Computers Chem. Vol. 17, No. 2, pp. 123–133, 1993
Printed in Great Britain

0097-8485/93 $6.00 + 0.00
Pergamon Press Ltd

GENMARK: PARALLEL GENE RECOGNITION FOR BOTH DNA STRANDS*

MARK BORODOVSKY† and JAMES MCININCH
School of Biology, Georgia Institute of Technology, Atlanta, GA 30332-0230, U.S.A.

(Received 24 September 1992; in revised form 24 November 1992)

Abstract—The problem of predicting gene locations in newly sequenced DNA is well known but still far from being successfully resolved. A novel approach to the problem based on the frame dependent (non-homogeneous) Markov chain models of protein-coding regions was previously suggested. This approach is, apparently, one of the most powerful "search by content" methods. The initial idea of the method combines the specific Markov models of coding and non-coding region together with Bayes' decision making function and allows easy generalization for employing of higher order Markov chain models. Another generalization which is described in this article allows the analysis of both DNA strands simultaneously. Currently known gene searching methods perform the analysis of the two DNA strands in turn, one after another. In doing this all the known methods fail in the sense that they generate false (artifactual) prediction signals for the given strand when the real coding region is located on the complementary DNA strand. This common drawback is avoided by employing the Bayesian algorithm which uses an additional non-homogeneous Markov chain model of the "shadow" of the coding region—the sequence which is complementary to the protein-coding sequence.

INTRODUCTION

Large-scale DNA sequencing calls for fast and efficient gene recognition methods since the search for new genes is at the top of the genome sequencing project priorities. A number of gene recognition methods have been suggested and implemented on different size computers as well as in networks (Fickett, 1982; Staden, 1984; Gribskov *et al.*, 1984; Almagor, 1985; Claverie & Bougueleret, 1986; Fichant & Gautier, 1987; Fields & Soderlung, 1990; Konopka & Owens, 1990; Uberbacher & Mural, 1991; Guigo *et al.*, 1992; see other references in Stormo, 1987; Gelfand, 1990). Nevertheless, there is sill a clear difference between the accuracy of the existing methods and the needs of biologists.

Recently, we have improved the method suggested by Borodovsky *et al.* (1986b) and have shown that the predictive accuracy of this method for fourth-order Markov chain models of coding and non-coding regions approaches a 10.0% false negative and 25.2% false positive rate for a control set consisting of 96 bp fragments of *Escherichia coli* DNA (Borodovsky & McIninch, 1993). Actually, such a test using short isolated fragments should be a rigid assay since in practice the average coding region has a larger size and the context information available from analysis of flanking regions increases the total accuracy of the prediction of the protein-coding region.

The current paper is devoted to further improvement of the Markov chain/Bayes method and making it applicable to the common case of searching for genes in newly sequenced DNA. In this case the DNA sequence is still absolutely symmetrical from the point of view of its possible functional meaning and one needs to search for genes along both DNA strands. None of the existing methods was developed for the analysis of two strands simultaneously. The usual idea was to apply the method developed for one strand analysis twice: once for the direct strand and in a second time for the complementary strand (Fickett, 1982; Gribskov *et al.*, 1984; Uberbacher & Mural, 1991). Sometimes this idea is implemented by sequence analysis software developers even without the authors' participation. It has been observed by many who applied these methods that strong false signals have appeared, for instance, on the panel characterizing the direct DNA strand, apparently, having been induced by a true coding region located on the complementary strand and vice versa (see Fig. 1, Table 1 and comments in the next section). Predictions of the coding regions that are made on the basis of this vague picture can be relatively easily done in the prokaryotic case when the length of open reading frame (ORF) which corresponds to the true coding region is as a rule much larger than the competitive one on the complementary strand. However, a fast search for gene locations, which is the most important task during high speed sequencing, has become quite difficult even in the prokaryotic case, to say nothing about eukaryotes.

* The preliminary version of this work was presented during the *Second International Workshop of Open Problems in Computational Molecular Biology*, Telluride Summer Research Center, Telluride, Colo., 19 July–2 August, 1992.

† Author for correspondence.

Table 1. Grail recognition system output for sequence ECARAC (>ECARAC, length = 1246; potential exons are listed in the following)

Position	Strand*	Strand probability	Frame	Quality	Open reading frame
821–909	r	0.53	3	Good	642–909
681–771	r	0.71	2	Excellent	653–830
498–641	r	0.64	3	Excellent	498–642
221–291	r	0.61	3	Excellent	102–306
1111–1146	f	0.54	3	Marginal	204–1146

The ORF (204–1146) containing a true coding region is listed here as marginal quality, which is certainly less promising than the other four suggested choices. Only additional analysis of the ORF lengths can give a strong indication in favor of the true coding region.

* f, Forward; r, reverse.

