Bayesian Approach to DNA Segmentation into Regions with Different Average Nucleotide Composition

Vsevolod Makeev[1], Vasily Ramensky[1], Mikhail Gelfand[2], Mikhail Roytberg[3], and Vladimir Tumanyan[1]

[1]Engelhardt Institute of Molecular Biology, Moscow, 117984, Russia
{makeev,ramensky,tuman}@imb.ac.ru
[2]VNIIGENETIKA, Moscow, Russia
mgelfand@ntl.ru
[3]Institute of Mathematical Problems of Biology, Puschino, Moscow Region, Russia
roytberg@impb.psn.ru

Abstract. We present a new method of segmentation of nucleotide sequences into regions with different average composition. The sequence is modelled as a series of segments; within each segment the sequence is considered as a random sequence of independent and identically distributed variables. The partition algorithm includes two stages. In the first stage the optimal partition is found, which maximises the overall product of marginal likelihoods calculated for each segment. To prevent segmentation into short segments, the border insertion penalty may be introduced. In the next stage segments with close compositions are merged. Filtration is performed with the help of partition function calculated for all possible subsets of boundaries that belong to the optimal partition. The long sequences can be segmented by dividing sequences and segmenting those parts separately. The contextual effects of repeats, genes and other genomic elements are readily visualised.

1 Introduction

1.1 Biological Motivation

Local nucleotide composition is believed important for many biological issues [1], [2] such as the isochoric organisation of the genome of higher eukaryotes [3], [4] compositional differences between exons and introns [5], [6], simple repeats (e.g. [7], tracts in splice sites [8] and binding sites [9] of DNA, GC islands in promoter sequences [10] and many others. Moreover, local nucleotide composition is accounted for in many algorithms developed to search different patterns in DNA sequences [11]. Usually the fixed length window is used and the results may undesirably depend on the length of the window.

O. Gascuel, M.-F. Sagot (Eds.): JOBIM 2000, LNCS 2066, pp. 57–73, 2001.

1.2 Current Algorithms for Compositional Segmentation

Basically our approach is similar to that of [12], see also [13], [14]. The main difference is that Liu and Lawrence use weights for configurations with different number of segments, favoring segmentations to longer segments. Instead, we use the two-stage procedure with filtration of boundaries, which allows us to study segments with the chosen length-scale. By refusing to use weights we avoid an approximate procedure of sampling and can profit from the faster implementation of dynamic programming technique (N^2 instead of N^3, where N is the overall length of the sequence).

Another approach is developed in [15]. It uses the traditional frequency count estimator to which the Bayesian estimators converge for large segments. This approach is less justified for small segments. Recently several algorithms appeared employing hidden Markov models to obtain the segmentation of nucleotide sequences into segments with different composition [16], [17], [18]. However, in these models the number of possible compositional states usually is set *a priori*. This assumption works well when some particular DNA segments are searched for (for instance GpC islands [18]. At the same time this approach is less justified for partition of new genomic sequences, with no special attention paid to any particular region.

2 Optimal Segmentation

2.1 Probabilistic Formulation

A symbolic sequence over an alphabet W of L letters is considered as a series of segments, each segment is a Bernoulli type random sequence. Each segment has the corresponding symbol counts vector $\mathbf{n} = (n_1,\ldots n_L)$, where n_j is a number of occurrences of the j-th symbol in the segment; **n** has the multinomial distribution.

The Bayesian approach we are using [16], [19] regards the estimated parameters as random variables. In the beginning these variables have some prior probability distribution, which may be chosen rather arbitrarily. These probability distributions are re-estimated from the data using the Bayes formula, and the posterior distribution is obtained (see formula (4) below). The results of Bayesian estimation are always some probability distributions of the estimated quantity. Bayesian and classical statistics, however, agree for large samples because Bayesian distributions converge to the maximal likelihood estimation for any reasonable prior [20].

2.2 Choosing the Prior

We addressed the issue of choosing the appropriate prior in detail in [21]. The success of the Bayesian techniques depends dramatically on the prior used, especially for small samples, and the proper choice of the prior should be conditioned by the context of the problem, as there are no formal recipes.

For the segmentation problem, the choice of the prior actually implies some ideas about the overall composition of polymer. The simplest choice is the Dirichlet prior [12], [17], which reflects *a priori* information on the sequence composition. A more complex is the Dirichlet mixture prior [11], [17], which is based on the idea that a segment composition can come from one of several compositional classes. One can also use the entropic prior [19] reflects the statistical homogeneity of the prior source data.

The Dirichlet prior allows pseudocount interpretation [12], according to which the additional "pseudocounts" are added to the observed counts for each segment. Thus such the prior can be included into the algorithm without substantial modification of the formulae. The Dirichlet prior introduces some *a priori* composition of the sequence, and sequence segments, the composition of which is in contrast with this *a priori* composition, are more likely extracted at the same significance level. Thus, by choosing the proper prior one can adjust the program to extraction of specific functional region with known composition.

However, we believe that the noninformative prior we use, is more suitable for the initial segmentation of newly sequenced genomes. In this case the problem is not searching for specific regions, but assessing the general structure of the sequence. In practice, to choose the prior some statistical observations are made on data banks, or on the composition of the same sequence averaged on a larger scale. Thus, some correlation of the sequence composition with other sequences or with other parts of the studied sequence is introduced into the statistical inference, which is not always desirable, when the sequence under study is entirely new. Informative priors are better fit to the problem of searching for specific regions with at least approximately known composition, such as codons, GC islands etc.

