BioSystems, 30 (1993) 161–171
Elsevier Scientific Publishers Ireland, Ltd.

Recognition of genes in DNA sequence with ambiguities

Mark Borodovsky and James McIninch

School of Biology, Georgia Institute of Technology, Atlanta, GA 30332-0230, and Institute of Molecular Genetics, Moscow, Russia

The search for genes in a newly sequenced DNA is a well known problem. Among other factors, the gene-searching process is hampered by a number of ambiguities which may remain unresolved experimentally for a long time. A computer method that is able to predict genes in a DNA sequence containing ambiguities has been developed, based on the non-homogeneous Markov chain technique. The reliability of the method has been tested using a set of sequences generated by a Monte-Carlo procedure and a set of 425 *E. coli* sequences with ambiguities introduced artificially.

Introduction

Usually, molecular biologists start analyzing DNA texts coming from sequencing long before these texts are ultimately refined and contain no ambiguities or insertions/deletions.

A computer method for prediction of coding regions might be useful in this situation, since it would provide additional information for resolving ambiguities, show insertions/deletions and frameshifts, and allow one to start investigating genes and translated protein sequences immediately. Such a method should be robust to the presence of ambiguities and insertions/deletions and should work properly without any information about positions of start and stop signals which may be easily masked by ambiguities.

Various methods for recognition of coding regions have been considered since the beginning of the sequencing era. They can be divided into two main groups: "search by content" and "search by signal". The latter group looks for specific boundary sites of coding regions, and from the formal point of view this group is close to other methods of recognition of short functional sites (referenced in Bishop and Rawlings, 1987; Gelfand, 1992). The search by content methods perform statistical analysis of longer DNA fragments that are suspected to contain protein-coding regions. Several approaches have already been exploited for this purpose (Shepherd, 1981; Shulman et al., 1981; Fickett, 1982; Staden and McLachlan, 1984; Staden, 1984; Gribskov et al., 1984; Konopka and Owens, 1990).

Nevertheless, the problem posed by analysis of raw sequences was not directly addressed yet. here we present a generalization of our previously published "search by content" method based on the non-homogeneous Markov chains technique (Borodovsky et al., 1986a, b). In the following paragraphs we introduce the Markov chain models for coding and non-coding regions of DNA written in both 4-letter and 15-letter "alphabets". We obtain actual Markov chain parameters by statistical processing of the training sets consisting of *E. coli* coding and non-coding sequences. Then we define the formula for calculating the probability of "coding" for a given DNA fragment (like the Fickett's TestCode coding potential (1982), but the characteristics presented here has a definite mathematical sense).

Finally we evaluate the reliability of the method in two ways: first, by using synthetic coding and non-coding DNA sequences; and second, by using a control set of 425 *E. coli* sequences.

0303-2647/93/$06.00 © 1993 Elsevier Scientific Publishers Ireland, Ltd.
Printed and Published in Ireland

Markov Chain Models

As it is well known, there are specific correlations between adjacent nucleotides in chromosomal DNA sequences. It means that if we have a nucleotide of type A in an arbitrarily chosen position, then we can find a nucleotide of type B in the next position with a frequency that might differ from the frequency of nucleotide B in the entire set of DNA sequences of this species.

That is why the Markov chain model of the DNA sequence is introduced. The Markov chain is a mathematical object with, say, K states. The necessary parameters of the Markov chain are: initial probabilities of states: $P_{0i}, i = 1, \ldots, K$; and probabilities of transitions between states: $P_{ij}, i, j = 1, \ldots, K$. According to this model we treat a DNA sequence as a realization of the process of transition between states (Markov chain dynamics).

Several attempts have been made to establish the numerical parameters of an ordinary (homogeneous) Markov chain model so that it would best describe the natural DNA. Finally, the ultimate result was not obtained. It was found that the DNA composition and features of nucleotide correlation vary among different species.

One of the most important results is that the model of DNA protein-coding region does not belong to the class of homogeneous Markov chains, but to the class of periodic non-homogeneous Markov chains (Borodovsky et al., 1986b). This class of non-homogeneous Markov chain models is described by three vectors of initial probabilities: P_0^1, P_0^2, P_0^3; and three matrices P^1, P^2, P^3 that contain transition probabilities P_{ij}^1, P_{ij}^2, P_{ij}^3, $i, j = 1, \ldots, 4$.

For what follows it is necessary to determine the appropriate Markov model for non-coding DNA as well. The homogeneous Markov chain model was found quite appropriate for this purpose.

The numerical parameters, for the first order Markov chain model of a nucleotide sequence are defined according to statistics of mono- and dinucleotides taken from the training sets of DNA sequences taken from a given species.