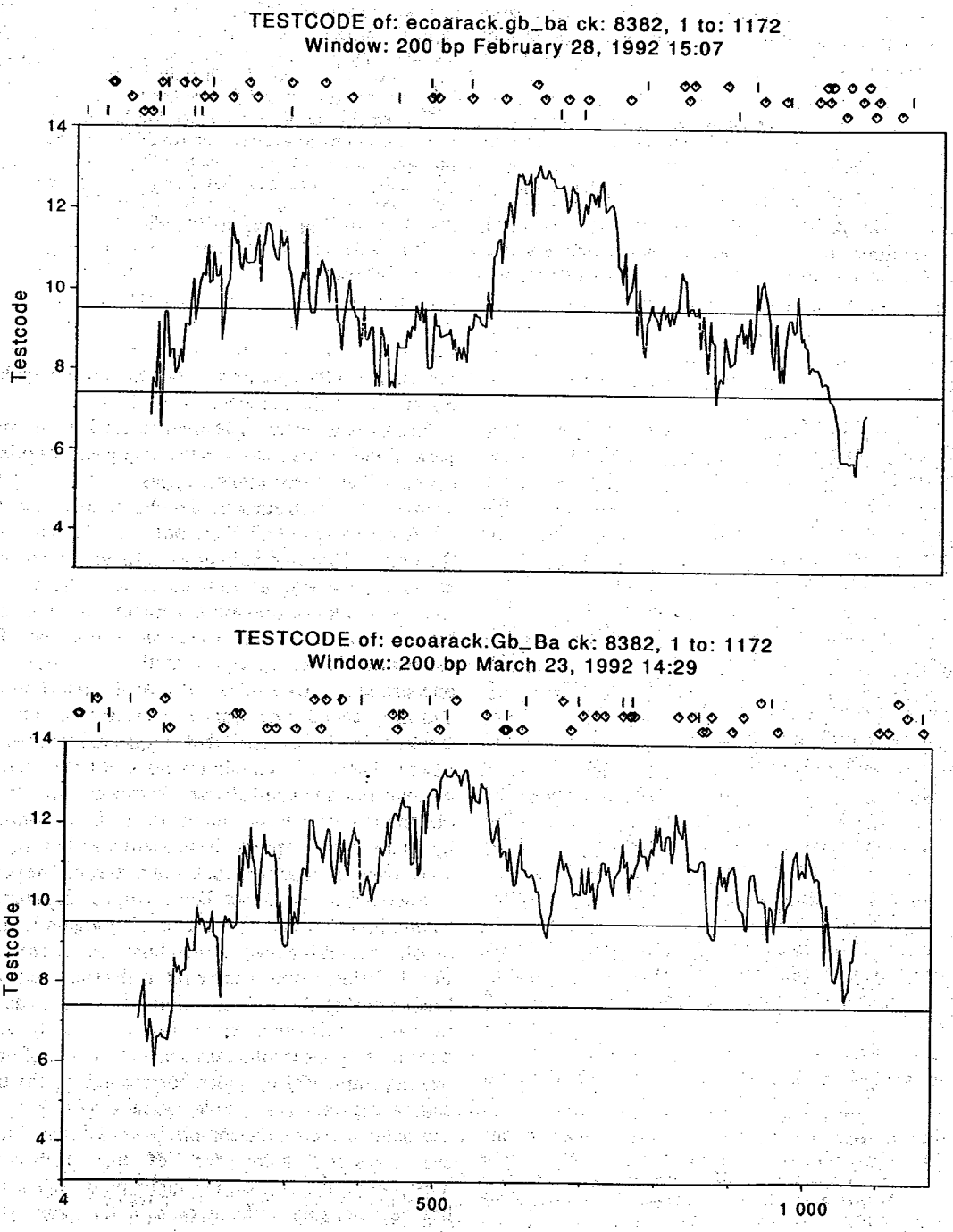

Fig. 1. *Caption on facing page.*

The Bayes' formalism employed in the our method (Borodovsky *et al.*, 1986b) is very flexible and allows one to easily enlarge the number of possible situations which should be distinguished from one another. We need to incorporate into the common Bayes algorithm the additional outcomes which appear in the simultaneous mode of analysis when the protein-coding region is located on the strand complementary to the one being analyzed. This is possible using additional non-homogeneous Markov chain models of a "shadow" of the coding region. Thus, in what follows, we describe the new version of the method which completely avoids the aforementioned drawback of generating false predictions in the region of a gene shadow.

ALGORITHM

Markov chain models

The previous papers (Borodovsky *et al.*, 1986a; Borodovsky & McIninch, 1993) contain a detailed explanation of constructing non-homogeneous (frame-dependent) Markov chain models of protein-coding DNA sequences and their utilization for Bayesian gene recognition algorithms. Figure 2 shows an illustration of the application of this "one-strand-only" non-homogeneous Markov chain method.

Let us consider the ECARAC sequence (name given in EMBL notation). This sequence contains a protein coding subsequence (270–1145) which encodes a low expression regulatory protein araC. The protein araC controls initiation and transcription of structural genes involved in the transport and metabolism of L-arabinose (araBAD operon) and in addition controls its own synthesis (Sancar *et al.*, 1980).

We have analyzed the sequence ECARAC using TestCode (Fickett, 1982) and GRAIL (Uberbacher & Mural, 1991) methods. The TestCode indicator function (Fig. 1) has shown a strong signal for the complementary sequence (bottom panel). The GRAIL coding recognition module also provides (Table 1, only the GRAIL output summary is presented) quite strong indications for the existence of several coding regions on the complementary strand as well. Note that the TestCode function is used as a component of GRAIL discrimination criteria. These examples are given not for the purpose of general critique of the methods mentioned. By the way, the GRAIL does not claim responsibility for the predictions for species other than human. We would just like to make clear some difficulties which can be caused by the shadows of the coding regions.

The similar false signals appear when GRAIL is applied to human DNA or methods described by Staden (1984) or Gribskov *et al.* (1984) are used for *E. coli*, for instance. The same is true when the "one-strand-only" Markov chain/Bayes method is applied (see the result below for ECARAC).

Let us now consider the *E. coli* sequence ECRECA. The ECRECA subsequence (238, 1296) encodes the protein recA which is involved in important cellular functions such as cell division, recombination-repair, mutagenesis and phage-induction. In non-induced cells recA protein is made only in small quantities, whereas in induced cells the gene is as actively transcribed as ribosomal RNA gene (Miada *et al.*, 1980).

The result of analysis done for sequence ECRECA by previously suggested one-strand-only Markov chain/Bayes method (Borodovsky *et al.*, 1986b) is presented in Fig. 2 in the way similar to Fig. 1, i.e. in the form of indicator function charts. Second-order Markov chain models of coding and non-coding regions were used. The size of the moving window is equal to 96 bp and step of the consequent moving is equal to 6 bp. The coding region is identified in the first reading frame (defined by the first nucleotide of the sequence ECRECA).

There is a clearly evident false signal on the bottom panel which corresponds to the complementary first reading frame. This artifact appears when the "one-strand-only" algorithm is applied to the sequence complementary to ECRECA and when the shadow of the true coding region is processed. In the previous paper (Borodovsky & McIninch, 1993) one can find more examples of the utilization of the "one-strand-only" analysis algorithm. It was shown that the intensity of the artificial signals decreases when higher order Markov chain models are employed; the decrease in the false signal intensity is also observed in the analysis of the DNA sequences coding for genes with low expressivity. Note, that the pattern of codon usage of highly expressible genes is strongly shifted to the limited number of so-called optimal codons in comparison with genes with the low expressivity (see references in Gouy & Gautier, 1982).

