboring nucleotides for both the coding and noncoding regions.

We define the quantities t^{CN} and t^{NC}, where C is the index of coding regions and N is the index of noncoding regions, by the equations:

TABLE 2. Conditional Probabilities $\overset{>}{P}(b|a)$ for E. coli DNA

First nucleotide	Total sample (A)				Noncoding region (B)				Coding region (C)			
	second nucleotide											
	T	C	A	G	T	C	A	G	T	C	A	G
T	0,267	0,235	0,185	0,309	0,309	0,220	0,208	0,266	0,234	0,247	0,173	0,346
C	0,235	0,222	0,251	0,296	0,234	0,234	0,273	0,264	0,231	0,215	0,235	0,319
A	0,242	0,222	0,325	0,214	0,253	0,207	0,318	0,222	0,232	0,236	0,329	0,207
G	0,229	0,294	0,244	0,233	0,238	0,270	0,246	0,246	0,224	0,305	0,243	0,228

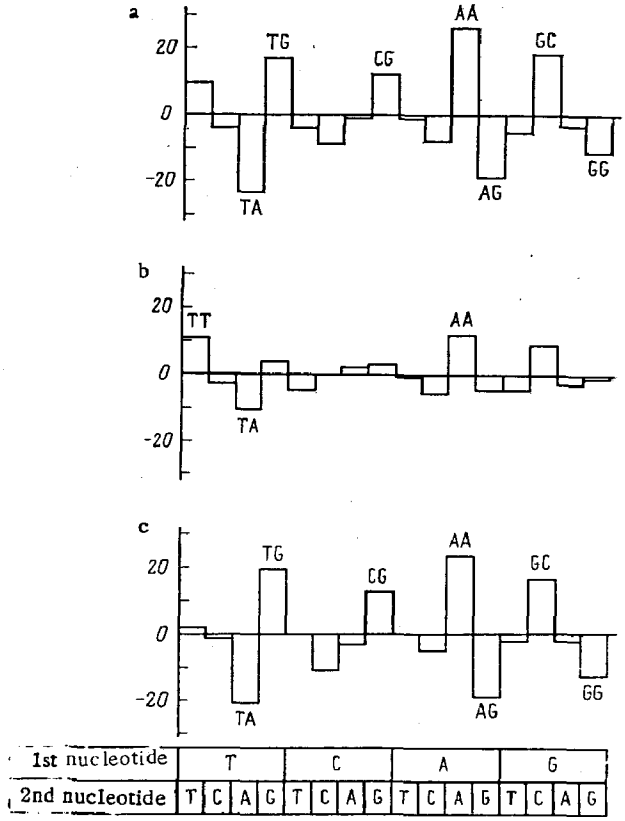

Fig. 1. Shift in dinucleotide frequency: a) relates to whole set of E. coli DNA fragments; b), to sample of noncoding regions; c), to sample of coding regions.

$$t^{NC}=\sum_{ab}(N^{N}(ab)-N^{N}(a)\cdot\overset{>}{P}^{C}(b|a))/(N^{N}(a)\cdot\overset{>}{P}^{C}(b|a))^{1/2},$$

$$t^{CN}=\sum_{ab}(N^{C}(ab)-N^{C}(a)\cdot\overset{>}{P}^{N}(b|a))/(N^{C}(a)\cdot\overset{>}{P}^{N}(b|a))^{1/2}.$$

If the character of the dependence of neighboring nucleotides in coding and noncoding regions is identical, random t^{NC} and t^{CN} quantities should have a χ^2_{12} distribution [14]. The actual t^{NC} and t^{CN} values are 1,556 and 1,568, respectively. Thus, the hypothesis of the agreement of correlation characteristics is rejected with a probability of no less than 0.999. We note that this conclusion contradicts a conclusion made earlier [4, 5].

We now move to a more detailed study of the statistical properties of the nucleotide sequences from coding regions. The available experimental material is sufficiently large and permits the determination of the statistical characteristics of the frequency of mono- and dinucleotides as a function of the position they occupy relative to the initiating codon. The multitude of all possible positions is broken down into groups, termed "frames." For example, nucleotides located at positions 1 + 3c, c = 0, 1, ... form the first mononucleotide frame; moreover, positions with c = 0 correspond to the first nucleotide of the initiating codon. Nucleotides found in positions 2 + 3c, c = 0, 1, ... form the second mononucleotide frame, while those found

TABLE 3. Positional Frequencies of Nucleotides

Frame	T	C	A	G
1	0,140	0,240	0,249	0,371
2	0,289	0,225	0,311	0,175
3	0,263	0,288	0,180	0,270
Σ	0,231	0,251	0,246	0,272

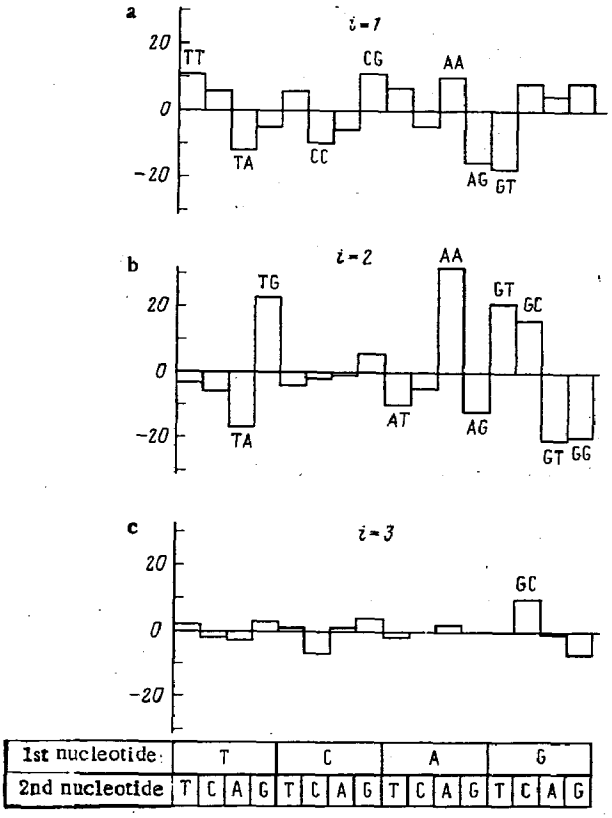

Fig. 2. Shift in positional frequency of dinucleotides: a) first dinucleotide frame; b and c) second and third frames, respectively.

in positions 3 + 3c, c = 1, 2,... form the third. Analogously, nucleotides found in positions 1 + 3c, 2 + 3c, c = 0, 1,... form the first dinucleotide frame, and so on. By the frequency of nucleotide a in the ith mononucleotide frame f_a^i is understood the ratio of the number of nucleotides of type a to the total number of nucleotides (N/3) present in the given frame. The values of f_a^i, i = 1, 2, 3, are presented in Table 3. The possible random error in the determination of these frequencies is estimated by the quantity σ, equal to 0.008. It is easily seen that the obtained frequencies indicate an uneven distribution of nucleotides of various types among frames, which was noted earlier [15, 16]. In Table 3 those quantities that deviate most from the average frequency of a given type of nucleotide in the coding regions are underlined.

These deviations may be associated with various factors. Thus, the decline in the content of T in the first frame and of A in the third is due to the prohibition of the codons TAA, TGA, and TAG. The increase in the content of G in the first frame and of C in the third emphasizes the preferential use of codons of the RNY type (R denote purine and Y, pyrimidine), which was earlier pointed out by Shepherd [17] and is associated with characteristics of the "primeval" genetic code. The relative increase in the content of T in the second frame may be due to the periodic series of synonymous codons coding nonpolar amino acid residues incorporated in the α helices of protein molecules [18]. Finally, the increase in the content of A and the decline in the content of G in the second frame is apparently due to the fact that the 14 codons with A in the second position correspond to seven amino acids, while the 15 codons containing G in the second position code only five different amino acids.

TABLE 4. Conditional Probabilities $P^i(b \mid a)$ for Coding Regions

First nucleotide	$i=1$				$i=2$				$i=3$			
	second nucleotide											
	T	C	A	G	T	C	A	G	T	C	A	G
T	0,386	0,264	0,207	0,143	0,246	0,253	0,100	0,401	0,148	0,228	0,236	0,392
C	0,329	0,167	0,271	0,233	0,240	0,276	0,173	0,311	0,146	0,201	0,257	0,399
A	0,329	0,198	0,378	0,092	0,212	0,260	0,325	0,206	0,128	0,239	0,261	0,367
G	0,199	0,264	0,331	0,207	0,417	0,411	0,051	0,120	0,137	0,296	0,241	0,326

The positionally uneven distribution of nucleotide frequencies in the coding regions of the E. coli genome is inconsistent with the model of a nucleotide sequence in the form of a uniform Markov chain. In fact, such a model implies that the probabilities of appearance of nucleotides of different types are the same at any position taken within the coding region, i.e., they equal P_a, a = T, C, A, and G. However, as we have just seen, the situation is otherwise. Therefore, the original model must be refined using positionally dependent statistics.

We denote the number of dinucleotides of type ab found in the ith dinucleotide frame by $N^i(a, b)$. We shall consider the simplest "positional" statistical model of the nucleotide text of the coding region in which neighboring nucleotides are assumed to be independent, but the probabilities of their appearance differ in different positions relative to the initiating codon. According to this model the expected number of dinucleotides of type ab that should be encountered in the dinucleotide frame with number i will be $N \cdot f^i \cdot f^{i+1}/3$ for i = 1, 2, and $N \cdot f^3 \cdot f^1/3$ for i = 3. Figure 2 presents the quantities $D_{(ab)}i = (N^i(ab) - \bar{N}^i(ab))/(\bar{N}^i(ab))^{1/2}$ for all i. These data permit testing the hypothesis of the possible use of the simplest positional model. If it is valid, the quantities $t^i = \sum_{ab} (D^i_{(ab)})^2$ should have a χ^2_9 distribution. The real values of t^i for i = 1, 2, 3 are 1,504, 3,719, and 246, respectively. This permits the assertion that, once again, neighboring nucleotides with a probability of 0.9999 cannot be considered as independent, i.e., there is a "short-range order" in their arrangement that is inadequately described by reference to positional irregularities in the frequencies of mononucleotides.

We improve the model using the probabilities of transition $P(b \mid a)$, which may now be dependent upon the position numbers occupied by the nucleotides. We then have three matrices $\hat{P}^i(b \mid a)$, the elements of which are determined by the formulas $\hat{P}^i(b \mid a) = N^i(ab)/N^i(a)$; a, b = T, C, A, G; i = 1, 2, 3. The obtained values are presented in Table 4.