In the case of a homogeneous Markov chain (training set of non-coding regions) the values of the initial probabilities are accepted to be equal to the normalized frequencies of the mononucleotides. The values of the elements in the transition matrices $P_{ij} = P(j|i)$ are assumed to equal the ratio $N(ij)/N(i)$ where i, j stand for types of nucleotides, $N(i)$ is the count of the nucleotide i, $N(ij)$ is the count of the dinucleotide ij. These results follow from the principle of maximum likelihood if one uses the Markov process realization for determining of the best-fit Markov chain model parameters. In the training set of coding regions statistics $N(i)$, and $N(ij)$ are calculated by separate counting the frequencies of mono- and dinucleotides in each of three "frames". It means that mononucleotides are counted separately in each position of codon and the dinucleotides also are split into three groups according to the codon position occupied by the left nucleotide.

The fundamental formula from Markov chain theory defines the probability that a given sequence of states (nucleotides) $i_1, i_2, \ldots, i_k$ can be observed in the realization of the Markov chain process starting from a given position in a sequence as

$$P(i_1, i_2, \ldots, i_k) = P_0(i_1)P(i_2|i_1)P(i_3|i_2)\ldots P(i_k|i_{k-1}). \tag{1}$$

DNA sequences in the training sets consist of four nucleotide symbols: T, C, A, and G (denoted here by $1, 2, 3, 4$). Our next step is to generalize formula (1) for the case when other types of nucleotides, representing ambiguities, are present in the fragment $i_1, i_2, \ldots, i_k$. Additional types of symbols are: Y (C or T), R (A or G), K (G or T), S (G or C), W (A or T), H (A or C or T), B (G or T or C), V (G or C or A), D (G or A or T), N (T or C or A or G).

In what follows it is convenient to refer to a *generalized symbol* X (x_1 or x_2 or x_3 or x_4), that means one of the ambiguity types listed above. $|X|$ denotes the number of nucleotides belonging to the generalized symbol X. The transition probabilities for the Markov chain with 15 generalized types of nucleotides are defined by the following formulas:

$$P(X|i_k) = \sum_j P(x_j|i_k),$$

$$P(i_k|X) = \sum_j P(i_k|x_j)P(x_j),$$

$$P(Y|X) = \sum_j P(y_j|X),$$

where X and Y are two generalized symbols.

Formula (1) can be used now for a sequence that contains all 15 nucleotide symbols (including ambiguity symbols). A similar formula can be simply written for a non-homogeneous Markov chain as well.

The numerical values of the initial probabilities as well as values of transition elements of homogeneous and non-homogeneous Markov chain matrices for *E. coli* DNA were determined in (Borodovsky et al., 1986b), where these data are considered in detail.

The Bayesian approach to gene recognition and its accuracy

Markov models for coding and non-coding regions of DNA described above supply us with, speaking metaphorically, statistical images of DNA sequences with specific functional meaning.

The problem of prediction of a protein-coding region within a newly sequenced DNA may be posed as a problem of search for a fragment (with a priori unknown boundaries) which would be close to a fragment generated by the Markov model of the coding region in a sense of some appropriate statistical measure.

In practice the problem is simplified by considering the set of fragments of a given length which cover the entire sequence. The fragments are analyzed one at a time, using the so called *moving window* technique.

Denote a nucleotide sequence fragment $j_1, j_2, \ldots, j_n$ by F. It is convenient to let n be a multiple of 3. We need several formulas for computation of auxiliary statistics based of the already defined Markov models. First the conditional probability that F belongs to a non-coding region is computed as

$$P(F|\text{NON}) = P_0(j_1)P(j_2|f_1)\ldots P(j_n|f_{n-1}). \tag{2}$$

The computation of the probability $P(F|\text{COD})$ that the fragment F belongs to a coding region is slightly more complicated. Actually, we can split this event into three mutually exclusive cases with the first nucleotide of the fragment F occupying the first, the second or the third position of a codon. Formula of the same type as (2) is applied in each case separately.

$$P(F|\text{COD}^1) = P_0^1(j_1)P^1(j_2|j_1)P^2(j_3|j_2)P^3(j_4|j_3)\ldots P^2(j_n|j_{n-1}),$$

$$P(F|\text{COD}^2) = P_0^2(j_1)P^2(j_2|j_1)P^3(j_3|j_2)P^1(j_4|j_3)\ldots P^3(j_n|j_{n-1}),$$

$$P(F|\text{COD}^3) = P_0^3(j_1)P^3(j_2|j_1)P^1(j_3|j_2)P^2(j_4|j_3)\ldots P^1(j_n|j_{n-1}). \tag{3}$$

The sum of the components $P(F|\text{COD}^1)$, $P(F|\text{COD}^2)$, and $P(F|\text{COD}^3)$ gives us the value of the probability $P(F|\text{COD})$.

Now we are close to our main objective which is to determine the value $P(\text{COD}|F)$ (or $P(\text{NON}|F)$, what is the same, since the sum $P(\text{COD}|F) + P(\text{NON}|F)$ is presumably 1). The designation $P(\text{COD}|F)$ (or $P(\text{NON}|F)$) stands for *a posteriori* probabilities of the event that the fragment F belongs to a coding (resp., non-coding) region given the sequence of F.