Figure 3 shows the result of analysis done for the sequence ECARAC. Fifth-order Markov chain models of coding and non-coding regions are used here. The moving window size and the step are the same as in Fig. 2. It can be seen that the coding region

Fig. 1 *(opposite)*. Gene prediction plots obtained by the TestCode method (GCG package realization). The diamonds and vertical bars shown above the graphs designate start codons and stop codons respectively. The graph in the top panel indicates the coding region for the direct DNA sequence ECARAC. The result could be interpreted as prediction of two coding regions in between three other regions which are characterized as "no opinion" (since the corresponding parts of the graph lay in intermediate zone). The graph in the bottom panel shows good prediction for the coding region located on the complementary DNA strand. The true coding region has been determined to be on the direct strand (Sancar *et al.*, 1980).

is identified in the third reading frame. The false signals that appear in the third reading frame of the complementary panels are not so intense as in case shown in Fig. 2. Actually, it is quite explainable from what was mentioned previously since the araC gene has the lower expression level than recA, and we are using higher order Markov chain models.

The explanation of false phenomenon comes from a well known observation (Shepherd, 1981) that coding regions have an excess of RNY type codons (R, purine; Y, pyrimidine). Since this formula is

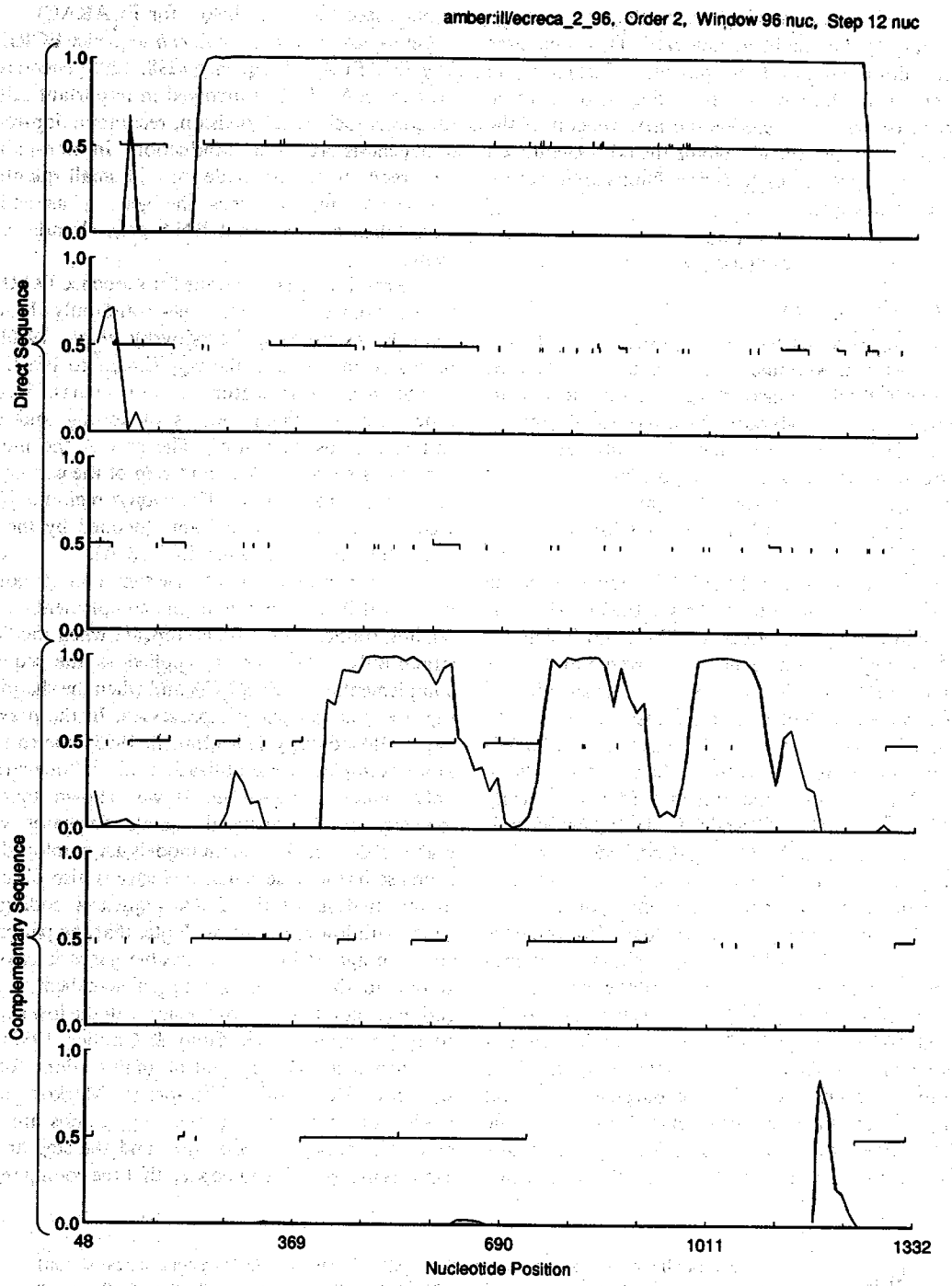

Fig. 2. Indication of protein-coding regions by one-strand-only Markov chain/Bayes method (sequence ECRECA). Six charts represent protein-coding region indicator-functions calculated for ECRECA sequence as described in the Algorithm section of the text. Second-order Markov chain models of coding and non-coding region have been used. The three top panels refer to the 1st, 2nd, and 3rd frames of reading triplets in the direct DNA sequence. The three bottom panels refer to the 1st, 2nd and 3rd frames of reading triplets in the reverse (complementary) DNA sequence.

self-complementary one can expect to find a number of RNY type triplets in the sequence fragment complementary to the real gene. These triplets fall into the same reading frame and eventually produce the false signals. The above mentioned tendency of the increasing of the artifactual noise for *E. coli* highly expressed genes corresponds to the fact that the RNY formula preferably describes optimal codons. On the other hand, this three letter pattern will mostly affect algorithms that use lower than third-order Markov chain models.

Now the modified procedure will be described for the case of first-order Markov chain models. The derivation of higher order models and their use in the algorithm is a rather straightforward generalization of the procedure given below.