It can be assumed that the character of the dependence between nucleotides belonging to different dinucleotide frames differs. Formally, this can be tested by the same method used to establish the differences in the character of the dependence between neighboring nucleotides in coding and noncoding regions of the E. coli genome. We define the quantities t^{ij} and t^{ji} by the equations

$$t^{ij} = \sum_{ij} (N^i(a, b) - N^i(a) \cdot \hat{P}^j(b \mid a))/(N^i(a) \cdot \hat{P}^j(b \mid a))^{1/2}, \quad (2)$$

where i and j are indices of dinucleotide frames. If the character of the dependence of neighboring nucleotides in frames i and j is identical, the quantities t^{ij} and t^{ji} should have a χ^2_{12} distribution. Insofar as t^{12} = 20,265, t^{21} = 9,946, t^{23} = 7,319, t^{32} = 12,230, t^{31} = 11,544, and t^{13} = 9,118, it can be asserted with a probability of no less than 0.9999 that the character of the dependence between neighboring nucleotides differs in different dinucleotide frames.

It should also be noted that the f^i_a quantities satisfy the equations

$$\sum_a f^i_a \cdot \hat{P}^i(b \mid a) = f^{i+1}_b, \quad i = 1, 2, \quad \sum_a f^3_a \cdot \hat{P}^3(b \mid a) = f^1_b.$$

Thus, f^i_a can be considered as final positional probabilities of the states P^i_a in a nonuniform first-order Markov chain with periodically repeating transitional matrices.

A comparison of the correlation parameters found for coding regions $\hat{P}^i(b \mid a)$ with their analogues in noncoding sequences, i.e., with $\hat{P}^N(b \mid a)$, is of interest. To do this we calculate from Eq. (2) the quantities t^{iN} and t^{Ni}, where i is the number of the dinucleotide frame taken in the coding region, and N is an index designating the noncoding region. We obtain t^{1N} = 1,867, t^{N1} = 1,881, t^{2N} = 3,612, t^{N2} = 4,685, t^{3N} = 3,206, and t^{N3} = 3,227. Again, turning to the χ^2_{12} distribution, we conclude that the group of correlation parameters of neighboring nucleotides in noncoding regions does not agree with any of the groups of positional correlational parameters in the coding regions.

The statistical characteristics found for nucleotide sequences from different functional regions of the E. coli genome permit the estimation of such quantities as entropy and redundancy, typically determined for symbolic sequences [3, 19, 20]. If such a sequence consists of independently appearing symbols, i.e., is represented by the simplest model, then according to [20] each of its symbols involves an entropy of

$$H^1 = -\sum_a P_a \cdot \log_2 P_a. \tag{3}$$

Substituting here instead of P_a the f_a quantities we find that $H_N^1 = 1.9978$ and $H_C^1 = 1.9975$ for noncoding and coding regions, respectively.

The presence of a correlation between neighboring nucleotides was accounted for, following [20], by the more complex formula

$$H^2 = -\sum_a P_a \cdot \sum_b P(b|a) \cdot \log P(b|a). \tag{4}$$

Replacing $P(b|a)$ by its estimates $\overset{>}{P}(b|a)$, we find that $H_N^2 = 1.9878$ $H_C^2 = 1.9769$ for coding and noncoding regions, respectively. More precise results for the coding regions can be attained if positional variations in the P_a and $P(b|a)$ quantities are accounted for. It can be shown that to do this Eqs. (3) and (4) should be replaced by the equations

$$H^{P1} = -\frac{1}{3}\sum_i \sum_a P_a^i \cdot \log_2 P_a^i, \tag{5}$$

$$H^{P2} = -\frac{1}{3}\sum_i \sum_a P_a^i \cdot \sum_b P^i(b|a) \cdot \log_2 P^i(b|a), \tag{6}$$

where P^i and $P^i(b|a)$ can be substituted by the known quantities f_a^i and $\overset{>}{P}{}^i(b|a)$. Then $H_N^{P1} = 1.9555$ and $H_C^{P2} = 1.7787$.

The magnitude of redundancy is associated with the magnitude of the entropy of the symbolic sequence by the following relationship [20]:

$$R = 1 - H/\log_2 M, \tag{7}$$

where M is the quantity of possible symbols (i.e., M = 4). Using the previously found values of H, we find that in the noncoding regions $R_N^1 = 1.1 \cdot 10^{-3}$ and $R_N^2 = 6.1 \cdot 10^{-3}$. In the coding regions $R_C^1 = 1.3 \cdot 10^{-3}$, $R_C^2 = 1.2 \cdot 10^{-2}$, and, furthermore, $R_C^{P1} = 2.2 \cdot 10^{-2}$ and $R_C^{P2} = 1.1 \cdot 10^{-1}$.

DISCUSSION

It is apparent from the results obtained that the statistical patterns of the alternation of nucleotides in the coding and noncoding regions of the E. coli genome differ greatly. It is naturally postulated that the character of these patterns reflects a specific path of evolution, selecting different rules of fixation of mutations in different functional regions.

It can be shown formally, using the widespread and rather rigorous steady-state and non-steady-state concepts [21], that a non-steady state of two types is observed in the E. coli genome: "Large-scale" non-steady state, associated with the alternation of coding and noncoding regions, as well as "small-scale" non-steady state, which is characteristic of the coding regions and is expressed in the positional dependence of the probabilities of appearance of nucleotides. This fact is important in defining the structure and parameters of statistical models for describing nucleotide sequences. Several such models were mentioned earlier. The adequate selection of statistical models is important for the development of effective algorithms for computer recognition of the functional zones of DNA. Furthermore, the form of the model influences substantially the result of computation of the probabilities of the random appearance in natural nucleotides of sequences of fragments with various types of symmetry, etc. [22].

It is clear from the foregoing that the structure and parameters of these models cannot be identical for different functional regions. Being unable within the framework of this article to enter into a detailed discussion of questions associated with the classification of models and the conditions of their applicability, we note only the following. Uniform Markov chains of any order [23, 24] lead to steady-state, i.e., positionally independent, distributions of the probability of appearance of nucleotides of a specific type, which does not in fact occur. For this reason uniform chains are of no use as static models of the structural sites of DNA.

The main cause of the development of non-steady-state static patterns in the coding regions of E. coli DNA were probably differences in the functional significance of nucleotides occupying different positions.

Thus, the selection of nucleotides in the first nucleotide frame almost unambiguously determines the amino acid sequence of the protein product, i.e., these dinucleotides bear the primary coding function. At the same time, the second dinucleotide frame includes nucleotides in the third codon position, the selection of which, which is associated with the so-called synonymous-codon selection, is of regulatory significance for the system of cellular biosynthesis [25, 26]. Therefore, the substantial difference between the correlation parameters of the first and second dinucleotide frames is not unexpected (Fig. 2a, 2c; Table 4). We believe that the correlation parameters $P^1(b|a)$, as well as $P^2(b|a)$ can indeed be used as "integral" characteristics for classifying the coding nucleotide sequences.

The presence of an admittedly comparatively weak but statistically significant correlation of the third codon nucleotide with the first nucleotide of the next codon provides the basis to speak of the statistical relationship of neighboring codons. This result is also supported by Lipman and Wilbur [11], who noted that the random rearrangement of codons in the structural region of DNA leads to an increase in entropy, which indicates a decline in the degree of correlation of the symbolic text.

With regard to the values of entropy and redundancy of nucleotide sequences, returning to Eq. (7) we note that the magnitude of $\log_2 M$ equals the entropy for a sequence of symbols that is generated in a series of independent events, in each of which any of M possible symbols appears with equal probability. The value of R for such a sequence is zero. It is known [3] that in linguistic texts (in modern languages) $R \approx 0.7$. This speaks of their substantial "interference resistance," or, in other words, the possibility of restoring a complete text in the presence of randomly distributed gaps.

We note, however, that the small magnitudes of R for noncoding regions cannot be considered a sign of the similarity of their local structure to the structure of a purely random sequence with zero R, insofar as the sample of noncoding regions contained sites with various functional properties, and in light of the foregoing this fact provides the basis to postulate that the primary structure of these regions is not statistically uniform.

The following can be stated concerning the coding regions. If a set of nucleotide fragments of length N taken from the coding regions is examined, it can be asserted [19] that only 2^{HN} different sequences out of the 2^{2N} possible variants of nucleotide sequences will be encountered among these fragments with a probability close to unity. In particular, on the basis of the actual value of H^{P2} we find that only one of the $4.6 \cdot 10^6$ randomly selected nucleotide sequences of 100 nucleotides length has the same statistical structure as a real coding region.

In conclusion it is appropriate to note that the "physical meaning" of the correlation parameters of neighboring nucleotides is not yet completely clear. In particular, Nussinov [5] showed that the magnitude of correlation parameters cannot be related to the values of the intramolecular energies of stacking-interactions in a thermodynamically equilibrium system. Such a fact is logical from the point of view of biology, since the issue is not one of the selection of the most stable sorts of macromolecules in a freely interacting mixture but of the selection of the most rapidly self-reproducing individual complexes of macromolecules (cells), which do not necessarily contain the most stable DNA molecules.

LITERATURE CITED

1. R. Grantham, FEBS Lett., 95, 1-11 (1978).
2. R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pave, Nucleic Acids Res., 8, r49-r61 (1980).
3. V. D. Gusev, V. A. Kulichkov, and T. N. Titkova, in: Computational Systems [in Russian], No. 83, Nauka, Sib. Otd., Novosibirst (1980), pp. 11-31.
4. R. Nussinov, Nucleic Acids Res., 8, 4545-4562 (1980).
5. R. Nussinov, J. Biol. Chem., 256, 8458-8462 (1981).
6. R. Nussinov, J. Mol. Evol., 17, 237-244 (1981).
7. R. Nussinov, J. Mol. Biol., 149, 125-131 (1981).
8. R. Nussinov, J. Mol. Evol., 20, 111-119 (1984).
9. R. Nussinov, Nucleic Acids Res., 12, 1749-1763 (1984).
10. T. F. Smith, M. S. Waterman, and J. R. Sadler, Nucleic Acids Res., 11, 2205-2220 (1983).
11. D. J. Lipman and W. J. Wilbur, J. Mol. Biol., 162, 363-376 (1983).
12. R. F. Doolittle, M. W. Hunkapiller, L. E. Hood, S. G. Devare, K. C. Robbins, S. A. Aaronson, and H. N. Antoniades, Science, 221, 275-277 (1983).
13. S. Wain-Hobson, R. Nussinov, R. J. Brown, and J. L. Sussman, Gene, 13, 355-364 (1981).
14. P. Billingsley, Ann. Math. Stat., 82, 12-40 (1961).