2.3 Segment Likelihood

Denote the set of letter probabilities (the segment composition) as $\sigma = (\theta_1,\ldots,\theta_L)$; it complies to the normalisation condition:

$$\sum_{k=1}^{L} \theta_k = 1 \tag{1}$$

The likelihood of the individual sequence to occur is

$$L(\sigma) = \prod_{k=1}^{L} \theta_k^{n_k} \tag{2}$$

Given the composition $\sigma = (\theta_1,\ldots,\theta_L)$, write the probability density function $p(\sigma)$, which is defined on the simplex $S = \left\{ \sigma : \theta_k \geq 0; \sum_{k=1}^{L} \theta_k = 1 \right\}$.This probability density should comply with the normalisation condition

$$\int_S d\sigma p(\sigma) = 1 \tag{3}$$

Given some prior distribution $p(\sigma)$, consider the tentative block with counts $\mathbf{n}$. The Bayes theorem brings about the estimated probability density function $p(\sigma|\mathbf{n})$:

$$p(\sigma|\mathbf{n}) = \frac{L(\mathbf{n}|\sigma)p(\sigma)}{P(\mathbf{n})} \tag{4}$$

where

$$P(\mathbf{n}) = \int_S d\sigma L(\mathbf{n}|\sigma)p(\sigma) \tag{5}$$

is the normalisation constant called the marginal likelihood [12].

Marginal likelihood reflects the overall probability of obtaining the given sequence in the two stage random process. First, the composition σ is picked up according to the prior distribution, and then the sequence is generated in the Bernoulli random process with the letter probabilities σ. If $p(\sigma)$ is the uniform distribution on the surface of the simplex S, then

$$P(\mathbf{n}) = \frac{(L-1)!}{(N+L-1)!} n_1 ! \ldots n_L ! \tag{6}$$

Surprisingly, this quantity is also obtained in the conceptually similar but different probabilistic model [22]. For the sequence with the length N, the overall numbers of each letter ($n_1,\ldots,n_L$; $\Sigma N_i = N$) are picked up from the uniform in this case discrete distribution, then the probability of obtaining the sequence in the shuffling procedure is calculated. Since we consider the segments as independent, the complete likelihood of the sequence segmentation into K segments with known boundary location writes

$$P = \prod_{k=1}^{K} P_k(\mathbf{n}_k) \tag{7}$$

This quantity is optimised over the set of all possible boundary configurations yielding the optimal segmentation.

2.4 Dynamic Programming

The maximisation algorithm is formulated as follows. Consider a sequence $S = s_1 s_2 s_3 \ldots s_N$ of length N, where $s_i \in \Omega$. For every segment $S(a,b) = s_a \ldots s_b$, $a \le b$ with the length N, we introduce the weight $W(a,b)$. In our case $W(a,b) = \ln P(S(a,b))$. Any particular segmentation R in m blocks is determined by a set of boundaries $R = \{ k_0 =$

0, k_1, …, k_{m-1}, k_m= *N* }, where k_i separates S_{k_i} and S_{k_i+1}; $k_0<k_1<\ldots<k_{m-1}<k_m$. Define the weight of segmentation *R* as

$$F(R)=\sum_{j=1}^{m}W\left(k_{j-1}+1,k_j\right) \tag{8}$$

For functions determined on the segmentations, we shall also use another set of variables, the indicators of the boundary positions q_k , k = 1,...,*N*. By definition, q_k = 1 if there exists a segment boundary after the letter *k*, otherwise zero. We shall use both variables, *F*(*R*) and $F(q_1,...,q_k)$, without special comments.

We are looking for the segmentation R^* that has the maximal weight. This is done in the recurrent manner. Denote by $R^*(k)$ the optimal segmentation of the fragment *S*(1,*k*), $1 \le k \le N$. It is trivial to find $R^*(1)$. In the case of known optimal segmentations $R^*(1)$,…,$R^*(k\text{-}1)$, the optimal segmentation $R^*(\mathrm{k})$ is found using the following recurrent expression

$$F\left(R^*(k)\right)=\max_{i=0,\ldots,k-1}\left[F\left(R^*(i)\right)+W\left(i+1,k\right)\right] \tag{9}$$

Here, we put $F(R^*(0)) = 0$. The recurrent relation (9) yields the algorithm. Since the building of segmentation $R^*(k)$ takes the time ~*k*, the total time can be estimated as N^2.

2.5 An Example

Figure 1 displays the segmentation of a 1000 bp long random sequence with uniform composition (all letter probabilities are equal to 0.25). One can see that usually it is segmented into very short segments. Surely this is not what we would like to obtain. Practically, the segments containing many identical letters are extracted, but the general homogeneity of the block is never exhibited.

It should be noted that if we used a different prior, we would obtain a different segmentation pattern. For instance, for the prior, which corresponds to the composition rich with A and T, borders separating segments with different numbers of ‘a’ and ‘t’ (such as a|tttt in the first line) are likely to disappear. Conversely, new boundaries can appear, as those separating segments with ‘g’ and ‘c’ (such as in the tgtt segment in the first line).