Table 1
The percentage of false negatives (coding predicted to be non-coding) in the artificial sample with the threshold $H =$ 0.75.

	48	96
0	19.4%	7.7%
10	24.1%	11.9%

Table 2
The percentage of false positives (non-coding predicted to be coding) in the artificial sample with the threshold $H =$ 0.75.

	48	96
0	15.5%	6.2%
10	16.5%	6.9%

Three components of the value $P(\mathrm{COD}|F)$ can be computed by the Bayes formula ($m = 1,2,3$)

$$P(\mathrm{COD}^m|F) = \frac{P(F|\mathrm{COD}^m)}{\sum_{j=1}^{3} P(F|\mathrm{COD}^j)P(\mathrm{COD}^j) + P(F|\mathrm{NON})P(\mathrm{NON})}, \tag{4}$$

The designations $P(\mathrm{COD}^i)$ and $P(\mathrm{NON})$ stand here for the so called *a priori* probabilities of the events COD^i, $i = 1,2,3$ and NON. The events COD^i happen when an arbitrary (unknown) fragment F is located within a coding DNA (with specified position of the first nucleotide in the codon), while the event NON happens if F is within the non-coding part. The simplest assumption is that $P(\mathrm{NON}) = 1/2$ and $P(\mathrm{COD}^i) = 1/6$ for $i = 1,2,3$.

Formula (4) gives a basic idea for the algorithm that would determine the values $P(\mathrm{COD}^i|F)$, $j = 1,2,3$ for any finite fragment of a DNA text.

The accuracy of the method has been tested using artificial nucleotide sequences. The Markov chain sequence generator has been designed for producing 15-letter alphabet sequences of coding and non-coding types (depending on whether the homogeneous or non-homogeneous Markov chain model has been used). The percentage of ambiguities could be varied. We have considered 0% and 10% ambiguity levels. Then the artificial sequence produced by a generator was divided into non-overlapping fragments — windows. Two different window sizes (48 or 96 nucleotides) were considered. Prediction of the protein coding capacity has been done for each window fragment according to the value $P(\mathrm{COD}|F)$ for this fragment. A threshold value H was established for the decision making. It means that if $P(\mathrm{COD}|F) \geq H$, F is predicted to belong to a coding region, while if $P(\mathrm{COD}|F) < H$, F is predicted to be non-coding.

The values of the type I error of the coding region prediction (coding recognized as non-coding) for the threshold value 0.75 are shown in Table 1.

Table 2 presents the type I error levels for the non-coding region prediction (non-coding recognized as coding) when decision making threshold again equals 0.75.

It is seen that prediction of sequences with ambiguities produces worse results and that the error of prediction of the non-coding regions is less sensitive to the presence of ambiguities. The enlargement of the window size allows us to achieve better reliability, a fact that could be expected, since a longer sequence fragment contains more information about the functional type of the sequence.

Evaluation of the method accuracy using the control set of *E. coli* DNA Sequences

A sample consisting of 425 *E. coli* sequences taken from EMBL Release 25 was used in order to estimate the method reliability in the case of natural DNA sequences. The levels of the type I errors made by predictions for coding region are presented in Table 3, while those for non-coding regions, in Table 4 ($H =$ 0.75). The histograms are presented on Fig. 1.

Figs. 2 and 3 present predictions made for a recently sequenced fragment of *E. coli* DNA, provided by F. Blattner (University of Wisconsin). The sequence region contains three unidentified open reading