15. J. W. Fikett, Nucleic Acids Res., 10, 5303-5318 (1982).
16. R. Staden, Nucleic Acids Res., 12, 551-567 (1984).
17. J. C. W. Shepherd, Proc. Natl. Acad. Sci. USA, 78, 1596-1600 (1981).
18. V. B. Zhurkin, Nucleic Acids Res., 9, 1963-1971 (1981).
19. K. Shannon, in: Papers on Information Theory and Cybernetics [Russian translation], Izd. Inostr. Lit., Moscow (1963), pp. 243-332.
20. L. L. Gatlin, Informational Theory and Living Systems, Columbia Univ. Press, New York (1975).
21. J. R. Pierce, Symbols, Signals and Noise, Hutchinson, London (1982).
22. W. M. Fitch, Nucleic Acids Res., 11, 4655-4663 (1983).
23. P. W. Garden, J. Theor. Biol., 82, 679-684 (1980).
24. H. Almagor, J. Theor. Biol., 104, 633-645 (1983).
25. T. Ikemura, J. Mol. Biol., 158, 573-597 (1981).
26. M. Gouy and C. Goutier, Nucleic Acids Res., 10, 7055-7074 (1982).

STATISTICAL PATTERNS IN THE PRIMARY STRUCTURES OF FUNCTIONAL REGIONS OF THE GENOME IN *Escherichia coli*

II. NONUNIFORM MARKOV MODELS

M. Yu. Borodovskii, Yu. A. Sprizhitskii, E. I. Golovanov, and A. A. Aleksandrov

UDC 576.315.42

Statistical models of the structural regions of the genome of E. coli in the form of uniform and nonuniform Markov chains are studied. The initial data for their construction were obtained by determining the statistical characteristics of the frequency of di- and trinucleotides in E. coli DNA. The degree of adequacy of models of various type is investigated. On their basis the correlation parameters in sequences of codons and amino acids residues are calculated.

The representation of a biopolymer molecule in the form of a sequence of monomers (nucleotides, amino acids) is naturally considered as a text, i.e., a symbolic sequence of elements of a finite alphabet. We showed [1] that the patterns of alternation of nucleotides in E. coli DNA are nonuniform in different functional regions, which is expressed in a difference in the frequencies of mono- and dinucleotides. Similar facts were established earlier for other DNA molecules [2].

Thus, statistical methods can be used to distinguish functionally significant structures in the text of the genome. Two paths are possible here. The first is to use a statistical description of the genomic text as a whole [3-5]. The simplest model of such type is the representation of the DNA in the form of a sequence of nucleotides appearing with equal probability and independently of one another. If such a model is taken as a foundation, all DNA fragments generated by this model with low probability should be assigned to the category of nonrandom fragments and, presumably, functionally significant. A deficiency of such a method is the need for further clarification of the type of function of the nonrandom fragments found.

The second route, which we shall take in the present study, requires a preliminary investigation resulting in the proposal of a statistical description of a specific class of functional regions. Then, fragments that in some sense satisfy this description are sought in the genomic text. The procedures for computing the statistical characteristics of the model and the subsequent search for functionally significant zones are usually performed with the computer [6, 7]. The foregoing questions are of growing importance with the increase in the degree of automation of the procedure for sequencing nucleic acids. The optimistic view of the prospect for using "local" statistical models is associated with the circumstance that they must distinguish a rather narrow set of all a priori possible nucleotide sequences. For example, it was shown [1] that of $4.6 \cdot 10^6$ randomly

Institute of Molecular Genetics, Academy of Sciences of the USSR, Moscow. Translated from Molekulyarnaya Biologiya, Vol. 20, No. 4, pp. 1024-1033, July-August, 1986. Original article submitted November 27, 1985.

0026-8933/86/2004-0833 $12.50

formed sequences of 100 nucleotides length, on the average only one possessing the statistical "structure" of the coding zone of E. coli DNA is formed.

The present paper investigates the structural regions of E. coli DNA taken from the third edition of the EMBL bank of nucleotide sequences. It was shown that in calculating the codon statistics, nonuniform Markovian models give results that are closer to reality compared with uniform models. Furthermore, a second order nonuniform model provided the basis for a determination of the correlation parameters of neighboring amino acid residues in the primary structure of E. coli protein molecules, the magnitudes of which indicate the presence of a short-range order in polypeptide chains.

METHODS AND RESULTS

Uniform and nonuniform Markov chains of zero, first, and second order are considered as statistical models of the structural regions of the E. coli genome.

A uniform Markov chain of zero order is specified by the magnitudes of the probabilities of its separate states, i.e., in the given case by the $P(a)$ quantities, where a symbolizes nucleotide type T, C, A, or G. (Here and below, unless stipulated otherwise, the letters a, b, c, l, m, and n denote a random nucleotide). A uniform first order Markov chain will be specified by a vector of initial probabilities of states $P^0(a)$ and by a matrix of transitional probabilities $||\, P(b\,|\,a)\,||$. Description of a uniform second order Markov chain requires specification of a vector of initial probabilities $P^0(ab)$ of 16 components, as well as a matrix of transitional probabilities $||\, P(c\,|\,ab)\,||$, which is 4×16 in size.

Earlier [1] we conducted a statistical analysis of the nucleotide texts coding E. coli DNA regions and obtained estimates for the magnitudes of $P(a)$ and $P(b\,|\,a)$ satisfying the conditions of the maximum plausibility method [8]. Insofar as all sequences taken began with an initiating codon AT, the initial distribution of probabilities are of the form

$$P^0(a)=1, \quad a=\mathrm{A}; \quad P^0(a)=0, \quad a\neq \mathrm{A}, \tag{1}$$

$$P^0(ab)=1, \quad ab=\mathrm{AT}; \quad P^0(ab)=0, \quad ab\neq \mathrm{AT}. \tag{2}$$

It can be shown that the uniform first and second order Markov chains we have defined possess ergodic properties, i.e., a final distribution of probabilities $P^j(a)$ exists. To do this it was immediately verified that the $P(a)$ quantities satisfy the systems of equations

$$\sum_a P(b)\,P(b\,|\,a)=P(b), \tag{3}$$

$$\sum_{ab} P(ab)\,P(c\,|\,ab)=P(c). \tag{4}$$

Here $P(ab)=\overline{P}(b\,|\,a)P(a)$, where $\overline{P}(b\,|\,a)=\sum_c P(b|ca)$. Fulfillment of Eqs. (3) and (4) mean that the $P(a)$ quantities are the final probabilities.

The probabilities of any nucleotide combination in the DNA structural regions, including triplet combinations, which are of special interest, can be calculated on the basis of each of the above models. Formally, if one proceeds from the zero order model, then

$$P(abc)=P(a)\,P(b)\,P(c). \tag{5}$$

It follows from the first order model that

$$P(abc)=P(a)\,P(b\,|\,a)\,P(c\,|\,b). \tag{6}$$

From the second order model,

$$P(abc)=P(ab)\,P(c\,|\,ab). \tag{7}$$

We now consider the class of nonuniform Markov models. Their utility is due to the positional unevenness of the frequencies of mono- and dinucleotides in the coding regions of the genome, which was noted earlier [1]. This unevenness is of a periodic character with a period of three nucleotides. A nonuniform zero order Markov chain in the given case will be specified by three vectors $P^i(a)$, i = 1, 2, 3. A vector with number i consists of the magnitudes of the probabilities of appearance of nucleotides in the ith position of the codon, i.e., of a triplet situated in the coding frame.

A nonuniform first order Markov chain is defined by the three vectors of the initial probabilities $P_0^i(a)$,

TABLE 1. Positional Frequency of Dinucleotides in Different Dinucleotide Frames [1]

Dinucleotide	Positional frequency of dinucleotides			Dinucleotide	Positional frequency of dinucleotides		
	first frame	second frame	third frame		first frame	second frame	third frame
TT	0,054	0,071	0,039	AT	0,082	0,066	0,023
TC	0,037	0,073	0,060	AC	0,049	0,081	0,043
TA	0,029	0,029	0,062	AA	0,094	0,101	0,047
TG	0,020	0,116	0,103	AG	0,023	0,064	0,066
CT	0,079	0,054	0,042	GT	0,074	0,073	0,037
CC	0,040	0,062	0,058	GC	0,098	0,072	0,080
CA	0,065	0,039	0,074	GA	0,123	0,009	0,065
CG	0,056	0,070	0,115	GG	0,077	0,021	0,088

$i = 1, 2, 3$, a = T, C, A, G, which correspond to the three $P^i(a)$ vectors just mentioned above. Furthermore, three matrices of transitional probabilities $||\, P^i(b \mid a)\,||$, $i = 1, 2, 3$, the elements of which are taken from Table 4 of the preceding paper [1], are required.

The description of a nonuniform second order Markov chain requires specification of three vectors of initial probabilities of 16 components, $P_0^i(ab)$, $i = 1, 2, 3$ (Table 1), as well as three matrices of transitional probabilities of 4×16 size $||\, P^i(c \mid ab)\,||$, $i = 1, 2, 3$. The elements of these matrices are presented in respectively the first, second, and third part of Table 2.

Direct verification shows that the values of the estimates of maximum probability for the elements of the vectors and matrices $P^i(a)$, $P^i(b \mid a)$, and $P^i(c \mid ab)$, $i = 1, 2, 3$ satisfy the equations

$$\sum_a P^i(a)\,P^i(b \mid a) = P^{i+1}(b), \quad i = 1, 2. \tag{8}$$

$$\sum_a P^3(a)\,P^3(b \mid a) = P^1(b),$$

$$\sum_{cb} P^i(ab)\,P^i(c \mid ab) = P^{i-1}(c), \quad i = 2, 3, \tag{9}$$

$$\sum_{ab} P^1(ab)\,P^1(c \mid ab) = P^3(c).$$

These equations are analogous to Eqs. (3) and (4). Therefore, the quantities $P^i(a)$ can be interpreted as positionally dependent final probabilities of individual nucleotides.

Representation of the text of coding regions in the form of nonuniform Markov chains leads to nonuniform values of the probabilities of abc triplets for different frames. In particular, we obtain the following formulas for the coding frame.

In the case of a zero order model:

$$P^1(abc) = P^1(a)\,P^2(b)\,P^3(c). \tag{10}$$

In the case of a first order model:

$$P^1(abc) = P^1(a)\,P^1(b \mid a)\,P^2(c \mid b). \tag{11}$$

In the case of a second order model:

$$P^1(abc) = P^1(ab)\,P^1(c \mid ab). \tag{12}$$

A cyclic rearrangement of the superscripts in Eqs. (10)-(12) provides equations for calculating the probabilities of abc triplets in the other two frames, which we denote by $P^2(abc)$ and $P^3(abc)$.

