Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 471–478
DOI: 10.15439/2014F88
ACSIS, Vol. 2

Change-Point Detection in Binary Markov DNA Sequences by the Cross-Entropy Method

Tatiana Polushina
Department of Clinical Science,
Faculty of Medicine and Dentistry,
University of Bergen, 7804,
NO-5020 Bergen, Norway
Email: t.polushina@gmail.com

Georgy Sofronov
Department of Statistics
Faculty of Science
Macquarie University
Sydney NSW 2109 Australia
Email: georgy.sofronov@mq.edu.au

***Abstract*—A deoxyribonucleic acid (DNA) sequence can be represented as a sequence with 4 characters. If a particular property of the DNA is studied, for example, GC content, then it is possible to consider a binary sequence. In many cases, if the probabilistic properties of a segment differ from the neighbouring ones, this means that the segment can play a structural role. Therefore, DNA segmentation is given a special attention, and it is one of the most significant applications of change-point detection. Problems of this type also arise in a wide variety of areas, for example, seismology, industry (e.g., fault detection), biomedical signal processing, financial mathematics, speech and image processing. In this study, we have developed a Cross-Entropy algorithm for identifying change-points in binary sequences with first-order Markov dependence. We propose a statistical model for this problem and show effectiveness of our algorithm for synthetic and real datasets.**

I. Introduction

THE eukaryotic genomes are packaged into nucleosomes, composed of approximately 147 base pairs. There are 4 different bases: adenine (A), cytosine (C), guanine(G) and thymine (T). We can consider different approaches to base partition that depend on chemical and physical structure. One type of separation is pyrimidines (T and C) and purine (A and G). The second type of separation is keto (T and G) and amino (A and C) groups. In this paper, we consider groups of complemental bases: GC and AT pairs.

In this study, we are interested in finding regions that differ from neighbouring ones in GC level. It is well-known that genomic sequences are nonhomogeneous with respect to GC level, differences in GC proportion may be over scale of 100 kb to megabases. These long segments are called GC-content domains or isochores [1], [2]. Many studies propose that the differences of GC proportion appear as an outcome from a selection process [3]. It is well-known that an average GC proportion in chromatin organization and, hence, gene regulation is significant [4]. So GC proportion has been revealed to correlate with genomic properties such as DNA bendability and regulated replication.

In the last years, this topic has been investigated by many researchers [5], [6], [7]. This stimulates the elaboration of computational techniques that are applied to large-scale biological experimental data. Positive relationships have been discovered between GC level and recombination in humans, birds, and plants [8], [9], [10], [11]. Spencer *et al.* [5] have discovered that recombination proportions are too fast-evolving to have permanent meanings on base composition.

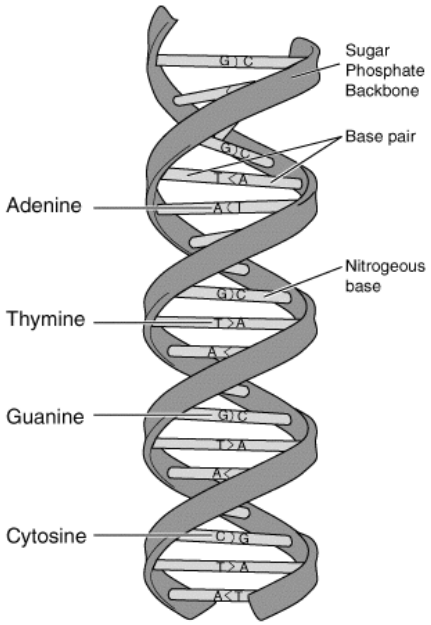

Fig. 1. The DNA structure. http://www.niherst.gov.tt/scipop/sci-bits/genetics-and-epigentics.htm

Positions in a DNA sequence at which nucleotides C or G are situated can be represented by a 1, and locations with T or A are situated can be represented by a 0. More formally, a sequence $a = \{a_1, \ldots, a_L\}$ of length L is given, where $a_m \in \{A, C, G, T\}$. The sequence may be transformed to a binary sequence $b = \{b_1, \ldots, b_L\}$ in which

$$b_m = \begin{cases} 1 & \text{if } a_m \in \{C, G\}, \\ 0 & \text{if } a_m \in \{A, T\}. \end{cases}$$

This work was carried out when the first author was at the Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, NO 7491, Trondheim, Norway.