3. Border Insertion Penalties

3.1 Fluctuations in Local Composition

One can see (Fig. 1) that usually the segments of optimal segmentation are very short. Moreover, the random uniform Bernoulli sequence is also divided into many segments. When the sequence consists of several random homogeneous domains, the optimal segmentation includes many borders that are located within the domains.

Redundant boundaries are present due to statistical fluctuation in local composition in random sequences.

```
aaga| t | c| t gt t | aaca| g| a| t t t t | g| aa| t | g| c| a| t t t | g| a| c| g| cac| t ggt gt | c| t
| aa| g| t t | g| cc| t at | c| aa| c| t | gcg| t t | accaca| t t t | g| c| aga| c| gag| t t | c| a| c
gc| aa| g| t | c| g| t t | a| gg| t gt t t | gag| c| a| ct c| aa| c| g| t at t | c| gaaga| t t | c| g|
t ct | a| cgcg| t | g| c| aaa| cct cc| t | a| c| t | g| aa| t t | ggg| cc| g| t at at | gg| cc| g| a
a| c| t | g| t ct c| g| aca| t | g| t ct t ct t c| aaa| ccc| gagag| t | cc| aa| t | cc| t | accaca
cacc| g| aa| t gt t | a| g| cct c| gag| t | aa| c| agaga| c| t at aa| gg| t t | c| a| gg| t t t | c
c| g| t ct t | a| g| c| aa| ggg| t t t gt | c| t | gag| t | cccgc| g| t t | ggg| c| t | cgccg| aat a
| ccc| a| t | g| ct c| a| t | g| at a| ggagg| cc| ggg| c| t | a| gt gt t t g| c| a| gt gt | aa| t | c
| g| a| gt g| c| a| t t | c| a| g| t | gcgg| cc| t t | a| cc| aa| g| t t | g| cc| t gt t t | a| c| aaat
aa| gt g| aa| t | cc| gg| a| gcgcg| t t | c| g| t t | g| c| gag| c| t | cggcgc| t | gg| caacac|
gg| aa| g| t t | g| aaa| g| c| a| g| t t t | aga| t t | ccc| t t | cc| g| t at aa| gg| at aa| t t gt |
ccgcg| t at a| g| t ct c| a| g| t | cgcc| gt t gt | a| cgc| t t | c| a| g| ccc| t | a| c| t | a| c| g
g| t at | cc| aga| cgc| a| t t | g| ccct cc| g| t | ccgcg| at a| ggcgg| t ct t t ct | a| t | c| g|
a| t ct | g| c| a| gg| t | a| cgc| gggt gg| aaca| g| t t | aa| c| t | aa| t cccct ct c| a| c| t t |
aa| t t | aa| cggcgc| a| gg| aa| cc| g| aaaca| c| g| t t | ccgcgc| t gt t | c| aa| t t | g| a| t
| aga| ccc| t | aa| gg| t t | aga| t t | c| at a| gg| at aa| t t | gagaa| cgc| t t t t at at | acaa
| g| t t | g| aa| c| gg| a| t | ggggg| t t | aa| gg| c| t | gg| c| a| t t | g| cccgc| t t t ct t ct | a
| t | cgc| t | a| g| c| t t | aa| c| g| aa| ggg| a| t | ccac| t | gcgagcg| ccc| at a| gg| t | a| g
| aacaa| g| cc| ggg| aaa| g| c| gagag| c| t | a| c| t t | g| a| t t t t | c| a| t | ccc| t | a| c| g
| t t | g| c| aaaa| cgc| t | aaat a| t t t t | gcg| at a| g| t t t t t gt t t | aca| t | gggt ggg| c| a
| t t t ct | gg| c| t t | caca| g| at a| ggg| t ct | g| c| t | a| g| t t | ccccacccc| t | g| cac| at
t at a| cac| g| t | a| t gt | c| a| t | g| a| t | g| ccgc| t | a| gg| aa| cct cc
```

Fig. 1. The optimal segmentation of a random sequence. Segment boundaries are marked with "|".

Thus it is advantageous to separate boundaries, which separate long regions with different compositions from those that reflect statistical fluctuations. This can be done by penalizing those segmentations that contains more boundaries. The penalty β for insertion a new border can be readily incorporated into functional (7)

$$P = \prod_{k=1}^{K} P_k\left(\mathbf{n}_k\right)\beta^K \tag{10}$$

For β<1 segmentations that contains less boundaries would have a preference. In our programs we included parameter $B = -\ln\beta$. The optimal value of B is chosen from computer simulations.

3.2 Random Sequences

It should be noted that the procedure we are using to estimate B is very heuristic. As it was pointed out by the referee of this paper, this parameter can be interpreted like a transition probability to enter in a new region. Then the problem could be presented as identical to modeling the sequence with a Hidden Markov Model (HMM) where the hidden states are the regions with homogeneous composition. This allows one to use one of many algorithms to estimate the hidden state transition in a HMM.