Equations (5)-(7) and (10)-(12) permit the calculation of the expected frequencies of codons belonging to 20 different synonymous groups in all three read frames. Comparison of the calculated values obtained on the basis of zero and first order models with experimental data leads to the results shown in Figs. 1-3 for the first, second, and third frames, respectively. Here the magnitudes of the differences between actual and calculated frequency values are depicted graphically. The units of measurement are the quantities of mean squared deviations, which are calculated from the model distributions of the probabilities of codons.

TABLE 2. Transitional Probabilities $P^i(c \mid ab)$, i = 1, 2, 3; a, b, c = T, C, A, G, for Nonuniform Second-Order Markov Chain

Dinucleotide	First frame				Second frame				Third frame			
	T	C	A	G	T	C	A	G	T	C	A	G
TT	0,272	0,388	0,158	0,367	0,154	0,183	0,239	0,423	0,350	0,317	0,243	0,090
TC	0,341	0,337	0,148	0,175	0,150	0,192	0,274	0,384	0,334	0,161	0,285	0,220
TA	0,449	0,551	0,000	0,000	0,172	0,276	0,276	0,276	0,369	0,167	0,405	0,059
TG	0,244	0,255	0,000	0,501	0,121	0,257	0,257	0,371	0,161	0,275	0,361	0,203
CT	0,113	0,106	0,033	0,748	0,148	0,204	0,167	0,463	0,326	0,247	0,212	0,215
CC	0,132	0,095	0,181	0,592	0,145	0,161	0,290	0,403	0,288	0,178	0,276	0,258
CA	0,135	0,189	0,197	0,478	0,154	0,256	0,231	0,385	0,290	0,204	0,360	0,146
CG	0,512	0,392	0,042	0,052	0,129	0,329	0,229	0,314	0,207	0,226	0,337	0,230
AT	0,268	0,386	0,035	0,312	0,121	0,273	0,242	0,348	0,411	0,275	0,190	0,124
AC	0,241	0,480	0,097	0,183	0,148	0,222	0,247	0,383	0,339	0,162	0,256	0,243
AA	0,141	0,292	0,427	0,140	0,099	0,227	0,267	0,406	0,289	0,252	0,373	0,086
AG	0,231	0,659	0,075	0,036	0,172	0,328	0,219	0,266	0,182	0,278	0,334	0,206
GT	0,342	0,164	0,205	0,288	0,151	0,247	0,260	0,342	0,468	0,234	0,175	0,123
GC	0,243	0,226	0,222	0,309	0,139	0,208	0,222	0,431	0,354	0,169	0,262	0,215
GA	0,248	0,208	0,386	0,158	0,222	0,222	0,333	0,333	0,343	0,189	0,395	0,073
GG	0,451	0,387	0,067	0,095	0,143	0,286	0,238	0,285	0,242	0,288	0,288	0,182

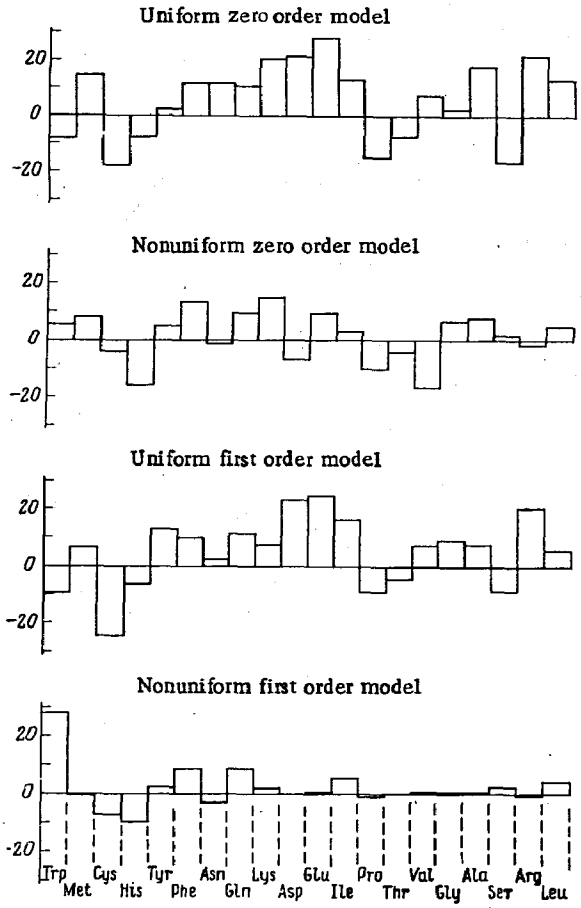

Fig. 1. Comparison of predicted and experimental frequencies of codons of twenty amino acid synonymous groups for different Markov models; first read frame (see text)

Evidently, all four statistical models give the best agreement for the third read frame (Fig. 3). This is due to the rather weak correlation of the third nucleotide codon with the first nucleotide of the next codon [1], which in the present case leads to a nearly equally probable distribution of the triplets of the third frame with respect to synonymous groups. Furthermore, all three figures show that the nonuniform first order model gives the best approximation for actual frequencies. We note that the degree of this agreement is less in the first read frame than in the second and third, which indicates the presence in the first read frame of additional features in the alternation of nucleotides that are not fully accounted for by a nonuniform first order Markov model.

As a whole, the data presented indicate the preferability of nonuniform models for the statistical description of the primary structure of the coding regions of E. coli DNA. This fact seems entirely natural insofar as uniform models, as apparent from Eqs. (5)-(7), do not reflect the dependence of the probability of a specific triplet upon the position occupied relative to the coding frame. In particular, they predict an identical probability of appearance of the terminating codons in all read frames.

The nonuniform second order model gives a precise agreement between the calculated values of frequencies of the groups of codons under examination with the actual codons; therefore, we present no figure for this model.

We shall examine the possibility of using the nonuniform Markov models presented above for describing texts consisting of the nucleotide triplets abc. We denote by $P(abc \to lmn)$ the probability of appearance of the lmn triplet after the abc triplet. According to the zero order model

$$P^0(abc \to lmn) = P^1(l)P^2(m)P^3(n) = P^0(lmn), \tag{13}$$

i.e., a dependence upon the specific form of the preceeding triplet abc is excluded. This means that if a nu-

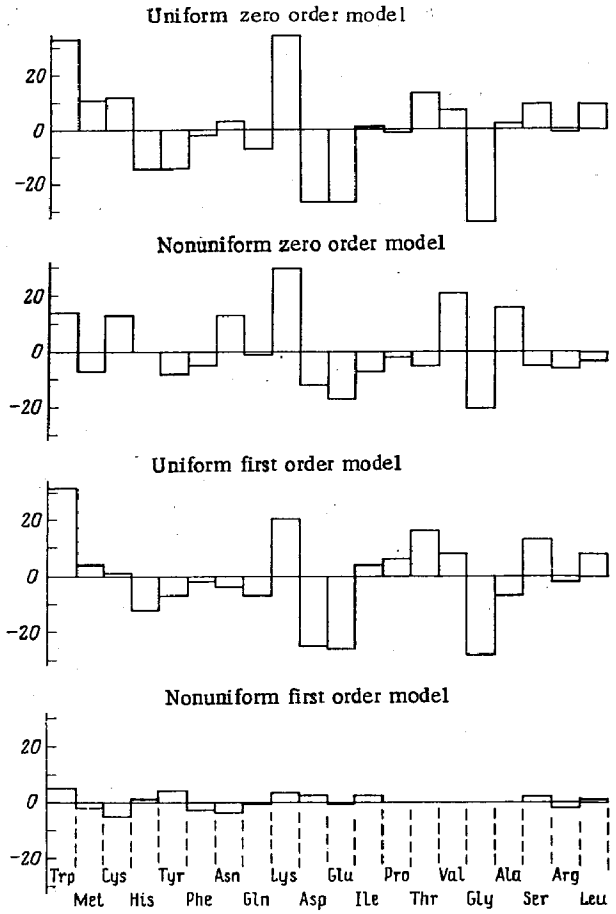

Fig. 2. Deviations in prediction of frequencies in second codon read frame.

cleotide sequence is represented by a nonuniform zero order model, the same sequence broken up into triplets will be described by a uniform zero order model that corresponds to the vector of probabilities $P^0(lmn)$.

On the basis of the first and second order models we obtain

$$P^1(abc \to lmn) = P^3(l|c)\,P^1(m|l)\,P^2(n|m), \tag{14}$$

$$P^2(abc \to lmn) = P^2(l|bc)\,P^3(m/|cl)\,P^1(n|lm). \tag{15}$$

The probabilities of the transition in Eqs. (14) and (15) depend upon the type of nucleotides c and bc. Thus, in this case the triplet sequences are represented in the form of uniform Markov first order chains. The elements of the corresponding matrices of transitional probabilities are determined from Eqs. (14) and (15).

We note that the quantities $P^1(abc \to lmn)$ for triplets taken in the coding frame can be considered as approximate estimates of the elements of a real matrix of transitional probabilities "codon → codon" of 64×64 size. The precise determination of the elements of this matrix requires a volume of information on the primary structures of E. coli nucleic acids that is presently unavailable to us.

A statistical description of the primary structure of E. coli protein molecules and, in particular, an estimate of the correlation parameters for neighboring amino acid residues can be obtained on the basis of the $P^1(abc \to lmn)$ quantities.

If $s_1, s_2, \ldots, s_k$ are the codon numbers from a synonymous group A_α of amino acid α, while $j_1, j_2, \ldots, j_r$ are the codon numbers from the synonymous group A_β of amino acid β, the probability of appearance of the pair $\alpha\beta$

$$P(\alpha\beta) = \sum_{s \in A_\alpha} \sum_{j \in A_\beta} P_s P(s \to j),$$

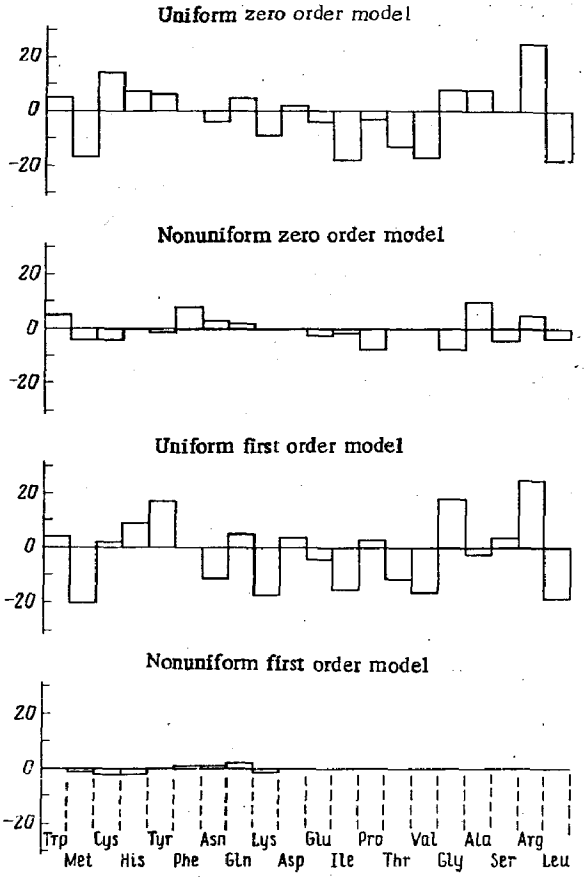

Fig. 3. Deviations in prediction of frequencies in third codon read frame.

where P_s is the probability of appearance of a codon with number s, determined by Eq. (7).