The model for non-coding DNA sequence is defined as a homogeneous Markov chain

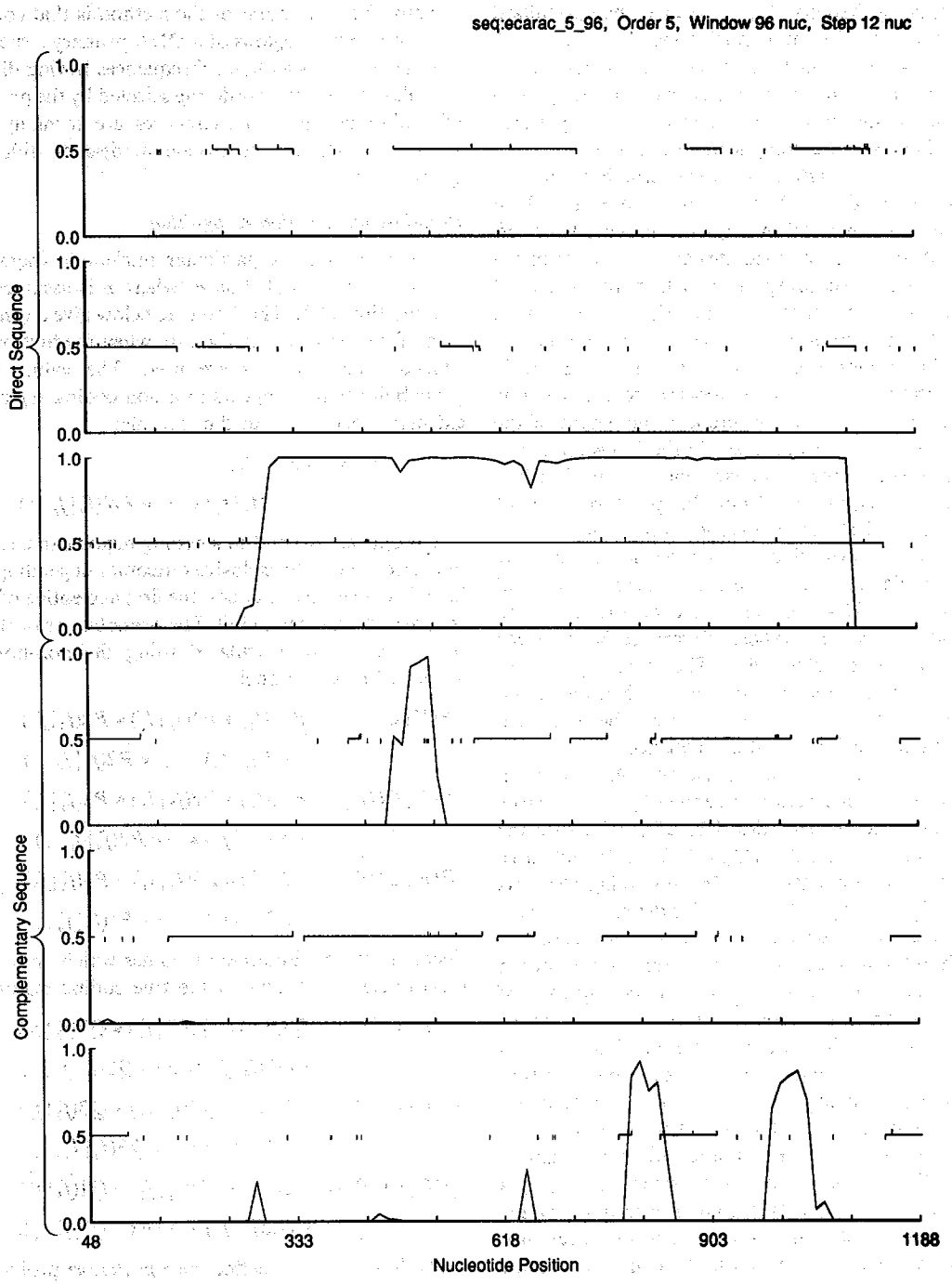

Fig. 3. Indication of protein-coding regions by one-strand-only Markov chain/Bayes method (sequence ECARAC). Fifth-order Markov chain models of coding and non-coding region have been used (see caption to Fig. 3).

(Borodovsky *et al.*, 1986a). Numerical values of the parameters of the first-order model (initial probability vector $\mathbf{PN}_0$ and transition matrix **PN**) are derived from the counts of mono-, and dinucleotides $N(i)$ and $N(ij)$, $i,j = 1,2,3,4$ (numbers are used instead of letters T, C, A, G) calculated from the training set of non-coding DNA sequences. The elements of the transition matrix PN_{ij} are assumed to be equal to $N(ij)/N(i)$ according to the maximum likelihood principle (Billingsley, 1961). The value of the initial probability PN_{0i} is taken as the normalized frequency of the mononucleotide i, $i = 1,2,3,4$. In a sense the model treats patterns of correlation between adjacent nucleotides as uniformly distributed along the sequence, although these patterns could have a more complicated distribution.

There is a certain concern whether such a phenomenological Markov chain model, which generates a model DNA sequence moving from left to right, in some sense violates an *a priori* symmetry presumption according to which both rightward and leftward sequence directions should be equal in respect to the initial training set of DNA sequences.

In fact, a model like this allows one to rigorously define the probability of an appearance of nucleotide string "$x_1, x_2, \ldots, x_k$" (where k is the length of the string) at any particular place in the DNA sequence. It can be shown that rightward and leftward Markov models are equivalent from the point of view of calculation of string appearance probability.

It was shown that a homogeneous Markov chain model does not give a satisfactory statistical description of DNA sequence that codes for a protein (Borodovsky *et al.*, 1986a; Tavare & Song, 1989; Kleffe & Borodovsky, 1992). The same papers considered a non-homogeneous periodic Markov chain model which describes more precisely the three step weak periodicity of a coding sequence.