Thus, the main objective of the tests presented below is demonstration, that the border insertion penalty is the powerful tool in extracting homogeneous segments with statistically different composition in the sequence. We demonstrate, that the dependence of the optimal *B* values on the segment length and composition is very weak, thus the same *B* values can be employed for segmenting of sequences of different biological origin. Good statistical tools would help to increase the model performance, but the relevance of the model is reliably demonstrated by our empirical calculations. In segmenting a random sequence, one would like to obtain the result which complies to certain requirements. The first of those implies that a homogeneous random sequence should be segmented as a single block. We performed several series of statistical tests with sequences of different length and composition. We tested sequences of 100, 1000, and 10,000 symbols with different compositional biases. For each length and composition a hundred of random sequences was generated, for each of which the minimal *B* providing segmentation into the single unit was found. For the sequences of the same length, a greater bias in composition usually implied a smaller critical *B* value. For strongly biased sequences there were observed examples of segmentation into the single unit with $B = 0$. Greater *B* values are usually required to remove the inner boundaries from the longer random sequences. The histograms of the critical *B* values for the sequences with uniform (all $p_i = 0.25$) composition for different lengths are shown in Fig. 2. The dependence of the critical *B* value upon both the sequence length and the compositional bias is remarkably weak. As a rule, *B* of about 3–5 is enough to provide segmentation of a random sequence into the single block (compare with Figs. 3, and 4 in the next section).

4 Block Random Sequences

When the *B* value is taken too large the boundaries between longer regions with a homogeneous composition are also removed. To limit *B* from the above we performed two series of tests. In the first series of tests sequences consisting of two random blocks were generated. The *B* value for which the single inner boundary was present in the maximum number of block-random sequences were found. Our calculations demonstrated that for the given difference between semi-blocks for small lengths the adequate segmentation is never found (for any *B*). With the increase in the length of the semi-block the adequate segmentation is found for some share of the random sequences. The number of such sequences increases rapidly with the length of the random semi-blocks. This can be explained from statistical reasons: larger sequences mean richer samplings for estimations. The best *B* value, for which the two-part segmentation is obtained, can be estimated by equalizing overestimation and underestimation errors.

Here, overestimation corresponds to the case when for the chosen *B* the sequence is segmented into the single block, whereas the same sequence is correctly segmented in two blocks for a smaller *B*; respectively, underestimation corresponds to the case when for the same *B* the sequence is segmented into more than two segments, and the adequate segmentation is achieved for some greater *B* value.

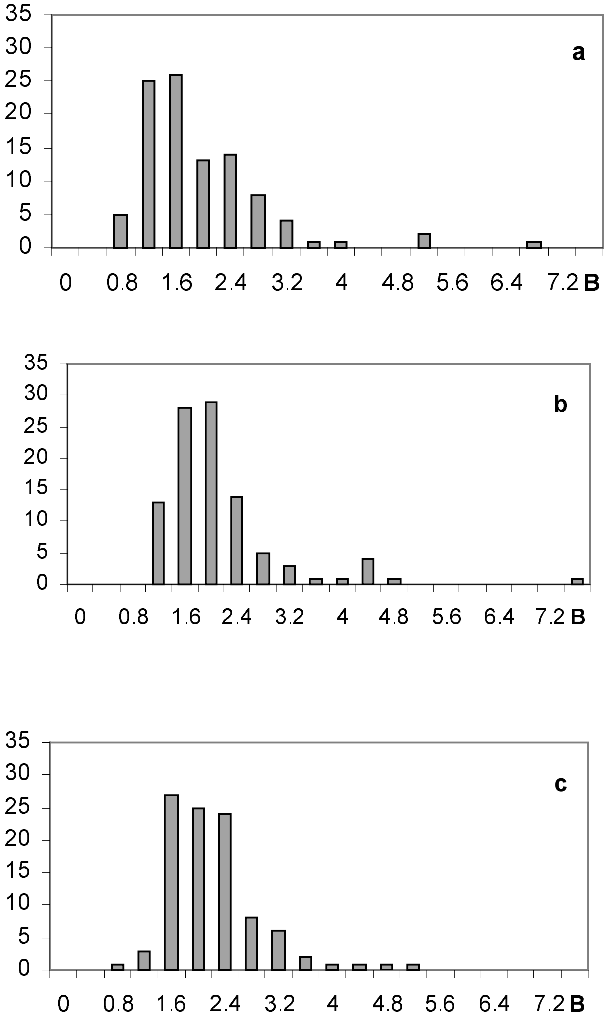

Fig. 2. Histogram of the minimal *B* necessary to segment a random homogeneous sequence into a single domain. 100 experiments. The uniform probability distribution. The sequence length is (*a*) 100bp; (b) 500 bp; (c) 1000bp

The optimal B value for 100 letter blocks with probabilities of (0.2, 0.2, 0.3, 0.3) and (0.3, 0.3, 0.2, 0.2) was about 2.7. From 100 random sequences tested, 60 were adequately segmented in two parts. Other 40 were not segmented into two parts for any B. The optimal value B = 2.7 was underestimation for six sequences and overestimation for seven ones. Thus, 45 sequences, a little less than a half, were adequately segmented. For B = 5, only two sequences were segmented in three blocks (both contained several identical letters on one of the ends), 25 of the sequences were segmented adequately with one boundary in the middle, and 75% sequences were segmented as a single unit.

However, if the bottom limit of B increases very slowly with the length, the upper limit increases dramatically faster. For instance, all the sequences made from two 1000–letter random blocks with the same probabilities as in the previous example are segmented in two blocks for any B from the range from 7.2 to 23.5 with the bottom limit for B less than 5 for 96 examples and less than 3 for 92 examples out of 100. This allows one to choose B rather comfortably when the short domains are not very important.