If it is assumed that the amino acid residues α and β appear in neighboring positions of the polypeptide chain independently of one another, the relation must be met:

$$P(\alpha\beta)=P(\alpha)P(\beta),$$

where $P(\alpha)=\sum_{s\in A_\alpha} P_s$, $P(\beta)=\sum_{j\in A_\beta} P_j$. In order to test the hypothesis of independence, we compute the quantities

$$t_{\alpha\beta}=N^{1/2}\frac{P(\alpha\beta)-P(\alpha)P(\beta)}{(P(\alpha)P(\beta))^{1/2}},$$

where N is the total number of codons in the examined set of nucleotide sequences.

The sum $S=\sum_{\alpha\beta} t^2_{\alpha\beta}$ for the case of the coding frame must have a χ^2 distribution with 361 degrees of freedom. According to the Fisher theorem [8], if the number of degrees of freedom n is great, the quantity $\xi=(2S)^{1/2}$ has a normal distribution with an average value $(2n-1)^{1/2}$ and a unit variance. When n = 361 the quantity $\langle\xi\rangle$ is 26.9. The actual value of ξ, calculated for the first frame, is 20.7. This result provides the basis for rejecting the hypothesis of the statistical independence of neighboring amino acid residues in the primary structures of E. coli proteins.

A similar test can also be performed for hypothetical amino acid chains corresponding to translation by the second and third read frames. In this case the quantity S has 400 degrees of freedom. The expected value of ξ is 28.2, while the actual is 47.3 and 86.3 for the second and third frames, respectively. Thus, the hypothesis of independence is also rejected.

We note that if the magnitude of deviation of ξ from the expected value $\langle\xi\rangle$ is considered a measure of the statistical dependence of neighboring amino acid residues, it is easily seen that this dependence is expressed to the weakest extent in real protein structures (first frame).

DISCUSSION

The obtained results may be of direct significance for answering the question of the applicability of models of the Markov chain type for the statistical description of real nucleotide texts. It was shown that uniform Markov models, which were also examined earlier [9-11], do not permit a satisfactory description of the nucleotide sequences of DNA coding regions. The cause of this lies in the positional unevenness of the nucleotide frequencies, which cannot be accommodated within the framework of uniform models.

A series of statistical characteristics of the primary structure of E. coli protein molecules was calculated for the structural region of DNA on the basis of the nonuniform second order Markov model. Comparison of the frequencies of possible pairs $\alpha\beta$ of neighboring amino acids, $\nu_{\alpha\beta}$, calculated on a nonuniform second order model with the $\overline{\nu}_{\alpha\beta}$ quantities obtained on the basis of the zero order model shows that the $\nu_{\alpha\beta}$ quantities for the majority of possible pairs of neighboring amino acids $\alpha\beta$ do not exceed the limits of $\overline{\nu}_{\alpha\beta} + 3\sigma$, where σ is the mean squared deviation of the quantity $\nu - \overline{\nu}$, determined by the zero order model. At the same time, the inequality $\nu_{\alpha\beta} > \nu_{\alpha\beta} + 3\sigma$ is met for the pairs Gln–Leu, Gln–Phe, Pro–Leu, Glu–Thr, and Lys–Thr, while the inequality $\nu_{\alpha\beta} < \overline{\nu}_{\alpha\beta} - 3\sigma$ is met for the pairs Gln–Glu, Gln–Gly, and Phe–Leu.

The χ^2 value calculated from the entire set of normalized values of the obtained deviations provides the basis to reject the hypothesis of the independence of neighboring amino acids with a level of significance of no less than 0.999. This result, obtained on the model, essentially coincides with the data obtained by Poroikov et al. [12] in an investigation of the correlation of neighboring amino acids on a sample of proteins from different families and organisms. However, these authors after an examination of more long-range correlations of amino acid pairs came to the conclusion that the primary structure of globular proteins is of a purely statistical character. Our results do not provide the basis for a similar judgement with respect to the primary structure of E. coli proteins. In the best case a "short-range" order exists here, and a "long-range" order is absent.

It is interesting to note that a "long-range" order exists in the nucleotide sequences of the structural regions of DNA and is expressed in the presence of positional frequencies of mono- and dinucleotides. In this connection it can be postulated that insofar as a deletion or insertion leading to a shift in the read frame causes a change in the parameters of long-range order characteristic of genes of actively functioning proteins, it is unlikely that a mutant polypeptide chain will give a selectively acceptable protein.

The question of the search for a statistical model to describe the primary structure of DNA now presents itself differently. It follows from the obtained results that there is no single model describing equally well both the coding and the noncoding regions [1]. The DNA text is non-steady state, and its different functional zones are described by different models. The global averaging of the frequencies of mono- and dinucleotides throughout the genome, for example on the basis of Nussinov's results [13], provides only an orientational, zero or second order uniform model. Such a model can be used to distinguish a broad set of nonrandom regions of the DNA text requiring further differentiation.

On the other hand, models of functional regions can be used to develop effective algorithms for distinguishing the corresponding DNA functional zones. In particular, this relates to the computer recognition of DNA coding regions in prokaryotes and eukaryotes using nonuniform Markov models, and this question merits special examination.

LITERATURE CITED

1. M. Yu. Borodovskii, Yu. A. Sprizhitskii, E. I. Golovanov, and A. A. Aleksandrov, Mol. Biol., 20, No. 4, pp. 1014-1023 (1986).
2. T. F. Smith, M. S. Waterman, and J. R. Sadler, Nucleic Acids Res., 11, 2205-2220 (1980).
3. W. M. Fitch, Nucleic Acids Res., 11, 4655-4663 (1983).
4. N. A. Kolchanov, V. V. Solov'ev, and A. A. Zharkikh, Dokl. Akad. Nauk SSSR, 273, 741-744 (1983).
5. E. V. Korotkov and M. A. Korotkova, Dokl. Akad. Nauk SSSR, 274, 748-750 (1984).
6. J. W. Ficett, Nucleic Acids Res., 10, 5303-5318 (1982).
7. R. Staden, Nucleic Acids Res., 12, 551-567 (1984).
8. H. Cramer, Mathematical Methods of Statistics, Princeton Univ. Press (1946).
9. P. W. Garden, J. Theor. Biol., 104, 633-645 (1980).
10. C. Fuch, Gene, 10, 371-373 (1980).
11. H. Almagor, J. Theor. Biol., 104, 633-645 (1983).
12. V. V. Poroikov, N. G. Esipova, and V. G. Tumanyan, Mol. Biol., 18, 541-546 (1984).
13. B. Nussinov, Nucleic Acids Res., 12, 1749-1763 (1984).

A POSSIBLE ENZYMIC ROLE OF IMPORTED $tRNA_1^{Lys}$ DURING THE SPLICING OF YEAST MITOCHONDRIAL TRANSCRIPTS

T. R. Soidla

UDC 577.217.33:576.311.347

A sequence of $tRNA_1^{Lys}$ imported into the mitochondria was complementary to first class intron donor boundaries and corresponds to the generalized ribozyme structure.

It has been shown that $tRNA_1^{Lys}$ is imported into the mitochondria of Saccharomyces cerevisiae but does not undergo aminoacylation in these organelles [1]. It has been found that $tRNA_1^{Lys}$ is complementary to a number of mitochondrial transcript sites which are of importance for splicing [2, 3]. In this case the role of this tRNA in the splicing process was not clear. On the basis of preliminary comparisons it was proposed that $tRNA_1^{Lys}$ is a ribozyme which participates in the cleavage of the primary transcript at the site of the intron donor boundaries. In this article the author discusses certain new reports in the literature which consider the role of $tRNA_1^{Lys}$ in splicing and demonstrates that the structure of this tRNA does in fact correspond to the typical ribozyme structure.

RESULTS AND DISCUSSION

It has previously been demonstrated that pre-mRNA sites, which are of importance for splicing, and $tRNA_1^{Lys}$ are complementary to each other [2, 3]. It was considered that the site of this tRNA, complementary to the first class mitochondrial intron donor boundaries, most probably participates in the splicing process. The presence of a complex and extremely conservative secondary structure is a common feature of first class introns [4]. They are encountered in the mitochondrial genome of fungi and also in the nuclear genome of protozoa and in the chloroplast genome of plants [5]. Certain introns of this class have the ability to achieve self-splicing in vitro [6-9], but in vivo, at least in the case of mitochondrial introns, splicing is dependent on factors which are encoded in the nuclear genome [7-11]. On account of the complementarity between $tRNA_1^{Lys}$ and the intron-containing mitochondrial transcripts, it was possible to combine data on the apparently heterogeneous boundaries of first class introns into a consistent system. In fact in the donor (5') boundaries of these introns only the last nucleotide of the exon was conservative, i.e., the boundary has a U ↓ X form [12]. However, after partial untwisting of $tRNA_1^{Lys}$ in more or less the same position in which the reverse transcriptase untwists the tRNA-primers [13] complementarity was found between the donor sites of all nine first class yeast mitochondrial introns and the untwisted portion of the $tRNA_1^{Lys}$. The data from the two previous reports of the author are presented in Fig. 1. On account of the complementarity revealed it was possible to find point donor boundaries in accordance with the following rule: complementarity, transition, the first m^2GU, boundary, complementarity are always found in the same order in the 3' → 5'-direction in the vicinity of the donor boundary. In cases in which the donor boundaries were subsequently more accurately defined they coincided with the boundaries predicted by the authors [9, 14]. However it seemed unlikely that a reverse transcriptase type of enzyme with the ability to cleave tRNA existed in the mitochondria. There are now two sets of data which support the hypothesis of the current author. For example it has been shown that introns in the mitochondrial genes may segregate at the level of DNA and in this case this phenomenon occurs frequently and involves the segregation of several introns concomitantly [15]. This phenomenon may be attributed to the reverse transcription of various pre-mRNAs of the gene with subsequent conversion between the gene and the DNA copy of one of the precursors of the mature mRNA. In addition a mitochondrial plasmid [16] in the ascomycete Neurospora crassa has recently been described which contains conservative nucleotide sequences typical of first class introns, and at the same time the plasmid has the dinucleotides UG and CA, which are characteristic of retroviruses, at the ends of its RNA-copy (the plasmid is totally transcribed). In addition the protein, which is encoded by this plasmid, has a site of weak homology with the re-

All-Union Scientific-Research Institute of the Hydrolysis of Plant Materials, Leningrad. Translated from Molekulyarnaya Biologiya, Vol. 20, No. 4, pp. 1034-1038, July-August, 1986. Original article submitted October 1, 1985.