The first-order non-homogeneous Markov chain model of a coding region is defined by three vectors of initial state probabilities: $\mathbf{P1}_0$, $\mathbf{P2}_0$, $\mathbf{P3}_0$, with the components $P1_{0i}$, $P2_{0i}$, $P3_{0i} = 1, 2, 3, 4$; and three transition matrices **P1, P2, P3**, containing elements $P1_{ij}$, $P2_{ij}$, $P3_{ij}$, $i,j = 1, 2, 3, 4$. The definition of these parameters is made on the basis of the maximum likelihood principle as well. The training set of coding sequences is concatenated into a long sequence of length N with stop-codons excluded. Counts of mono- and dinucleotides $N(i)$ and $N(i,j)$ are divided into three components depending on a position that nucleotide i occupies in a codon. The new statistics are designated as $N^m(i)$ and $N^m(i,j)$ $m = 1,2,3$. Value of the element P^m_{ij} is assumed to be equal to $N^m(ij)/N^m(i)$, value of the initial probability $P^m_0(i)$ is equal to the $N^m(i)/(N/3)$. A similar procedure, counting of $k+1$-tuples and k-tuples split into three subsets (according to the position of the first nucleotide of the k-tuple or $k+1$-tuple in the codon) is accomplished for the definition of parameters of non-homogeneous Markov chain model of the order k.

The non-homogeneous Markov chain model of a shadow of the coding region is now easy to construct. Let us reserve letter Q for all similar designations of parameters that appear for the non-homogeneous Markov chain model of the shadow of the coding region: $\mathbf{Q1}_0$, $\mathbf{Q2}_0$, $\mathbf{Q3}_0$, **Q1**, **Q2**, **Q3** and so on. One can operate with the training set of the shadows of true coding regions or find Q^m_{ij} analytically, combining values of statistics $N^m(ij)$ and $N^m(i)$ which are known from the true coding sequences training set analysis.

Thus, the main point of the method is that coding and non-coding regions of a DNA primary structure are treated as nucleotide subsequences having different rules of nucleotide ordering selected by the process of evolution. These subsequences are formally described by Markov stochastic models of different types.

Protein-coding region recognition

Let us consider a particular nucleotide fragment "$f_1, f_2, \ldots, f_n$" denoted as F (where n is assumed to be a multiple of 3). The formulae below give a general idea of the algorithm in the case when the first-order Markov chain models are used. The value of a probability that F appears in a non-coding region is calculated according to the formula:

$$P(F \mid NON) = PN_0(f_1) * PN(f_2 \mid f_1) * \cdots * PN(f_n \mid f_{n-1}). \quad (1)$$

The appearance of F in a coding region can be split into three mutually exclusive outcomes depending on in which position of a codon the first nucleotide of the fragment F happens to fall. The probabilities of these outcomes can be calculated using the non-homogeneous Markov model

$$\begin{aligned} P(F \mid COD_1) &= P1_0(f_1) * P1(f_2 \mid f_1) * P2(f_3 \mid f_2) * P3(f_4 \mid f_3) * \cdots * P2(f_n \mid f_{n-1}) \\ P(F \mid COD_2) &= P2_0(f_1) * P2(f_2 \mid f_1) * P3(f_3 \mid f_2) * P1(f_4 \mid f_3) * \cdots * P3(f_n \mid f_{n-1}) \\ P(F \mid COD_3) &= P3_0(f_1) * P3(f_2 \mid f_1) * P1(f_3 \mid f_2) * P2(f_4 \mid f_3) * \cdots * P1(f_n \mid f_{n-1}). \end{aligned} \quad (2)$$

There are three additional outcomes which appear if F falls into the shadow of the true coding region

$$\begin{aligned} Q(F \mid COD_1) &= Q1_0(f_1) * Q1(f_2 \mid f_1) * Q2(f_3 \mid f_2) * Q3(f_4 \mid f_3) * \cdots * Q2(f_n \mid f_{n-1}) \\ Q(F \mid COD_2) &= Q2_0(f_1) * Q2(f_2 \mid f_1) * Q3(f_3 \mid f_2) * Q1(f_4 \mid f_3) * \cdots * Q3(f_n \mid f_{n-1}) \\ Q(F \mid COD_3) &= Q3_0(f_1) * Q3(f_2 \mid f_1) * Q1(f_3 \mid f_2) * Q2(f_4 \mid f_3) * \cdots * Q1(f_n \mid f_{n-1}). \end{aligned} \quad (3)$$

The final step is to define the *a posteriori* probabilities $P(COD_m \mid F)$ and $Q(COD_m \mid F)$ which characterize the coding property of the fragment F being read in six possible ways. Three components $P(COD_m \mid F)$

$m = 1, 2, 3$ are determined according to the Bayes' formula

$$P(COD_m \mid F) = \frac{P(F \mid COD_m) * P(COD_m)}{\sum_j P(F \mid COD_j) * P(COD_j) + \sum_j Q(F \mid COD_j) * Q(COD_j) + P(F \mid NON) * P(NON)}. \quad (4)$$

The designation $P(COD_m)$ stands for the *a priori* probability of the event COD_m, $m = 1, 2, 3$, which means that any as yet unspecified fragment F falls into a coding region (and its first nucleotide is located in a codon position defined by index m).

The same kind of formula defines probabilities of different shadow phases when the fragment F is observed

$$Q(COD_m \mid F) = \frac{Q(F \mid COD_m) * Q(COD_m)}{\sum_j P(F \mid COD_j) * P(COD_j) + \sum_j Q(F \mid COD_j) * Q(COD_j) + P(F \mid NON) * P(NON)}. \quad (5)$$

The designation $Q(COD_m)$ stands for the *a priori* probability of the event COD_m, $m = 1, 2, 3$, that an as yet unspecified fragment F falls into a coding region shadow (and that the first nucleotide of F is located in a codon position defined by index m).

The designation $P(NON)$ stands for the *a priori* probability of the event—*NON*, that an as yet unspecified fragment F falls into a non-coding region. The natural assumption here is that $P(NON) = 1/2$ and that $P(COD_m) = Q(COD_m) = 1/12$ for $m = 1, 2, 3$. Formulae (4)–(5) determine six coding—in-frame *a posteriori* probabilities—for any one given fragment F of the DNA sequence.