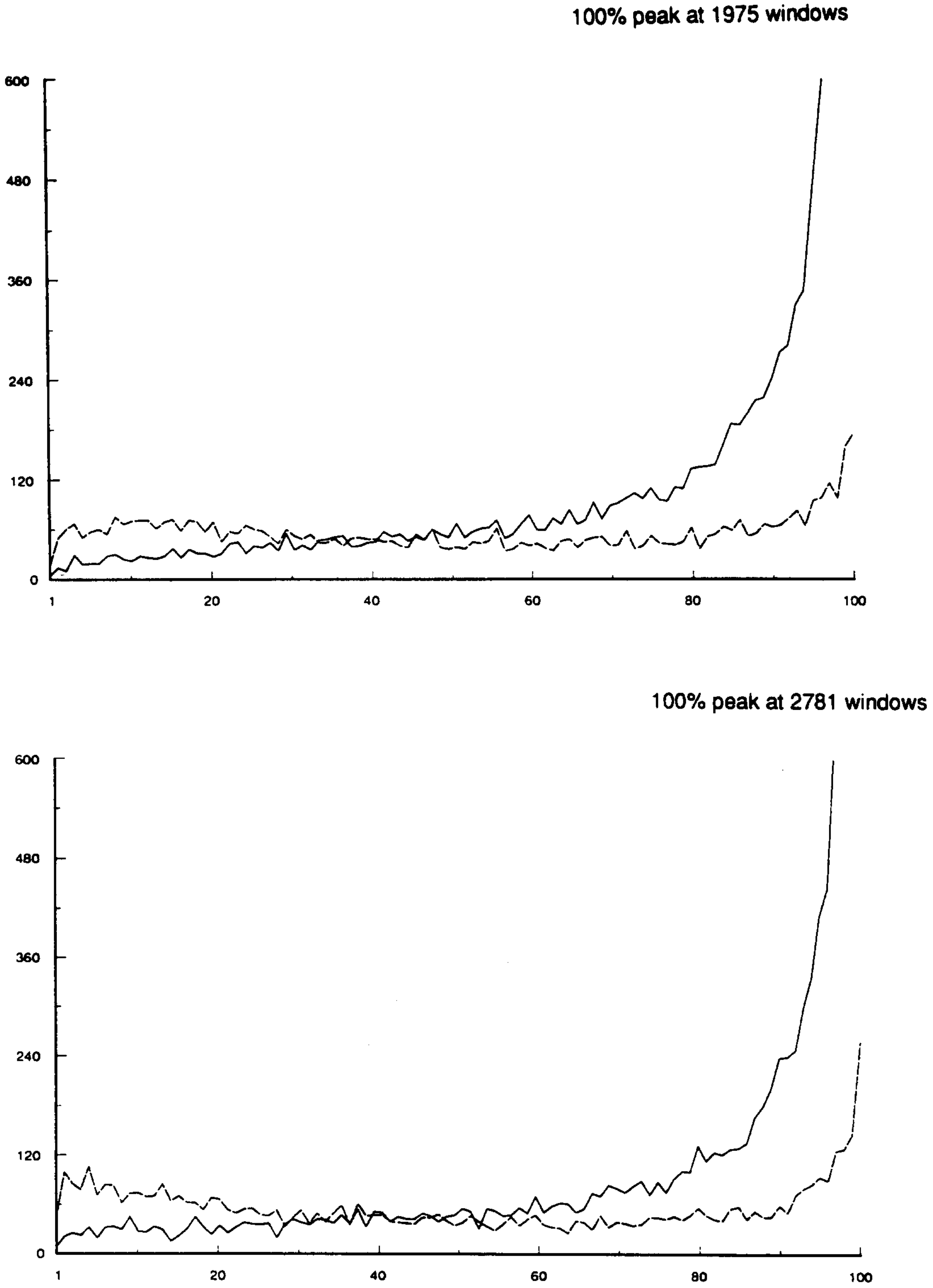

Fig. 1. Histograms of the probability function for two sets of 12,914 coding and 5,628 non-coding 48-nucleotide fragments of *E. coli* DNA. Solid line — the number of coding fragments which produced the given value of the probability function, dashed lines — same for non-coding fragments. (A) No ambiguous nucleotides. (B) The data with 10% of artificially introduced ambiguous nucleotides.