0026-8933/86/2004-0841 $12.50

STATISTICAL PATTERNS IN THE PRIMARY STRUCTURES OF FUNCTIONAL REGIONS IN THE GENOME OF *Escherichia coli*.

III. COMPUTER RECOGNITION OF CODING REGIONS

M. Yu. Borodovskii, Yu. A. Sprizhitskii, E. I. Golovanov, and A. A. Aleksandrov

UDC 576.315.42

A method is proposed for the rapid analysis of nucleotide sequences that enables determination of the structural regions of DNA, establishment of the read frame of the triplet codon, and also estimation of the potential degree of expression of the protein product. The idea behind the method originates from earlier-obtained resulted on the difference in the statistical properties of coding and noncoding *E. coli* DNA fragments. This difference is, in the final analysis, manifested in the character of Markov models for the coding and noncoding regions of DNA, which are used in the recognition algorithm. The results are presented of calculations made on an Iskra-226 personal computer.

At the present time, when effective methods for deciphering DNA primary structures are widely available [1, 2], problems of the functional interpretation of the nucleotide texts that are read should be given high priority. The results already obtained indicate that the situation is quite complex. As a rule, a single function is performed by different combinations of nucleotides in DNA primary structures. The groups of synonymous codons coding the same amino acid, promotor regions, translation initiation sites, and splicing sites can be mentioned as the simplest examples [3-5].

However, the appearance of a permissible nucleotide combination in the DNA primary structure still does not mean that the corresponding function is actually realized. This requires a fully defined "context." In particular, a nucleotide triplet will act as a codon only if it is found in a coding region and is situated in a specific phase (coding frame) with respect to the initiating codon. Thus, it can be assumed that methods considering the "context" and analyzing extended DNA regions will have a special role in the interpretation of primary structures [6]. Included in this category are methods for recognizing DNA coding regions proposed earlier [7-12] and oriented towards the use of a computer. They are based on the statistical characteristics of the frequency of mono- and trinucleotides. In some cases, they are calculated in advance with a sample of previously known DNA coding sites (training sample). It should be emphasized that by coding site (fragment) or gene we refer to the site of single-stranded DNA bearing information on the actual functional polypeptide. The contextual features of the genes are associated with the amino acid structure of the coded protein, as well as with the mechanism of translation [10].

Without going into a comparative analysis of earlier-developed [8-12] approaches, we note the following. The method of Staden and McLachlan [8] was used in the study of Sanger et al. [13] to determine the precise dimensions of the DNA coding regions in λ phage. The method of Fickett [9] is suitable for answering the question as to whether some specific DNA fragment is a functioning gene (or part thereof). It was used in the study of Tramontano et al. [14] to analyze the "open read frames" of the genetic code, which were found in the genomes of various organisms on DNA strands complementary to coding regions. Gribskov et al. [11] indicate that the forementioned methods [8, 10, 11] can be used to determine the correct read frame, to estimate the potential degree of gene expression, and also to indicate errors in deciphering the primary structure of DNA that are of the form of the insertion or deletion of some number of nucleotides (not a multiple of three).

Institute of Molecular Genetics, Academy of Sciences of the USSR, Moscow. Translated from Molekulyarnaya Biologiya, Vol. 20, No. 5, pp. 1390-1398, September-October, 1986. Original article submitted January 30, 1986.

 0026-8933/86/2005-1144$12.50

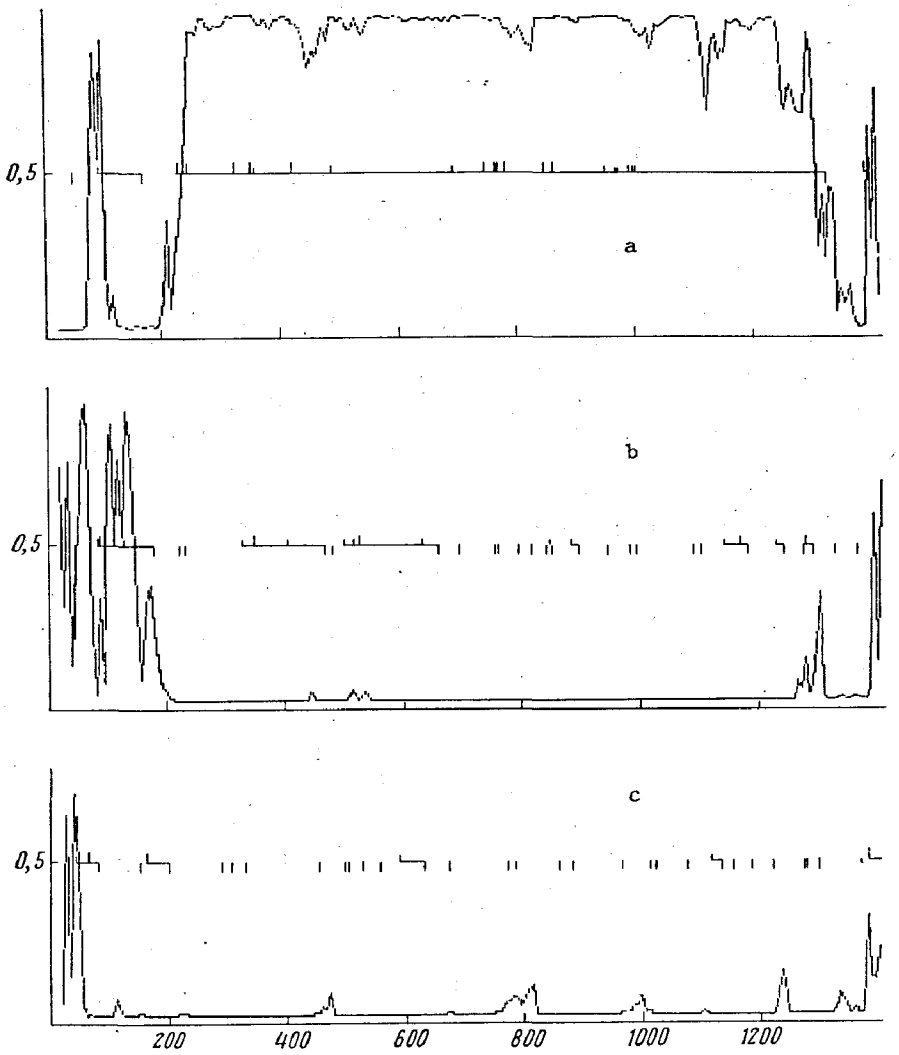

Fig. 1. Plots of indicator functions of coding regions computed for three read frames (a-c) of ECRECA sequence. V = (1, 16).

Our statistical analysis [15] of the coding and noncoding regions of the primary structure of *E. coli* DNA opened new possibilities for development of a method for recognizing coding regions that proved to be the missing link in several approaches associated with the statistics of mononucleotides [9, 10] and trinucleotides [8, 10, 11]. Furthermore, the route we take — representation of DNA functional zones by uniform and nonuniform Markov chains [16] — makes it possible to give a logical exposition of the theory behind such methods.

DESCRIPTION OF METHOD

The problem of recognizing the coding regions in *E. coli* (and other prokaryotes) using purely computer methods amounts to a "marking" of the DNA into alternating zones, some of which are "closest" to the image of a coding region present in the computer memory, and others, to the image of a noncoding region. By images here are understood certain mathematical models, the parameters of which were obtained by means of a statistical analysis of the corresponding samples of nucleotide sequences.

We shall move directly to the algorithm. We consider the nucleotide fragment (a_1, a_2, ... a_n), subsequently abbreviated as α. It is convenient to take n as a multiple of three. We designate by $P(K|\alpha)$ the probability that if a site identical to α is found in the DNA sequence, this site will belong to a coding region, and by $P(N|\alpha)$, the probability that this site will belong to a noncoding region. The quantity $P(K|\alpha)$ is made up of three quantities $P(K_1|\alpha)$, $P(K_2|\alpha)$, and $P(K_3|\alpha)$. $P(K_i|\alpha)$ is the probability that α will belong to a coding region, and at the same time nucleotide a_i occupies the ith position in some codon. To calculate the probabilities $P(N|\alpha)$ and $P(K_i|\alpha)$, i = 1, 2, 3, we must know the parameters of the mathematical models of the coding and noncoding regions.

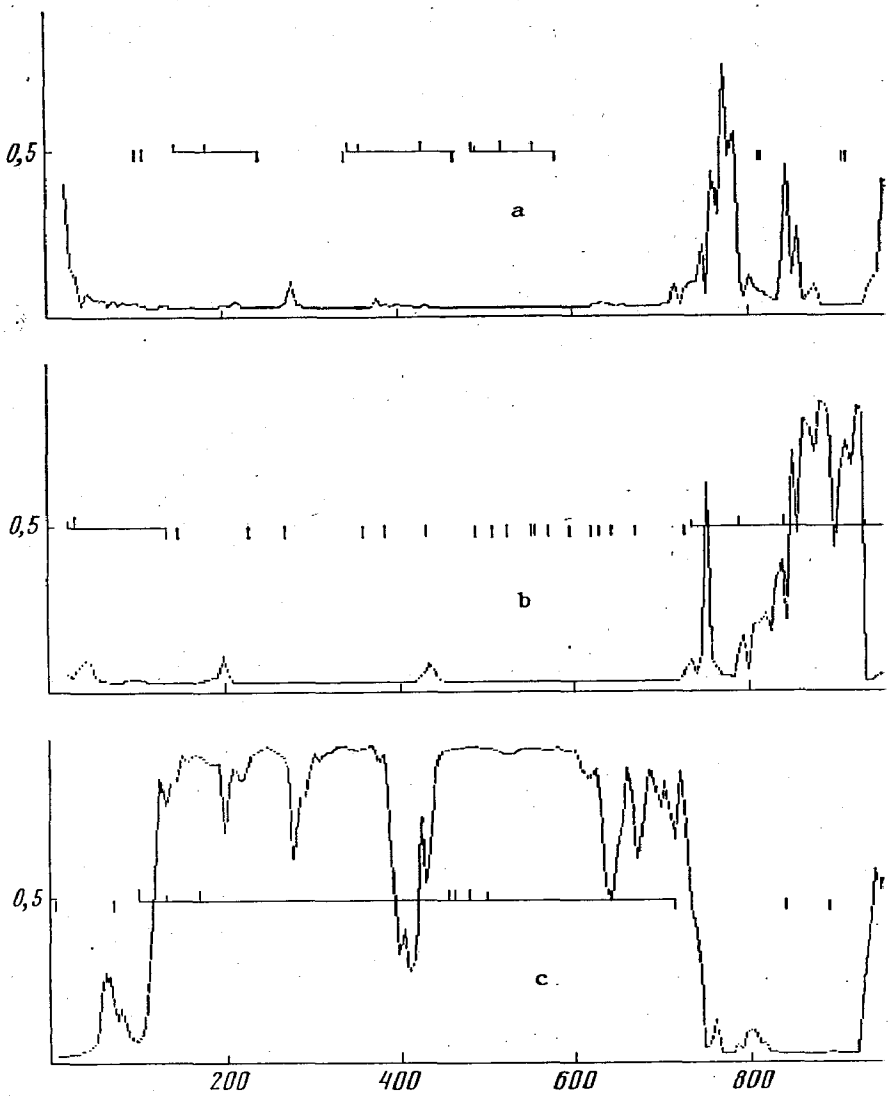

Fig. 2. Plots of indicator functions of coding regions calculated for three read frames (a-c) of ECLEXX sequence. V = (1, 16).