Obviously, the longer the segments, the smaller the difference, which the method is possible to resolve. The sequence made of two 1000 blocks with compositions (0.24, 0.24, 0.26, 0.26) and (0.26, 0.26, 0.24, 0.24) is almost never (in <10% of all cases) segmented adequately. However, the single boundary in the sequence with the same (0.24 vs. 0.26) block probabilities can be reliably resolved for the segment length of 10,000. The error for the boundary location also grows with the decrease in the compositional difference.

Similar results were obtained in the experiments on island recognition (Figs. 3 and 4). Random sequences were generated and in each experiment islands of contrasting compositions were put at random into the uniform sequence with the length of 1000, 2000, and 5000 length. The island length amounted to 0.1 of the total sequence length. We scanned over all BIPs with the 0.5 step. We put that the island was recognized if the strictly two inner boundaries were found in the sequence allowing for an error of 10% of island length. The example of the dependence on the BIP is shown in Fig. 3.

One can see, that if the compositional contrast is significant enough, to provide an acceptable recognition of the segment, the *B* parameter can be picked up rather comfortably. Fig. 4 shows the example of trade off between the compositional difference and the length of the segment.

In the next series of tests we generated sequences consisting of 10 blocks with different compositions and determined the *B* values for which the correct number of blocks was obtained. Then, we monitored the positions of boundaries in the segmentation with this optimal *B*. For the 100–letter blocks some false-positive boundaries (often near the ends of the sequence) were added, and some blocks with close compositions were merged. The chance of block merging is much greater for the blocks whose composition is close to the uniform (p_i = 0.25 for all *i*s). Again the percentage of correctly segmented sequences was greater for the sequences consisting of the longer blocks. In all examples of 10,000–letter blocks segmented with $B = 3.5$ we obtained the adequate segmentation.

All in all, it seems that choosing *B* between 3 and 5 allows one to eliminate most of fluctuations and consider the domains obtained as random with a fixed composition. For $B >> 5$, almost only the adequate boundaries are left, but some presumably divergent blocks are merged.

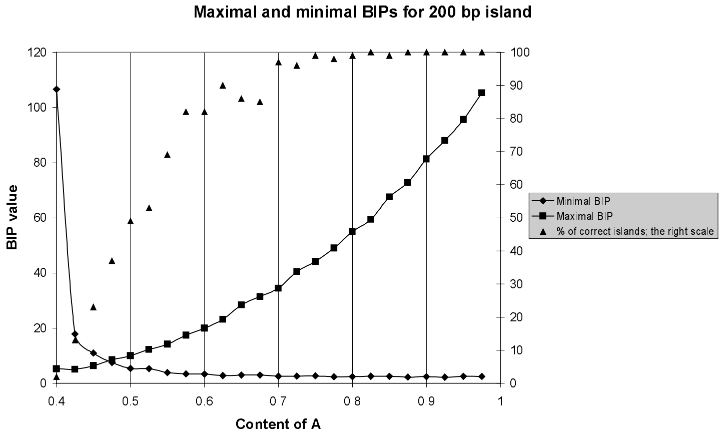

Fig. 3. Searching for an island of contrasting composition within a long sequence. The island with the length 200bp is located at the random place in the random sequence of the uniform composition with the length 2000 bp. Along the X-axis is the A-content of the island, the other three letters have the identical content. For each composition scanning by the BIP value is performed from bottom to top. The minimal BIP is the average BIP (averaged over 100 experiments) at which the island is correctly recognised (if any). The maximal BIP is the average BIP corresponding to the limiting value, above which the sequence is segmented into the single segment. The right scale shows the percentage of sequences for which the island was correctly recognised allowing for 20bp error in the boundary positions.

5 Filtration of Boundaries

5.1 Partition Function

Another way to study the relative significance of a boundary is the partition function. With the help of this function one can calculate a score, which reflects how the addition of this particular boundary influences weights of segmentations.

Given the probability of each segmentation, the partition function of the segmentations can be determined a standard way [23] by summing the probabilities of all possible partitions:

$$Z(N) = \sum_{q_1} \dots \sum_{q_{N-1}} \Pi(q_1, \dots q_{N-1}) \tag{11}$$

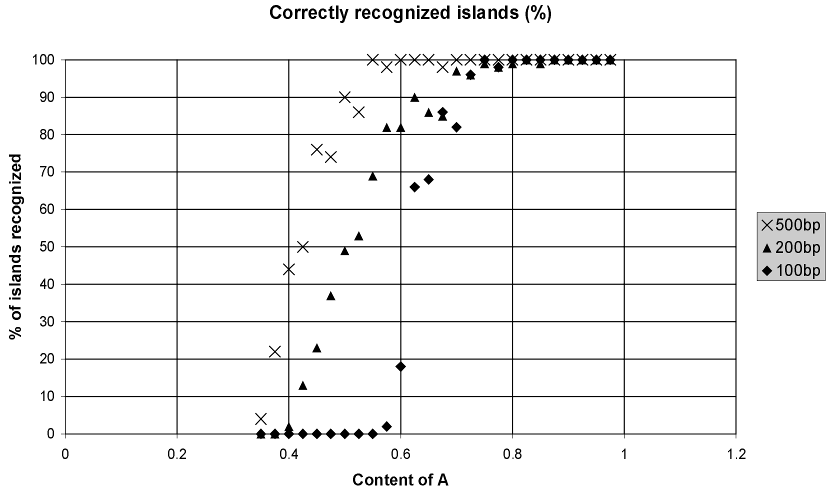

Fig. 4 Dependence of the percentage of correctly recognized "islands" on the island" composition and the island length. The islands of the lengths of 100, 200, and 500 bp were placed at random into the sequence with the uniform composition, the length of the islands was always 0.1 of the total sequence length. Along the X-axis is the A-content of the island, the other three letters have the identical content. *B* =5 in all experiments.

here q_k equals unity if there is a segment boundary after the letter k in the sequence, zero otherwise; the overall $\mathbf{q} = (q_1,\dots,q_N)$ determines a segmentation which has the probability $\Pi(\mathbf{q})$.