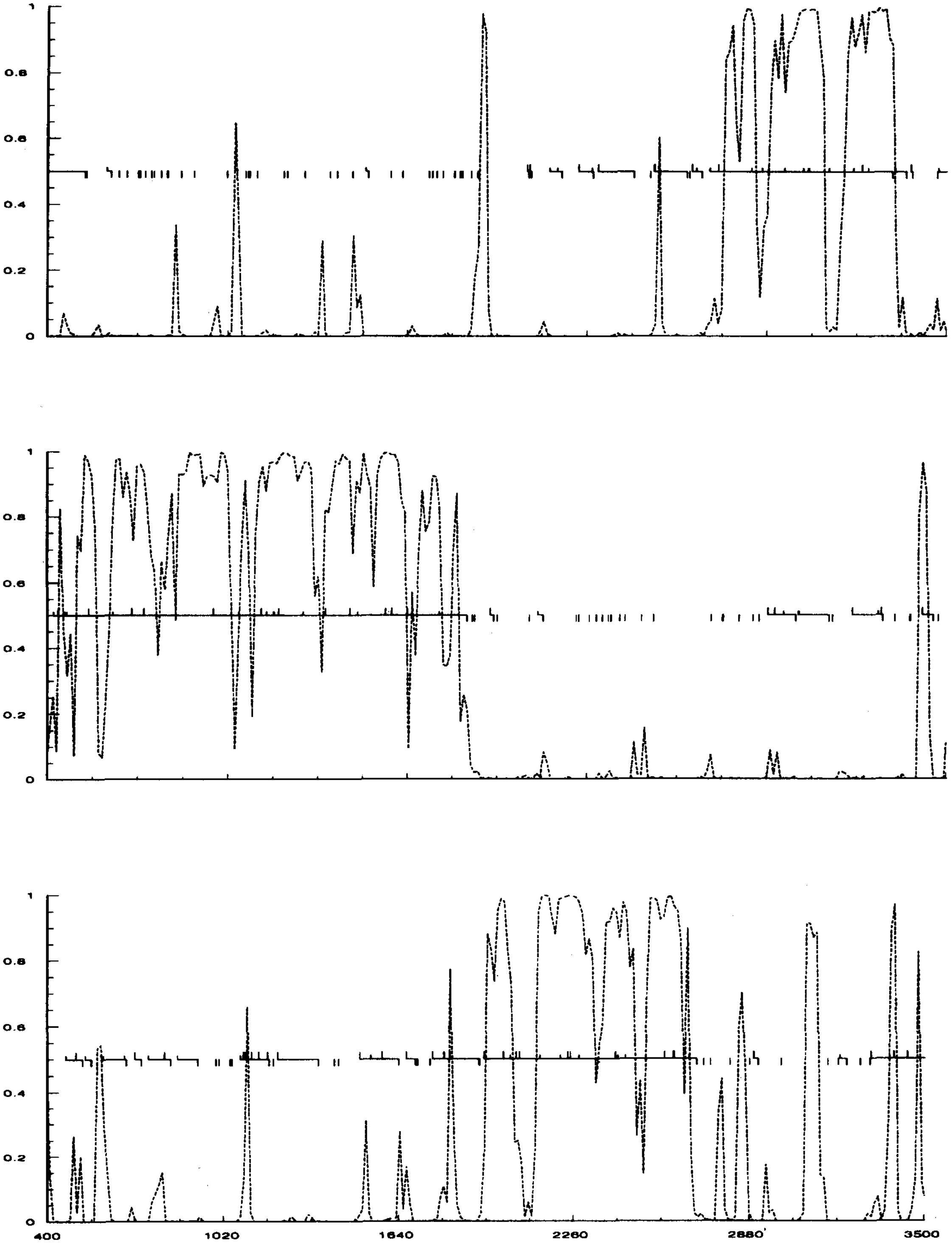

Fig. 2. Probability values as a function of the moving window position for a *E. coli* DNA. First order Markov chain model has been used. Window length is 48 nucleotides. (A) No ambiguous nucleotides. (B) The data with 10% of artificially introduced ambiguous nucleotides.