The model of noncoding region is specified by a uniform first order Markov chain. The vector of the initial distribution of probabilities (according to [15], Table 1) has four components: $P_0(T) = 0.231$, $P_0(C) = 0.259$, $P_0(A) = 0.261$, $P_0(G) = 0.248$. The matrix of transitional probabilities for this chain was presented earlier (Table 2b [15]). Nonuniform Markov chains of three different orders, r = 1, 2, 3, can be used as a model of the coding region. The greater r the closer the statistical characteristics of the model and object [16]. However, this must be "paid" for by the introduction of additional parameters and an increase in the time of the calculations. Therefore, we will not dwell on any one (better) model but will present below the results for all three.

The model type is not important for clarifying the nature of the algorithm. We shall take for definition a nonuniform first order Markov chain. Such a chain is specified by three vectors of the initial probabilities $P_0^i(a)$, a = T, C, A, G, and by three matrixes of transitional probabilities of sizes $4 \times 4 \| P^i(b|a) \|$, a, b = T, C, A, G, i = 1, 2, 3. The numerical values of the component vectors and elements of the matrixes are presented in our previous paper and in Tables 3 and 4, respectively.

The first step in the algorithm is the calculation of four accessory statistical quantities for the α fragment. One of them, $P(\alpha|N)$, determines the probability of randomly finding a fragment identical to α in a noncoding region. According to an equation known from the theory of Markov chains [17]:

$$P(\alpha|N) = P_0(a_1)P(a_2|a_1)\cdot \ldots \cdot P(a_n|a_{n-1}). \quad (1)$$

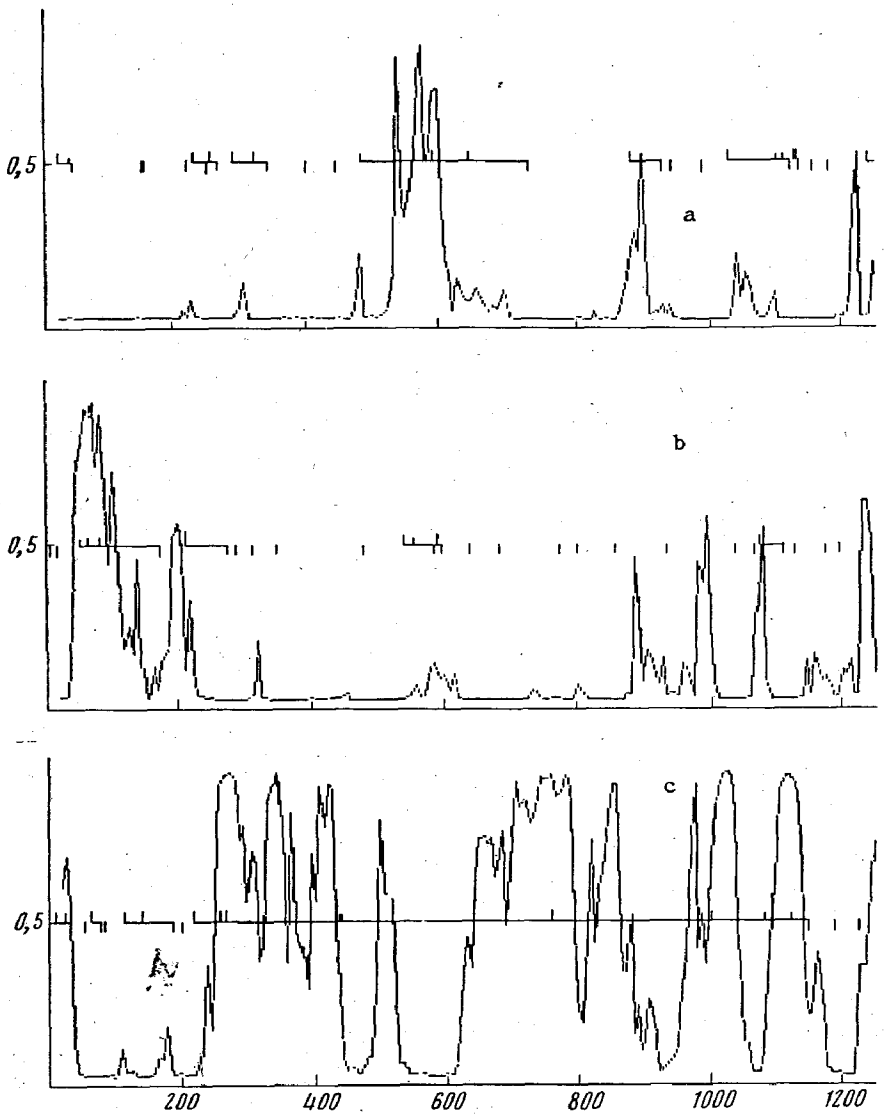

Fig. 3. Plots of indicator functions of coding regions calculated for three read frames (a-c) of ECARAC sequence. V = (1, 16).

The other three quantities we denote $P(\alpha|K_1)$, $P(\alpha|K_2)$, and $P(\alpha|K_3)$. $P(\alpha|K_i)$ is the probability of randomly finding fragment α in a coding region and in such position that nucleotide a_1 is in the ith position of some codon. We have

$$P(\alpha|K_1) = P_0^1(a_1)P^1(a_2|a_1)P^2(a_3|a_2)\cdot\ldots\cdot P^2(a_n|a_{n-1}), \quad (2)$$

$$P(\alpha|K_2) = P_0^2(a_1)P^2(a_2|a_1)P^3(a_3|a_2)\cdot\ldots\cdot P^3(a_n|a_{n-1}), \quad (3)$$

$$P(\alpha|K_3) = P_0^3(a_1)P^3(a_2|a_1)P^1(a_3|a_2)\cdot\ldots\cdot P^1(a_n|a_{n-1}). \quad (4)$$

The probabilities $P(N|\alpha)$ and $P(K_i|\alpha)$ can now be calculated. Moreover, it will be sufficient to determine the quantities $P(K_i|\alpha)$, since it can be assumed with a high degree of precision that $P(N|\alpha) = 1 - P(K_1|\alpha) - P(K_2|\alpha) - P(K_3|\alpha)$. On the basis of Bayes' formula [17], we find

$$P(K_i|\alpha) = \frac{P(\alpha|K_i)P(K_i)}{\sum_i P(\alpha|K_i)P(K_i) + P(\alpha|N)P(N)}, \quad i = 1,2,3. \quad (5)$$

Here $P(N)$ and $P(K_i)$ are the so-called a priori probabilities of events N and K_i. They characterize the probability of membership of some fragment α to a noncoding or coding region (with one of the positions of the first nucleotide fragment α specified above). In this case, the specific primary structure of α is disregarded. It is naturally assumed that $P(N) = 1/2$, $P(K_i) = 1/6$, $i = 1, 2, 3$.

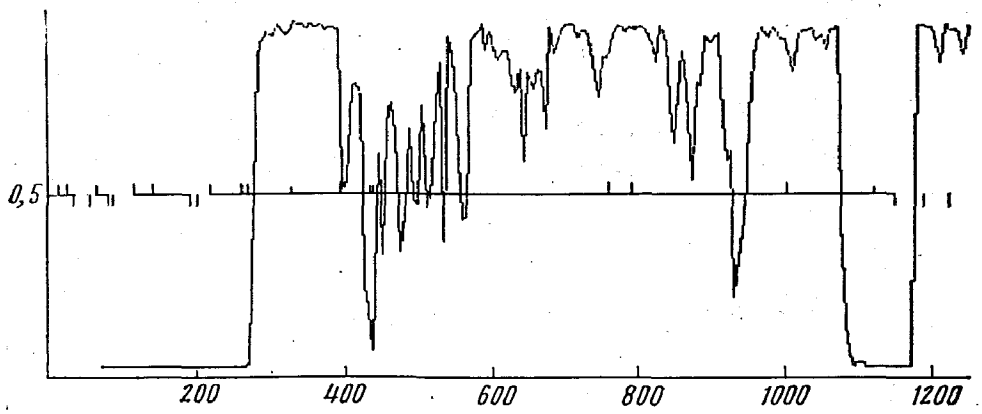

Fig. 4. Plot obtained for the same conditions as in Fig. 3c but using algorithm. V = (2, 48).

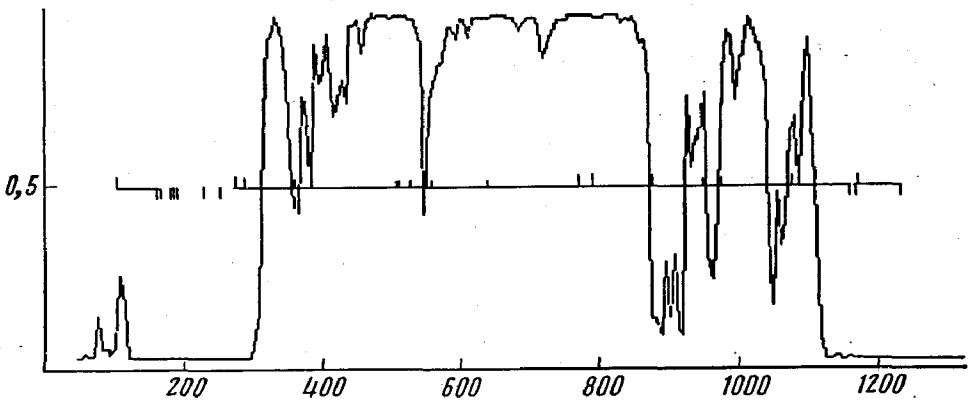

Fig. 5. Plot of coding-region indicator function for sequence ECRRNBZ, containing "open frame." V = (2, 32).