The value

$$P(NON \mid F) = \frac{P(F \mid NON) * P(NON)}{\sum_j P(F \mid COD_j) * P(COD_j) + \sum_j Q(F \mid COD_j) * Q(COD_j) + P(F \mid NON) * P(NON)}. \quad (6)$$

gives an *a posteriori* probability of the event that a given fragment F belongs to non-coding region. The total of $P(COD_m \mid F)$ and $Q(COD_m \mid F)$, $m = 1, 2, 3$, is designated a $P(COD \mid F)$. We assume that $P(COD \mid F) + P(NON \mid F) = 1$ thus, the case when fragment F is partially coding and partly non-coding is not considered.

IMPLEMENTATION

Graphical output

The above algorithm was implemented using the ANSI-C language as a program on an Amiga 3000 and IBM-386 personal computers and on an IBM RX 6000 computer. The training sets of protein-coding and non-coding regions consisted of 479,589 and 245,307 bp respectively. The limited size of the current training set did not allow the use of models of order higher than five.

The sliding window size and the step size of a sliding are parameters of the algorithm. In our calculations we used the window sizes: 16, 32 or 48 codons and the sliding step equal to 6 or 12 nucleotides. The probability values $P(COD_m \mid F)$, $Q(COD_m \mid F)$ $i = 1, 2, 3$ were calculated for every nucleotide fragment F found in the window opening and referred to this fragment's middle point. The sequence of these probability values forms the six indicator functions.

Thus, when the nucleotide sequence is read only from the one strand, the gene searching has been performed simultaneously on the two complementary strands. The graphical output of the algorithm is designed in the form of six panels on a page (the top three for the direct strand and the bottom three for the reverse one). Each panel corresponds to one of six possibilities of reading subsequent triplets in a DNA sequence. The vertical axes represent values of the probability $P(COD_m \mid F)$ and $Q(COD_m \mid F)$ while the horizontal axes represent nucleotide positions along a DNA sequence. Positions of translation start and stop codons are marked by small vertical ticks at the 0.5 level. Upward ticks denote the position of the start codons ATG and GTG (the tick's length for GTG is half that of the ATG one). Downward ticks show the positions of the translation termination triplets TAA, TAG, and TGA. Every start codon gives an origin to an open reading frame which is marked by solid line. Note that the direction of the reading of the protein code in the bottom three panels is leftward.

The examples of the above described graphics are given in Fig. 4 and Fig. 5 for the same *E. coli* sequences ECRECA and ECARAC. The result of using the second-order algorithm for the ECRECA sequence analysis is shown in Fig. 4. It is seen again that the coding region is identified in the first reading frame of the direct sequence. No false signal now appears in the first reading frame panel of the complementary sequence. A similar result for the sequence ECARAC is shown in Fig. 5.

Accuracy of the recognition procedure

Figures 2–5 give just an example of the graphical output of the method. The statistical evaluation of the quality of the recognition procedure has been done by the estimation of the average false positive and false negative rates.

For this purpose two control sets of *E. coli* DNA, coding and non-coding regions being 373,845 and 131,538 bp in size, were considered. Each of these sequence sets was divided into a number of non-overlapping fixed size fragments (the fragment length was taken equal to 48, 96, and 144 base pairs). For each fragment the value of the *a posteriori* probability of its protein-coding function was calculated by a particular version of the recognition algorithm. Five versions of the algorithm were used for calculations corresponding to different orders of Markov chain models (from first through fifth). The set of probability values obtained from the utilization of a given version of the algorithm can be presented in the form of a histogram on the interval (0,1). These data are represented in tabular form. In Tables 2 and 3 we