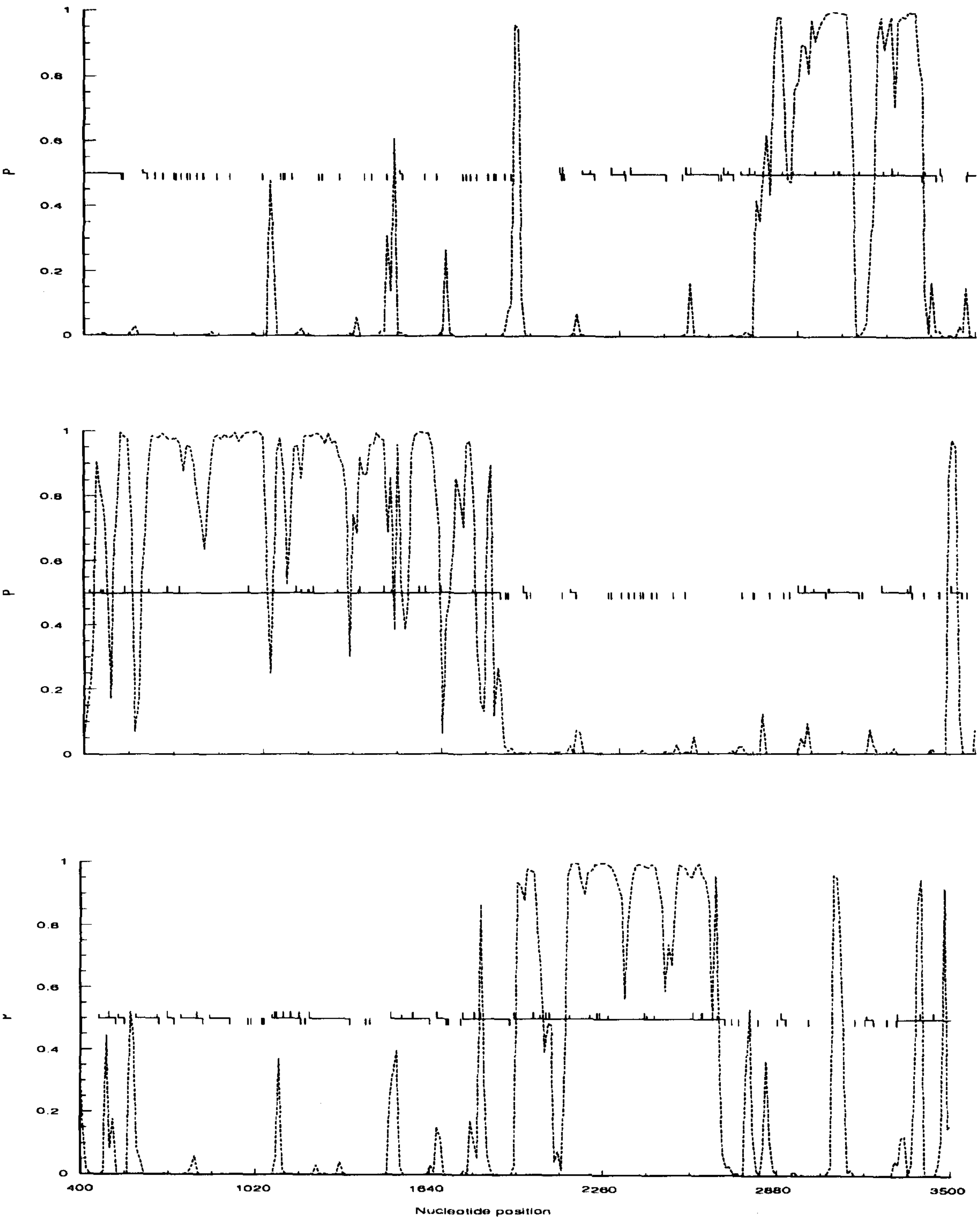

Fig. 2 — continued.

Table 3
The percentage of false negatives (coding predicted to be non-coding) in the *E. coli* sample with the threshold $H =$ 0.75.

	48	96
0	24.9%	17.4%
10	27.5%	19.8%

Table 4
The percentage of false negatives (coding predicted to be non-coding) in the *E. coli* sample with the threshold $H =$ 0.75.

	48	96
0	32.8%	28.0%
10	32.4%	27.4%

frames: from position 422 to position 1846 in the second frame, from 1944 to 2696 in the third frame and from 2713 to 3312 in the first frame.

The transition to the better quality of the gene regions identification can be observed when the percentage of ambiguities decreases and the window length increases.

Discussion

The numerical results obtained in the testing of the control set of sequences generated by the Markov chain generator indicate that sufficient accuracy can be achieved even in the case of a window of 48 nucleotides wide (Tables 1 and 2).

The reliability of the prediction drops insignificantly (in a 5% range) if the relative number of ambiguities does not exceed 10%. Note that rough DNA sequences that come from an experiment usually contain 1–3% of ambiguities.

The results from the experiment with artificial sequences can be considered as on optimistic estimate of the real accuracy of the method. More realistic estimate of the method accuracy is determined by application of the method to natural DNA sequences (Tables 3 and 4). Here one can see that the false negative error rate (coding as non-coding) increases by 2–8%, and the false positive error rate (non-coding as coding) increases by 18–20%.

Also, we have observed two phenomena that should be explained. The first one is that the prediction error that has been found for synthetic sequences for the threshold 0.75 is not close to the value 0.75. That is caused by the fact that the Bayes probability 0.75 determined for a fragment does not correspond to the portion of all fragments under consideration having the value of the Bayes function less than 0.75. The same statement is valid for all other values of the Bayes probability if one is comparing any local Bayes probability taken as a threshold value with the total amount of the prediction errors.

Another question will arise if one notices that the false positive error defined in the presence of ambiguities happens to be smaller than the false positive error rate determined for the sequence that does not have ambiguities (Table 4). The explanation is that the distribution of the Bayes probability on the set of the non-coding fragments has a specific shape. As soon as the percentage of ambiguities increases, the distribution begins to shrink to the middle of the region (dashed lines on Fig. 1A and 1B), so the portion of the fragments with the Bayes probability exceeding 0.75 decreases instead of expected increasing.

As a whole, the present model demonstrates that a 32-codon window is better for practical purposes than a 16-codon window. The former will give at least 80.2% reliability for prediction of coding and 72.6% reliability for prediction of non-coding (for isolated DNA fragments with the ambiguity level 10%). Note that the use of isolated fragments for estimates of the error rate produces a "pessimistic" judgment about the reliability of the method. In practice, when the results of the analysis of the adjacent DNA fragments are combined together in the process of the decision making, the random fluctuations are suppressed and the final error rate decreases.