The quantities $P(K_i|\alpha)$ and $P(N|\alpha)$ are determined in a similar manner in those instances where a specific region is modeled by a nonuniform Markov chain of zero or second order. In particular, for r = 0 the product of positional probabilities of nucleotide encounter appears in Eqs. (1)-(4) (see [15], Table 3). For the case of r = 2 the data are taken from Tables 1 and 2, published earlier [16].

Thus, the formal aspect of the method reduces to the computation by Eqs. (1-5) of the values of $P(K_i|\alpha)$, i = 1, 2, 3 for some multitude of fragments covering the investigated DNA sequence.

RESULTS AND DISCUSSION

The foregoing algorithm was realized on an Iskra-226 personal computer. The program permitted the use of any of the three Markov models of the coding region. The parameter l — the size of the "window" for scanning the sequence — could be taken equal to 16, 32, or 48 codons. The "window" was successively shifted by two codons, and probabilities $P(K_i|\alpha)$, i = 1, 2, 3 were calculated at each position for the fragment α landing in the opening of the "window." These values were arranged in accordance with the center of the obtained fragment. Subsequently, for the abbreviated designation of the computational variant we shall use the notation V(r, l), indicating the order of the model of the coding region and the width of the "window."

In order to clarify the selection of objects for analysis, we note the following. It was shown [18] that, on one hand, the patterns of the nonrandom use of synonymous codons in bacterial genes may be associated with the degree of their expression in the cell. On the other hand, the selection of codons also affects the statistical characteristics of the nucleotide sequences of coding regions [11]. Therefore, it is of interest to obtain and compare the results of genes possessing according to [11, 18] substantial differences in the rules of selection of synonymous codons.

TABLE 1. Values of $\bar{p}$ and s_1 Obtained for Different Genes Using Different Variants of Algorithm

Order of Markov model	Parameter	*araC*			*lexA*			*recA*		
		Number of codons in "window"								
		16	32	48	16	32	48	16	32	48
0	$\bar{p}$	0,52	0,64	0,70	0,76	0,94	0,95	0,79	0,93	0,96
	s	0,31	0,33	0,33	0,22	0,08	0,13	0,20	0,15	0,15
1	$\bar{p}$	0,51	0,53	0,51	0,86	0,96	0,97	0,95	0,99	0,99
	s	0,36	0,40	0,42	0,21	0,13	0,12	0,09	0,07	0,09
2	$\bar{p}$	0,61	0,69	0,76	0,90	0,94	0,90	0,98	0,96	0,97
	s	0,34	0,33	0,30	0,22	0,22	0,28	0,14	0,18	0,16

For this purpose, three nucleotide sequences were selected — ECRECA, ECLEXX, and ECARAC (the designations adopted in the description of the EMBl bank are used here), of 1390, 943, and 1246 nucleotides length, respectively. The sequence ECRECA at site (238, 1296) contains the recA gene of the regulator protein of the *E. coli* SOS system, possessing the ability for intensive expression [19]. Fragment (102, 707) of the ECLEXX sequence contains the gene lexA, which is a repressor of the SOS-system of protein synthesis [20]. The ECARAC sequence contains the gene araC (270, 1145), which codes the protein repressor of the arabinose operon [21].

Figure 1 presents the plots of the quantities $P(K_i|\alpha)$, $i = 1, 2, 3$, obtained for the ECRECA sequence at $V = (1, 16)$ for three possible read frames. The numbers on the horizontal read out the number of nucleotides from the start of the sequences. The segment of the vertical axis corresponds to the interval (0, 1). Here and in subsequent figures, the solid lines at the 0.5 level denote triplet chains lacking terminating codons. They begin from the long or short vertical slashes, which denote the triplets ATG or GTG, respectively, and end in the long slashes shifted downward, which indicate the position of the terminating triplets.

Figures 2 and 3 give plots of the quantities $P(K_i|\alpha)$ obtained in the case $V = (1, 16)$ for the sequences ECLEXX and ECARAC. It is easily noted that coding regions are recorded in Figs. 1a, 2c, and 3c. Moreover, the behavior of the indicator functions is of a differing character. It is apparent that the degree of "resolution" of the coding regions worsens from ECRECA to ECARAC. Formally, this may be reflected using the values $\bar{p}_i$, $i = 1, 2, 3$ — the means of the $P(K_i|\alpha)$ numbers taken within the limits of the real coding region — as well as the quantities s_i, which are the mean squared deviations of $P(K_i|\alpha)$ from P_i in the same interval.

In the given case $\bar{P}_1 = 0.95$ and $s_1 = 0.09$ for the recA gene, $\bar{P}_3 = 0.86$ and $s_3 = 0.21$ for the lexA gene, and $\bar{P}_3 = 0.51$ and $s_3 = 0.36$ for the araC gene. The same quantities calculated from other algorithm variants are presented in Table 1. According to the table, the greatest degree of resolution for the araC gene is attained in the case $V = (2, 48)$. A plot of the corresponding indicator function, but only for the coding frame, is presented in Fig. 4.

Returning to Figs. 1-3, it should be noted that if an error of the type of a deletion or insertion of some number of nucleotides other than three arose during the interpretation of DNA it would be manifested in the "jumping" of the coding-region indicator from one frame to another.

Thus, it can be concluded from the foregoing control parameters that the application of the described algorithm gives sufficient information for the identification of coding regions. The obtained results also permit the postulate that at fixed algorithm parameters the degree of resolution of the coding region will be higher the greater the maximally possible rate of expression of the coded protein. In this case, the $\bar{p}$ and s quantities are informative characteristics of the coding region associated with the rate of expression of the protein product.

We shall consider, for example, the still incompletely functionally identified sequence ECRRNBZ, containing the ribosomal operon. It was described earlier [22], and the open read frame on the (275, 1141) interval was identified in it. The plot of the indicator function obtained by calculation for $V = (2, 32)$ for fragment (1, 1300) in the second read frame is presented in Fig. 5. The values of $\bar{p}_2$ and s_2 for the interval (275, 1141) are 0.73 and 0.34, respectively. This provides the basis to believe that fragment (275, 1141) codes a still unidentified *E. coli* protein, possessing a high degree of expression.

In conclusion, we note the following in comparison with the earlier methods [8-11]. With regard to the variant V = (1, 16), which we considered in greatest detail here, the presented algorithm is not inferior in terms of the degree of resolution of the coding regions to methods described earlier [8, 10, 11]. However, compared with the work of Staden [10], which described a method using triplet statistics and a "window" size of no less than 15 codons, our method required less work to prepare the initial information. Specifically, instead of three tables each containing 64 frequencies of codon encounter, we need three tables with 16 values of transitional probabilities [16]. The other algorithm, using triplet statistics [11], presupposes the use of a wider "window" of from 25 to 50 codons, which increases the expenditure of computer time. Staden and Fickett [9, 10] also proposed algorithms associated with the positional statistics of mononucleotides. Corresponding to these approaches is the use of a nonuniform Markov chain of zero order as a model of the coding region. Therefore, we can turn to Table 1, which shows that when l = 16 such an algorithm lacks sufficient precision and, as in the case in [11], the size of the "window" must be increased. Finally, the use of $\bar{p}$ and s makes it possible to classify the genes with respect to the features of construction of nucleotide sequences associated with the degree of protein expression. The latter question clearly requires separate discussion.

It is also appropriate to note that, as shown by calculations, the application of uniform Markov chains [16] as models of the DNA coding regions for the recognition algorithm presented above does not yield applicable results. This again confirms our conclusions [16] concerning the low precision of the statistical description of the nucleotide structures of coding regions using uniform Markov chains.

LITERATURE CITED

1. A. M. Maxam and W. Gilbert, Proc. Natl. Acad. Sci. USA, 74, 560-564 (1977).
2. F. Sanger, S. Nicken, and A. R. Coulson, Proc. Natl. Acad. Sci. USA, 74, 5463-5467 (1977).
3. D. K. Hawley and W. R. McKlure, Nucleic Acids Res., 11, 2237-2255 (1983).
4. E. B. Keller and W. A. Noon, Proc. Natl. Acad. Sci. USA, 81, 7417-7420 (1984).
5. G. D. Stormo, T. D. Schreider, and L. M. Gold, Nucleic Acids Res., 10, 2972-2996 (1982).
6. N. A. Kolchanov, V. V. Solov'ev, and A. A. Zharkikh, Dokl. Akad. Nauk SSSR, 273, 741-744 (1983).
7. V. B. Zhurkin, Nucleic Acids Res., 9, 1963-1967 (1981).
8. R. Staden and A. D. McLachlan, Nucleic Acids Res., 10, 141-156 (1982).
9. J. W. Fickett, Nucleic Acids Res., 10, 5303-5318 (1982).
10. R. Staden, Nucleic Acids Res., 12, 551-567 (1984).
11. M. Gribskov, J. Devereux, and R. R. Burgess, Nucleic Acids Res., 12, 539-549 (1984).
12. J. C. W. Shepherd, Proc. Natl. Acad. Sci. USA, 78, 1596-1600 (1981).
13. F. Sanger, A. R. Couson, G. F. Hong, D. F. Hill, and G. B. Petersen, J. Mol. Biol., 162, 729-773 (1982).
14. A. Tramontano, V. Scarlato, N. Barni, M. Cipollaro, A. Franze, and A. Cascino, Nucleic Acids Res., 12, 5049-5059 (1984).
15. M. Yu. Borodovskii, Yu. A. Sprizhitskii, E. I. Golovanov, and A. A. Aleksandrov, Mol. Biol., 20, 1014-1023 (1986).
16. M. Yu. Borodovskii, Yu. A. Sprizhitskii, E. I. Golovanov, and A. A. Aleksandrov, Mol. Biol., 20, 1024-1033 (1986).
17. Yu. A. Rozanov, Random Processes [in Russian], Nauka, Moscow (1971).
18. M. Gouy and C. Goutier, Nucleic Acids Res., 10, 7055-7074 (1982).
19. A. Sancar, C. Stashelek, W. Konigsberg, and W. D. Rupp, Proc. Natl. Acad. Sci. USA, 77, 2611-2615 (1980).
20. B. E. Markhan, J. W. Little, and D. W. Mount, Nucleic Acids Res., 9, 4149-4161 (1981).
21. C. G. Miada, A. H. Horwitz, L. G. Cass, J. Timko, and G. Wilcox, Nucleic Acids Res., 8, 5267-5274 (1980).
22. J. Brosius, T. J. Dull, D. D. Sletter, and H. F. Noller, J. Mol. Biol., 148, 107-127 (1981).