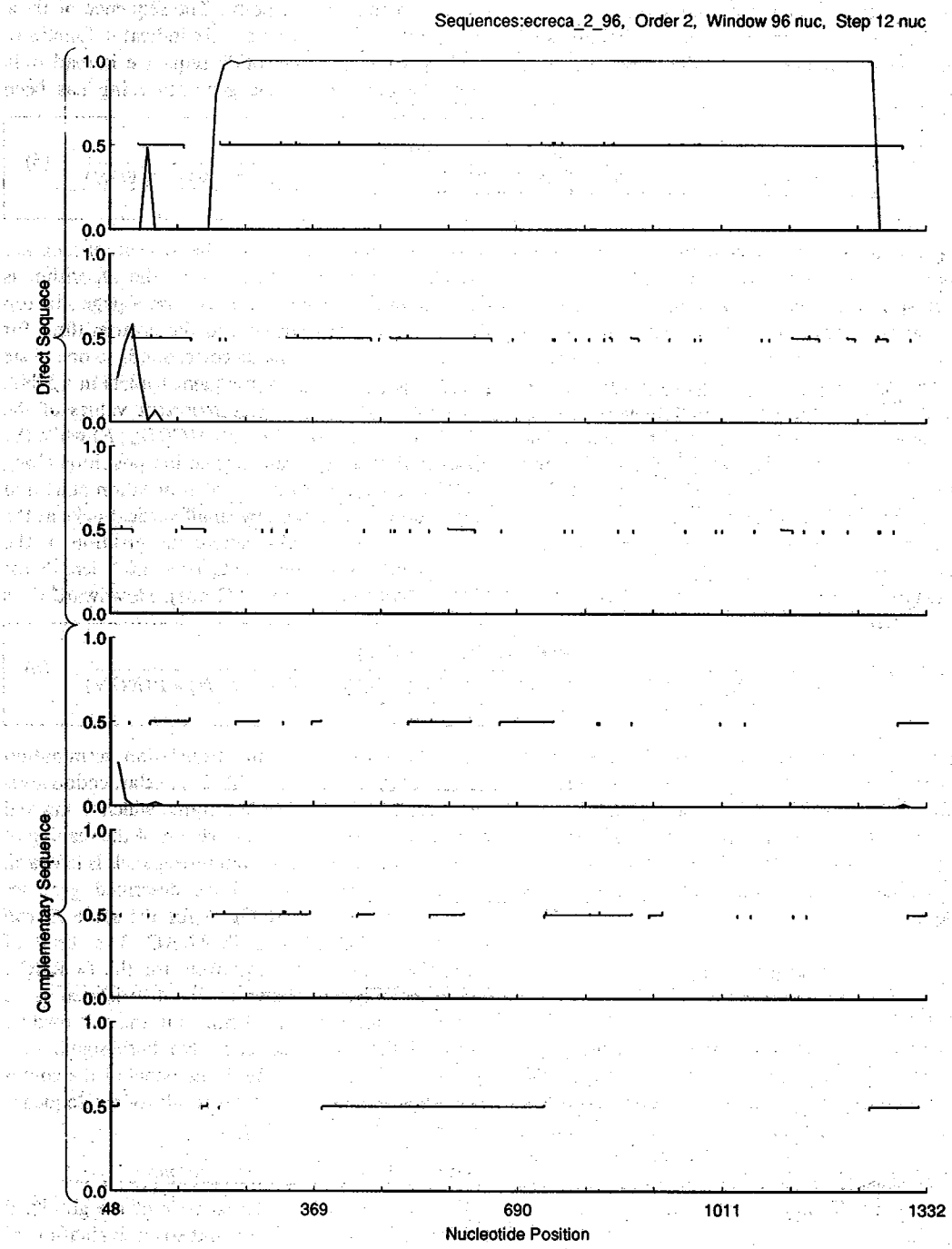

Fig. 4. Indication of protein-coding regions by the both-strand Markov chain/Bayes method (sequence ECRECA). Six charge representing protein-coding region indicator functions obtained by second-order both-strand-together method (see caption to Fig. 3).

Table 2. Distribution of protein-coding probability function in the set of coding fragments

Order	Cumulative histogram value					
	0.0–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–1.0
1	0.185	0.037	0.034	0.037	0.049	0.659
2	0.107	0.016	0.014	0.020	0.021	0.822
3	0.106	0.011	0.014	0.018	0.014	0.837
4	0.104	0.010	0.016	0.013	0.017	0.841
5	0.117	0.012	0.013	0.105	0.017	0.825

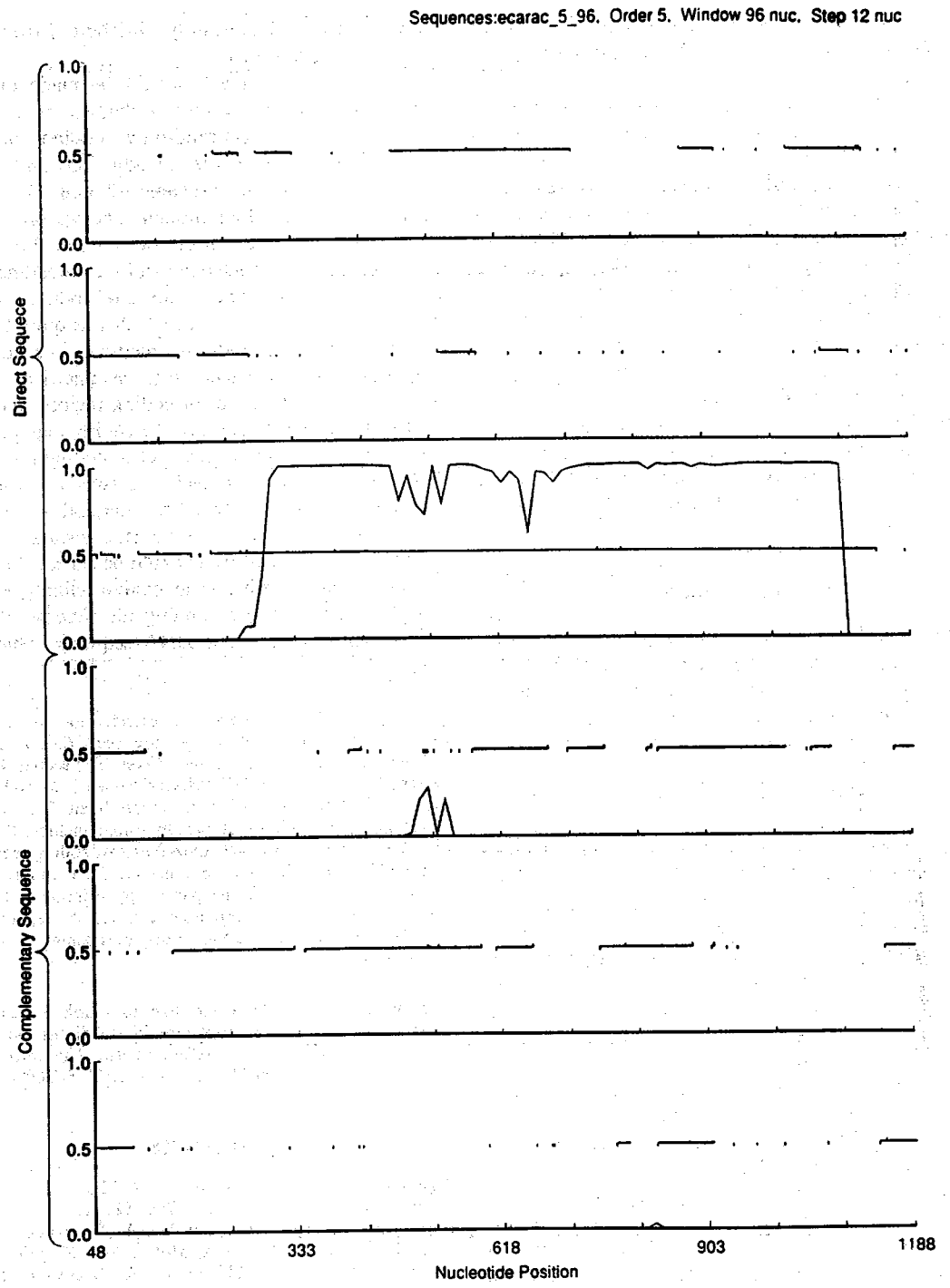

Fig. 5. Indication of protein-coding regions by the both-strand-together Markov chain/Bayes method (sequence ECARAC). Six charts representing protein-coding region indicator functions obtained by fifth-order both-strand-together method (see legend to Fig. 3).

Table 3. Distribution of protein-coding probability function in the set of non-coding fragments

Order	Cumulative histogram value					
	0.0–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–1.0
1	0.720	0.033	0.026	0.028	0.021	0.172
2	0.757	0.012	0.013	0.009	0.014	0.196
3	0.764	0.011	0.007	0.013	0.010	0.196
4	0.766	0.010	0.007	0.012	0.007	0.198
5	0.772	0.007	0.015	0.009	0.007	0.190

list the relative frequencies of a falling of calculated probability value into six subintervals which cover the interval (0,1).