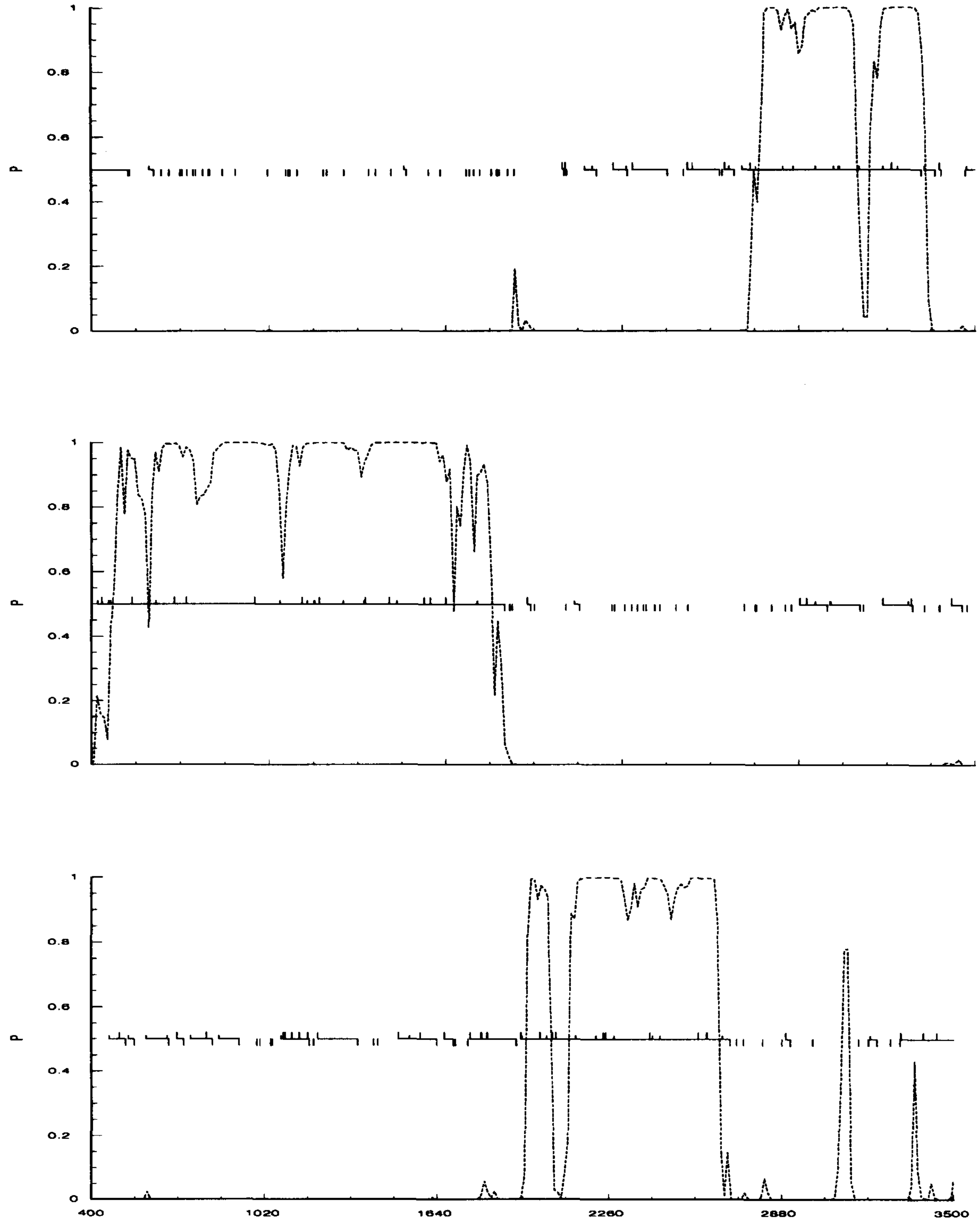

Fig. 3. Probability values as a function of the moving window position for a *E. coli* DNA. First order Markov chain model has been used. Window length is 96 nucleotides. (A) No ambiguous nucleotides. (B) The data with 10% of artificially introduced ambiguous nucleotides.