With the partition function at hand we can calculate the probability of a boundary to be located after a particular letter k, via calculating two partition functions of the regions to the left and to the right of this border, *ZL* and *ZR* respectively:

$$\Pi(k) = \frac{ZL(k)ZR(N-k)}{Z(N)} \qquad (12)$$

5.2 Dynamic Programming

The partition function (11) may be rewritten as follows [23]:

$$Z(N) = \sum_{q_1}\dots\sum_{q_{N-1}} e^{F(q_1,\dots q_{N-1})} \qquad (13)$$

To calculate the probability of a boundary after the letter k, we need also the partition functions of the segments to the left and to the right of this boundary:

$$ZL(k)=\sum_{q_1}\ldots\sum_{q_{k-1}} e^{F(q_1,\ldots,q_{k-1})} \tag{14}$$

$$ZR(k)=\sum_{q_k}\ldots\sum_{q_{N-1}} e^{F(q_k,\ldots,q_{N-1})} \tag{15}$$

Recurrent formulae to calculate *ZL*(*k*) and *ZR*(*k*) are analogous to (10) and are obtained through the formal substitution of operations. Summation is used instead of taking the maximum, and multiplication is used instead of summation [23]. Then the following relations replace (10):

$$ZL(k)=\sum_{j=0}^{k-1} e^{W(j+1,k-1)} ZL(j) \tag{16}$$

$$ZR(k)=\sum_{j=k}^{N} e^{W(k,j)} ZR(j) \tag{17}$$

with the respective boundary conditions *ZL*(0) = *ZR*(*N*+1) = 1; *W*(*k*-1, *k*) = *W*(*N*, *N*+1) = 0.

The obvious modification of the dynamic programming calculates the partition function in the case when only the given set of boundaries is allowed.

5.3 Filtration Strategy

For the best result one should combine calculation of optimal segmentation with filtration. At the first stage the optimal segmentation with some *B* is found. Then the cutoff value is chosen and all the boundaries with probabilities (11) lower than that cutoff value are removed. The resulting set of boundaries usually is not optimal in the sense that some boundaries from the resulting set are removed when the step one is repeated. So an additional round of optimisation (with the same *B*) is performed, removing some boundaries. Iterations converge rapidly to the stable set of boundaries all of which have the partition function probabilities greater than the cutoff value.

5.4 Partition Function Cutoff vs. Border Insertion Penalties

Both partition function cutoffs and border insertion penalties serve for the same purpose: to merge segments with close composition absorbing local fluctuation of composition. However, two segmentations into the same number of domains obtained via these two methods are not the same. Segmentation with given *B* is the best segmentation from all the segmentations with the fixed number of domains (this is clearly seen by taking log in (7)). Thus if one obtains two segmentations with the same number of domains, one using *B* and the other using some filtration cutoff level, then the segmentation obtained via border insertion penalty would have a greater score. However, the difference is not the critical and in general two segmentations agree.

5.5 An Example of the Large Scale Segmentation

We have segmented first 200,000 bp of the complete sequence of *Plasmodium falciparum* chromosome II available at www.tigr.org. This sequence contains a telomeric region, a long subtelomeric repeat and several genes, which consist of exons with various length. This genome is rich with A+T (in average 80%). In contrast because a gene should code for the protein its AT content cannot be to high, since some codons contain G and C in the significant positions of the triplet. Thus one can hope that some long segments in this sequence are related to the coding sequences.

To evaluate this hypothesis, 200,000 bp sequence was segmented with $B = 3$. Filtration with 0.999 cutoff level was performed. All the segments longer than 500 with the G+C content greater than 0.2 were taken.

From the chromosome description all exons with the length greater than 200 were taken. The results of the comparison are shown in the Table. The telomeric repeat and the long 21999 bp subtelomeric repeat found in this sequence is clearly seen. Among remaining 41 long GC containing segments 30 coincide with long exons with different precision (marked with ‘‘y”). In one case (177,844) the segment contains two exons. In one case (149,594) three subsequent segments cover one long exon. Seven long exons hadn't any corresponding segments. These are false negative of our ‘‘prediction”.

There are five examples, when long GC containing segments did not contain long exons. We searched for long ORFs in such segments. Indeed, three segments out of five contained long ORFs (starting at 27,864; 112521; and 192,716). These are good candidates for the new-found genes. Two segments (at 23,281 and 171,499) did not contain long ORFs. These are false positives of our ‘‘prediction”. One can see that our simple model of a long exon (long GC segment containing long ORF) describes surprisingly well the situation found in *Plasmodium falciparum* chromosome II.