978-83-60810-58-3/$25.00

From mathematical point of view we can designate a bound of segments with different GC ratio as a break-point or a change-point. Biological applications of the change-point problem, in particular, to DNA sequences, have been extensively considered in literature (see, for example, [12], [13], [14], [15], [16]). Note that the multiple change-point problem is a flexible model, which can be applied in many areas such as economics, finance, environmental control [17], [18], [19], signal detection, quality control [20], health and surveillance [21], [22]. Various techniques to the change-point problem with independent observations have been developed [16], [23], [24], [25], including stochastic optimization methods [13], [26], [27], [28], [29], [30], [31] and Markov chain Monte Carlo (MCMC) algorithms [14], [32], [33], [34], [35]. The Cross-Entropy (CE) method for independent case was developed in [13].

We can formulate a more general change-point problem for a sequence of dependent observations. The case of the Markov dependence in biological sequences was investigated in different articles. Polansky [36] considered cases with known and unknown number of change-points. The author applied the likelihood ratio, the bootstrap for estimation p-values for these cases, the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) with unknown number of change-points. Zhang and Siegmund [37] proposed a new penalty component in the modified BIC. Avery and Herderson [12] investigated a problem of prediction of the occurrence of the definite sequence in DNA. For this purpose they considered the first-order, the second-order and the higher-order Markov chain models. Then the authors [38] developed a nonparametric method based on the approach of Pettitt [39]. Krauth [15], [40] used the exact Fisher test and the finite conditional tests for the multiple change-point problem in binary first-order Markov sequences. In this paper, we develop the CE method for identifying change-points in the first-order Markov dependence in binary sequences for artificial and real data.

We use the genome of the Bacteriophage *lambda*, a virus of the intestinal bacterium *Escherichia coli*, and a part of the Human Major Histocompatibility Region. Consideration of individual chromosomes is one of the most common approaches in the literature [41], [42]. Particularly it is very important for the analysis of the cancer genome [43].

The paper is structured as follows. Section 2 provides a statement of the multiple change-point problem in mathematical terms. In Section, 3 we describe a general framework of Cross-Entropy method. Section 4 contains developing the Cross-Entropy algorithm for the multiple change-point problem in dependent case. In Section 5, we discuss the results of numerical experiments.

II. The Multiple Change-Point Problem in Binary Markov Sequences

In mathematical terms we can describe the general multiple change-point problem as follows. A binary sequence $b = (b_1, \dots, b_L)$ of length L is given. A segmentation of the sequence is specified by giving the positions of the change-points $c = (c_1, \dots, c_N)$ and the number of change-points N, where $1 = c_0 < c_1 < \cdots < c_N < c_{N+1} = L$. This means that a change-point is a boundary between two neighbouring regions, and the value c_n is the sequence position of the rightmost character of the segment to the left of the n-th change-point.

In this model we assume that characters within each region are generated by Bernoulli trials with first-order Markov dependence. The probability distribution, which depends on the segment, can be represented by a transition matrix

$$\begin{pmatrix} \theta_0 & 1-\theta_0 \\ \theta_1 & 1-\theta_1 \end{pmatrix},$$

where $\theta_0 = P(X_{m+1} = 0 \mid X_m = 0)$, $1-\theta_0 = P(X_{m+1} = 1 \mid X_m = 0)$, $\theta_1 = P(X_{m+1} = 0 \mid X_m = 1)$, $1-\theta_1 = P(X_{m+1} = 1 \mid X_m = 1)$.