Table 2 shows the histogram data sets obtained by the fifth-order version of the recognition algorithm for 3894 true coding fragments of 96 bp length. It is seen that if the threshold of decision making is set at a 0.5 level then the rate of false negative predictions (coding identified as a non-coding) is 25.6% for the algorithm using first-order Markov chains and this rate decreases to 14.2% for the case when the fourth-order Markov chain model is used. Another important parameter: the false positive rate for non-coding fragments (non-coding identified as coding) does not change significantly: from 22.1% for the first-order to 21.7% for the fourth-order method. This conclusion comes from Table 3 which shows protein-coding probability values calculated by five algorithm versions for 1370 non-coding fragments of 96 bp length.

DISCUSSION

The results presented above have shown a reasonable accuracy of the Bayesian algorithm based on Markov chain models. This algorithm is able to reveal the coding DNA strand and true reading frame where a gene is located and generate a distinctive identifying signal.

The elimination of the false signals is a significant feature of the algorithm. It makes it easier to choose the best decision when the competitive ORFs appear in complementary DNA strands (prokaryotic case). It should help even more in the case of eukaryotic sequence analysis when supportive information such as ORF location is less useful. That is why any signal (true or false) that appears in the course of the analysis requires the application of additional resources in order to clarify its real nature.

The current version of the algorithm is very sensitive to the pattern of nucleotide correlations. To achieve good results the sequence to be analyzed should be taken from the same statistical population as the training set is. So, one cannot expect that the algorithm trained on the *E. coli* sequence set will be successfully applied to the sequence taken from the genome of the other species. For instance, even for the case of bacteriophage lambda the algorithm trained on *E. coli* works properly only for the first 21,000 bp. It does not produce any satisfactory results in the late genes regions which, as it is well known, has a significantly different pattern of nucleotide correlations.

One important remark should be made on the comparison of the accuracy figures that were determined for one-strand-only version of the method (mentioned in the Introduction) and both-strands-together version (Tables 2 and 3). These figures seem to be close enough. The reason is that the accuracy of the one-strand-only method was evaluated on the restricted control sets of coding and non-coding regions. The shadows of coding regions were not taken into account, which was quite favorable for the one-strand-only method. The current both-strands-together method is "symmetrical" with respect to the control set of coding regions and the control set of the shadows of the coding regions—so we can use only one of them (set of coding regions). Consequently, the final level of predictive accuracy of the both-strands-together method obtained should be considered as a realistic accuracy parameter. This accuracy level should be expected in the real-life situation when one cannot eliminate the potential opportunity of finding the shadow of the coding region in the new DNA sequence which is analyzed.

Program availability—The above described method can be used for the analysis of newly sequenced *E. coli* DNA through the Georgia Tech E-mail server. The sequence can be sent to the program GENMARK which is available at the E-mail address genmark @ ford. gatech.edu. The output of the program which is sent back by E-mail includes the list of ORFs that have been recognized as real coding regions. The optional PostScript output file gives an opportunity to obtain the full six frame picture by printing out this file on a PostScript compatible printer. A version of GENMARK for human DNA sequences is being developed at the present time.

Acknowledgements—We would like to thank James W. Fickett and Andrzej K. Konopka for fruitful discussions and an anonymous reviewer for useful criticism. This work was supported in part by NIH grant NO. 1R01 HG00783-01.

REFERENCES

Algamor H. (1985) *J. Theor. Biol.* **117,** 127.
Billingsley P. (1961) *Ann. Math. Stat.* **82,** 12.
Borodovsky M. Yu., Sprizhitsky Yu. A., Golovanov E. I. & Alexandrov A. A. (1986a) *Molek. Biol.* **20,** 833.
Borodovsky M. Yu., Sprizhitsky Yu. A., Golovanov E. I. & Alexandrov A. A. (1986b) *Molek. Biol.* **20,** 1144.
Borodovsky M. Yu. (1990) *Computer Analysis of Genetic Texts* (Edited by Frank-Kamenetsky M. D.), pp. 81–112. Moscow.

Borodovsky, M. & McIninch J. (1993) *Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*. In press.
Claverie J. M. & Bougueleret L. (1986) *Nucleic Acids Res.* **14,** 179.
Fichant G. & Gautier C. (1987) Statistical method for predicting protein coding regions in nucleic acid sequences. *Comput. Appl. Biosci.* **3,** 287.
Fickett J. W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* **10,** 5303.
Fields C. A. & Soderlund C. A. (1990) *Comput. Appl. Biosci.* **6,** 263.
Gelfand M. (1990) *Biotech. Software* **7,** 3.
Gouy M. & Gautier G. (1982) *Nucleic Acids Res.* **10,** 7055.
Gribskov M., Devereux J. & Burgess R. R. (1984) *Nucleic Acids Res.* **12,** 539.
Guigo R., Knudsen S., Drake N. & Smith T. (1992) *J. Mol. Biol.* **226,** 141.
Kleffe J. & Borodovsky M. (1992) *Comput. Appl. Biosci.* **7,** 433.
Konopka A. K. & Owens J. (1990) *Gene Anal. Tech. Appl.* **7,** 35.
Kozhukin C. G. & Pevzner P. A. (1991) *Comput. Appl. Biosci.* **7,** 39.
Lapedes A., Barnes C., Burks C., Farber R. & Sirotkin K. (1990) *Computers and DNA* (Edited by Bell G. I. & Marr T.), pp. 157–182. Addison–Wesley, Reading. Mass.
Miada C. G., Horwitz A. H., Cass L. G., Timko J. & Wilcox G. (1980) *Nucleic Acids Res.* **8,** 5267.
Sancar A., Stashelek C., Konigsberg W. & Rupp W. D. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77,** 2611.
Shepherd J. C. W. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78,** 1596.
Staden R. (1984) *Nucleic Acids Res.* **12,** 505.
Stormo G. D. (1987) *Nucleic Acid and Protein Sequence Analysis: a Practical Approach* (Edited by Bishop M. J. Rawlings C. J.), pp. 359–385. IRL Press, Oxford.
Tavare S. & Song B. (1989) *Bull. Math. Biol.* **51,** 95.
Uberbacher E. C. & Mural R. J. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88,** 11261.