The fact that the coding regions are more compositionally uniform than the uncoding regions has been reported by several authors, who used different segmentation procedures and different biological material. It was reported in [24] for compilations of coding and non-coding sequences from different organisms were studied. The HMM results published in [11] also allows one to make such conclusion. The attempt to solve an inverse problem, that is to find coding regions by the segmentation procedure was published in [25] for *Rickettsia prowazekii*. The authors of [25] used an entropic segmentation, which is similar to our approach in the case of long segments (see [21] for comparison), however, they encoded DNA sequence in 12 letter alphabet associated with the codon triplet preferences. Thus their algorithm is more related to the statistical method of gene finding and in a sense is related to the HMM method, published in [11].

Special methods of statistical gene finding may have a better performance as compared to our general segmentation method. Here we want to note, that in *Plasmodium* *falciparum* genome, the majority of long exons can be extracted *only* by compositional segmentation, which we believe implies that in reality there are not so many examples of homogeneous regions in genomes, besides coding regions and repeats. The preliminary data that we obtained for *Leishmania major* confirm this conclusion.

6 Discussion

When the initial segmentation with the non-informative prior obtained, it becomes interesting to cluster the composition of the resulting segments. Firstly, our preliminary calculations on *Plasmodium falciparum* and *Leishmania major* genomes demonstrated that there are limited number of classes of long, statistically homogeneous regions in genomes of lower eukaryotes. These are long exons, divergent repeats, and other low-complexity regions, such as AT-rich strands. Short exons and intergenic regions are segmented into many short segments.

Table 1. Comparison of long GC-rich segments with coding regions in initial 200,000bp segment of *P.falciparum*

SegBeg	SegEnd	SegLen	Score	GC	ExBeg	ExEnd	ExLen	
0	1152	1152	-0.236931	0.4635	telomeric repeat			
1152	23151	21999	1392.596229	0.3237	long sub-telomeric repeat			
23281	24034	753	63.886702	0.2922				n
25106	27469	2363	139.753512	0.3275	25232	29035	3803	y
27861	29136	1275	70.190261	0.3318	27864	29183	1319	long ORF
29658	31160	1502	140.942212	0.2843	29837	31168	1331	y
32952	33956	1004	41.621573	0.3516	33030	33965	935	y
35868	37186	1318	120.276477	0.2868	35927	37249	1322	y
38300	39105	805	39.909483	0.3379	38287	39132	845	y
41439	42558	1119	46.975472	0.3512	41515	42573	1058	y
					45286	46344	1058	n
47186	49879	2693	499.468944	0.2042	48923	49861	938	y
51774	53139	1365	132.519112	0.2806	52456	53202	746	y
54351	55083	732	28.625235	0.3538	54418	54936	518	y
57445	58207	762	21.842848	0.3727	57344	58228	884	y
63358	64231	873	41.500332	0.3414	63360	64376	1016	y
66723	67531	808	41.432232	0.3354	66729	67545	816	y
					69370	69771	401	n
73492	74070	578	58.532224	0.2734	73441	74094	653	y
77158	78522	1364	160.958018	0.2595	77251	78360	1109	y
81349	83887	2538	317.825703	0.2537	81291	83900	2609	y
86838	87409	571	45.914259	0.296	86832	87400	568	y
					87675	88110	435	n

91330	98457	7127	568.747181	0.3022	91318	98532	7214	y
103364	105245	1881	57.568751	0.3732	103385	105238	1853	y
109570	110596	1026	141.488399	0.2407	109564	110202	638	y
112528	113189	661	76.080559	0.2602	112551	113167	616	long ORF
117633	118436	803	129.753598	0.2204	117558	118167	609	y
120595	124170	3575	401.15354	0.2666	120524	124102	3578	y
					127994	128314	320	n
129441	133827	4386	696.118369	0.2253	129688	133570	3882	y
					135523	137139	1616	n
					139955	140191	236	n
141536	147622	6086	1096.926023	0.2087	141625	147564	5939	y
149524	151241	1717	207.214863	0.2574	149524	156981	7457	y
151257	153666	2409	425.008465	0.2109	cont			y
153781	157075	3294	595.192248	0.208	cont			y
158090	159707	1617	277.641574	0.214	158137	159660	1523	y
160425	161310	885	138.391867	0.2249	160514	161242	728	y
					166144	168051	1907	n
168815	170150	1335	105.964553	0.3004	168838	170136	1298	y
171499	172084	585	82.778237	0.2359				n
177123	178030	907	86.610189	0.2811	176628	178300	1672	y
178844	181560	2716	470.085051	0.2135	178955	180924	1969	y
cont					181103	181526	423	
183428	185860	2432	228.672468	0.285	182228	189115	6887	y
192622	194059	1437	271.130882	0.2011	192716	193324	608	long ORF
194795	198041	3246	597.246378	0.2055	194826	196916	2090	y
198302	199601	1299	162.87607	0.2525	198353	199570	1217	y

Although these results are preliminary and should be tested on a greater number of genomes, it is very likely that segment compositions in native sequences belong to the limited number of classes. In this case the advanced segmentation algorithm, for instance Hidden Markov Models, which uses the number of compositional states as the parameter becomes entirely relevant [18]. Knowing natural compositional classes will facilitate constructing a good informative prior, which allows one to reliably annotate genomic sequences with fast segmentation method.

Moreover, the comparative positioning of the regions with distinct composition can be an interesting for assessing evolution, one of whose basic processes is the relocation of parts of genomes between chromosomes or within the same chromosome [26].