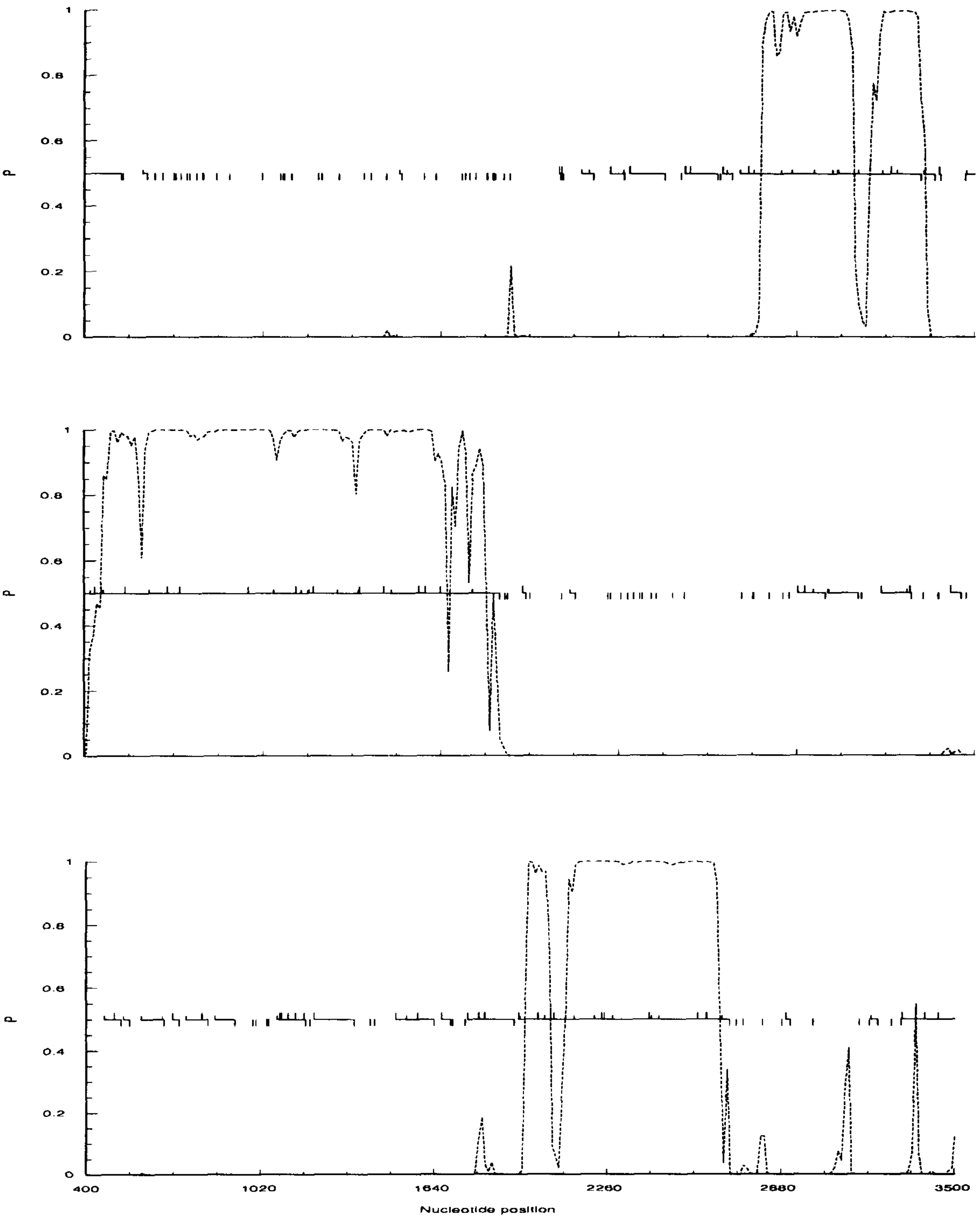

Fig. 3 — continued.

Further generalizations of the method for the case of higher order Markov chain models and simultaneous analysis of two complementary DNA strands are presented in (Borodovsky and McIninch, 1993ab).

References

Bishop, M.J. and Rawlings, C.J. (eds.), Nucleic Acid and Protein Sequence Analysis: A Practical Approach, 1987. (IRL Press, Oxford, Washington, D.C.).

Borodovsky, M., Sprizhitsky, Yu., Golovanov, E. and Alexandrov, A., 1986a, Statistical features in the *E. coli* genome primary structure. II Non-stationary Markov chains. Molecular Biology 20, 833–840.

Borodovsky, M., Sprizhitsky, Yu., Golovanov, E. and Alexandrov, A., 1986b, Statistical features in the *E. coli* genome primary structure. III Computer recognition of protein-coding regions. Molecular Biology 20, 1144–1150.

Borodovsky, M. and McIninch, J., 1993a, Prediction of gene locations using DNA Markov chain models. Proceedings of the Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis (St. Petersburg, FL) (to appear).

Borodovsky, M. and McIninch, J., 1993b, Parallel gene recognition for both DNA strands, Computers & Chemistry (to appear).

Fickett, J.W., 1982, Recognition of protein coding regions in DNA sequences. Nucl. Acids Res. 10, 5303–5318.

Gelfand, M.S., 1992, Computer functional analysis of nucleotide sequences: problems and approaches, in: Mathematical Methods of the Analysis of Biopolymer Sequences (DIMACS, vol. 8), S.G. Gindikin (ed.) (AMS, Providence, RI), pp. 19–62.

Gribskov, M., Devereaux, J. and Burgess, R.R., 1984, The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucl. Acids Res. 12, 539–549.

Konopka A. and Owens J., 1990, Complexity charts can be used to map functional domains in DNA. Gene Anal. Techn. Appl. 7, 35–38.

Shepherd, J.C.W., 1981, Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justifications. Proc. Natl. Acad. Sci. USA 78, 1596–1600.

Shulman, M.J., Steinberg, C.M. and Westmoreland, N., 1981, The coding function of nucleotide sequence can be discerned by statistical analysis. J. Theor. Biol. 88, 409–420.

Staden, R. and McLachlan, A.D., 1982, Codon preference and its use in identifying of the protein coding regions in DNA sequences. Nucl. Acids Res. 10, 1541–1563.

Staden, R., 1984, Computer methods to locate signals in nucleic acid sequences. Nucl. Acids Res. 12, 552–567.