Thus, the segmentation can yield a lot of valuable biological information. We believe that our method suits best for the initial segmentation of newly sequenced genomes, with no *a priori* information on the composition. Another possible application is initial segmentation into long regions (with large B) as a preprocessing before pattern search procedure, which often uses general composition as a statistical reference. The power of such procedure can be increased by referring the algorithms not to the general composition of the sequence but to the composition of the region.

Compositional dependences can be helpful in searching for many functionally important patterns. However, in this case a more specialized algorithms appear to be more powerful. We believe that HMM is the best for such the case. Our approach is faster than HMM, since we use no sampling procedure, which requires thoughtful convergence control [18]. Thus, our algorithm can be helpful in studies of long complete genomes. In this case the application of informative prior constructed with the reference to the specific region can improve the results.

Acknowledgements. The authors wish to thank D. Haussler for stimulating discussion and for the reference to the work of Liu and Lawrence (1998), R. Guigo for discussion and for the copy of the proceedings of the Halkidiki (1997) workshop (Lawrence, 1997), D. Cellier for the inspiriting discussion and G. Shaeffer and M. Regnier for many valuable comments. We also wish to thank the anonymous referee for many valuable remarks. This study has been partially supported by Russian Human Genome Program 18/48 (V.E.R., V.Ju.M., V.G.T), and 99-0153-F(018) (M.A.R.), INTAS grant 99-1476, Lyapunov Institute Project, and MGRI project 244.

References

1. Karlin, S., Brendel, V.: Patchiness and correlation in DNA sequences. Science **259** (1993) 677–680.
2. Li, W.: The study of correlation structure of DNA sequences: a critical review. Computer & Chemistry **21(4)** (1997) 257–278.
3. Bernardi, G.: The isochore organization of the human genome. Annual Review of Genetics **23** (1989) 637–661.
4. D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C., Bernardi, G.: Correlation between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. **32** (1991) 504–510.
5. Guigo, R. Fickett, J. W.: Distinctive sequence features in protein coding, genic noncoding and intergenic human DNA. J. Mol. Biol. **253** (1995) 51–60.
6. Herzel, H., Grosse, I.: Correlation in DNA sequences: The role of protein coding segments. Phys. Rev. E. **55** (1997) 800–810.
7. Li, W., Kaneko, V.: DNA Correlations. Nature **360** (1992) 635–636.
8. Gelfand, M. S.: Prediction of function in DNA sequence analysis. Journal of Computational Biology **2** (1995) 87–117.
9. Gelfand, M. S., Koonin, E. V.: Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucl. Acid. Res **27** (1995) 2430–2439.
10. Pedersen, A. G., Baldi, P., Chauvin, Y. Brunak, S.: The biology of eukaryotic promoter prediction. Computer & Chemistry **23** (1999) 191–207.
11. Krogh, A., Mian, I. S. Haussler, D.: A hidden Markov model that finds genes in E.coli DNA. Nucl. Acid. Res **22** (1994) 4768–4778.
12. Liu, S. L., Lawrence, C. E.: Bayesian Inference of Biopolymer Models. Bioinformatics **15** (1999) 38–52.

13. Lawrence, C. E.: Bayesian Bioinformatics. 5th international conference on intelligent systems for molecular biology, Halkidiki, Greece (1997).
14. Liu, S. L., Lawrence, C. E.: Bayesian inference of biopolymer models, Stanford Statistical Department Technical Report (1998).
15. Roman-Roldan, R., Bernaola-Galvan, P. and Oliver, J. L.: Sequence compositional complexity of DNA through an entropic segmentation method. Phys. Rev. Lett. **80** (1998) 1344.
16. Churchill, G. A.: Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. **51** (1989) 79–94.
17. Durbin, R., Eddy, Y. S., Krogh, A. Mitchison, G.: Biological Sequence Analysis. Cambridge, Cambirdge University Press (1998).
18. Muri, F., Chauveau, D., Cellier, D.: Convergence assessment in latent variable models: DNA applications. In C. P. Robert (ed.) Lectural Notes in Statistics, Vol. 135, Discretization and MCMC convergence assessment., Springer. (1998) 127–146.
19. Wolpert, D. H., Wolf, D. R.: Estimating functions of probability distributions from a finite set of samples. Phys. Rev. E. **52** (1995) 6841–6854.
20. Rozanov, Y. M.: Teoriya veroyatnosti, sluchainye processy i matematicheskaya statistika (russ: Probability Theory, Stochastic Processes and Mathematical Statisitics). Moscow, Nauka (1985).
21. Ramensky, V.E., Makeev, V.Ju., Roytberg, M.A., Tumanyan, V.G.: DNA segmentation through the bayesian approach. Journal of Computational Biology., **7** (2000), 215–231.
22. Shaeffer, G. (1999) Personal communication.
23. Finkelstein, A. V., Roytberg, M. A.: Computation of biopolymers: A general approach to different problems. BioSystems **30** (1993) 1–19.
24. Ossadnik, S.M., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Peng, C.-K., Simons, M., Stanley, H.E.: Correlation approach to identify coding regions in DNA sequences. Biophysical Journal **67** (1994) 64–70.
25. Bernaola-Galván, P., Grosse, I., Carpena, P., Oliver, J., Román-Roldán, R., Stanley, H.: Finding borders between coding and noncoding DNA regions by an entropic segmentation method. Phys. Rev. Let., **85,** (2000) 1342–1345.
26. Ono, S.: Evolution by gene duplication. Springer. (1970)