Thus, the likelihood function of N, $c = (c_1, \dots, c_N)$, and

$$\theta = (\theta_{00}, \theta_{10}, \dots, \theta_{0n}, \theta_{1n}, \dots, \theta_{0N}, \theta_{1N}),$$

is given by

$$\begin{aligned} f(N, c, \theta) &= P(X_1 = b_1) \\ &\times \prod_{n=0}^{N} \theta_{0n}^{\mathbf{I}_{00}(c_n, c_{n+1})} (1-\theta_{0n})^{\mathbf{I}_{01}(c_n, c_{n+1})} \\ &\times \theta_{1n}^{\mathbf{I}_{10}(c_n, c_{n+1})} (1-\theta_{1n})^{\mathbf{I}_{11}(c_n, c_{n+1})}, \end{aligned}$$

where $\mathbf{I}_{ij}(c_n, c_{n+1})$ is the number of times i ($i = 0, 1$), is followed by j ($j = 0, 1$) in the segment bounded by sequence positions $c_n + 1$ and c_{n+1}.

In order to simplify optimization, we consider the log-likelihood function at point $x = (N, c, \theta)$, having observed $b_1, \dots, b_L$,

$$\begin{aligned} \pi(x) &= \ln P(X_1 = b_1) \\ &+ \sum_{n=0}^{N} \Big(\mathbf{I}_{00}(c_n, c_{n+1}) \ln \theta_{0n} \\ &+ \mathbf{I}_{01}(c_n, c_{n+1}) \ln(1-\theta_{0n}) \\ &+ \mathbf{I}_{10}(c_n, c_{n+1}) \ln \theta_{1n} + \mathbf{I}_{11}(c_n, c_{n+1}) \ln(1-\theta_{1n}) \Big). \end{aligned} \tag{1}$$

III. The Cross-Entropy Method

From mathematical point of view the multiple change-point detection problem can be interpreted as a maximization problem of the log-likelihood function defined in (1).

Let F be a real valued performance function on $\mathcal{X}$, where $\mathcal{X}$ is a finite set of states. We want to find the optimum of F over $\mathcal{X}$, and the state corresponding to this value (which is a vector of positions of change-points). We can apply stochastic optimization methods for this optimization problem, in particular, the CE method.

The CE method is a technique for the estimation of rare event probabilities [44], [45], [46]. This estimation problem can be reformulated as an optimization problem. Thus we

define a set of indicator functions $\{I_{\{S(x)\geq\gamma\}}\}$ on $\mathcal{X}$ for different levels $\gamma \in R$. Let $\{f(\cdot,u)\}$ be a family of probability density functions (pdfs) on $\mathcal{X}$ with a real-valued parameter u. Following [45], we associate the optimization problem with the problem of estimating

$$l(\gamma) = \mathbf{P}_u(S(X) \geq \gamma) = \sum_x I_{\{S(x)\geq\gamma\}} f(x,u) = \mathbf{E}_u I_{\{S(X)\geq\gamma\}},$$

where γ is a known or unknown parameter and $\mathbf{P}_u$ is the probability measure under which the random state X has the pdf $f(\cdot,u)$.

The problem of estimating l is not trivial. Adaptive changes to the pdf are based on the Kullback-Leibler (or the CE) distance. Thus it allows to create a sequence $f(\cdot,u_0), f(\cdot,u_1), \ldots, f(\cdot,u^*)$. The final pdf $f(\cdot,u^*)$ corresponds to the density at an optimal point. This means that the CE method creates a sequence of pairs $\{(\gamma_t,u_t)\}$, which converges quickly to a close neighbourhood of the optimal tuple (γ^*,u^*). More specifically, we should set up u_0 and simulation parameters, and then we carry out the following procedure [45]:

1) **Adaptive updating of** γ_t. For a fixed u_{t-1}, let γ_t be a $(1-\rho)$-quantile of $\widehat{S}(X)$ under u_{t-1}. A simple estimator $\widehat{\gamma}_t$ of γ_t is
$$\widehat{\gamma}_t = \widehat{S}_{(\lceil(1-\rho)N_2\rceil)},$$
where, for a random sample $X_1,\ldots,X_{N_2}$ from $f(\cdot,u_{t-1})$, $\widehat{S}_{(i)}$ is the i-th order statistic of the performances $\widehat{S}(X_1),\ldots,\widehat{S}(X_{N_2})$.
2) **Adaptive updating of** u_t. For fixed γ_t and u_{t-1}, derive u_t from the solution of the CE program
$$\max_u D(u) = \max_u \mathbf{E}_{u_{t-1}} I_{\{\widehat{S}(X)\geq\gamma_t\}} \ln f(X,u).$$

IV. The Cross-Entropy Method for the Multiple Change-Point Problem

Let N be the number of change-points and c be a set of the change-points, which is a nondecreasing N-dimensional vector.

We apply the CE algorithm that uses normal distributions to simulate the change-point positions. The CE method updates the parameters in each step and updating is continued until a convergence state is achieved. A variance-based stopping criterion is used to estimate the fit of the combinations of change-points in each step.

Our study differs from previous [13] in the following aspects. Firstly, we consider a change-point problem for a sequence of dependent observations. Secondly, we apply the BIC (Bayesian information criterion) [47], [48] in order to estimate the number of change-points, which is usually unknown. The combination that minimizes F (our performance function) under the corresponding N is considered as the optimal solution. Therefore, we replace the problem of maximization of log-likelihood function with minimization problem of the BIC.

TABLE I
Parameters θ in Example 1

positions	θ_1	θ_2
1–2000	0.9	0.5
2001–4000	0.4	0.15
4001–6000	0.1	0.6
6001–8000	0.6	0.9
8001–10000	0.2	0.4
10001–12000	0.4	0.2
12001–14000	0.2	0.7
14001–16000	0.6	0.5
16001–18000	0.4	0.9
18001–20000	0.2	0.2
20001–22000	0.7	0.5

For each change-point vector c in the sample, we obtain the maximum likelihood estimate of parameters with respect to the each of the segments and evaluate the performance function F. The performance function, the BIC, which we minimize is

$$F = -2\pi(x) + k\ln(L), \tag{2}$$

where $\pi(x)$ is the log-likelihood as in (1) of the sequence. We use the standard penalty

$$k\ln(L) = (3N+2)\ln(L).$$

In each iteration an *elite* sample is defined as the best performing combinations of change-points with respect to the performance function score. The process is carried out until a specific stopping criterion is achieved.

In each step, the simulation parameters are updated accordingly. The main steps of our algorithm are described in Algorithm 1.

We should specify N_1, ρ, ε , the parameters of the algorithm as well as the initial values for the simulation parameters μ and σ^2. Note that we choose the parameters under the conditions which guarantee convergence of the algorithm [49].

V. Numerical Results

In this section, we include results of numerical experiments that illustrate the performance of the CE method. In the first example, we consider a synthetic sequence with a known distribution, which allows us to provide direct comparison of estimated and true profiles in terms of the Root Mean Squared Error (RMSE). The second and the third examples use real DNA sequences and we do not have any information about the structure of dependence. We apply a test of independence for these examples.

A. Example 1: Artificial data

Let $(b_1, b_2, ..., b_{22000})$ be a sequence of random variables generated with the parameters from Table I.

At first, we assume that we do not know the number of change-point and apply our algorithm for different N. We run our algorithm with the following simulation parameters: the elite proportion value $\rho = 0.1$ and the sample size $N_1 = 1500$.

Algorithm 1 Algorithm for change-point detection

1: Choose initial sets for

$$\mu^{(0)} = \left(\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_N^{(0)}\right)$$

and

$$(\sigma^2)^{(0)} = \left((\sigma_1^2)^{(0)}, (\sigma_2^2)^{(0)}, \ldots, (\sigma_N^2)^{(0)}\right).$$

The length of both vectors is N. Set $t = 1$.

2: Generate a random sample $c^{(1)}, c^{(2)}, \ldots, c^{(N_1)}$ from the normal distributions with parameters $\left(\mu^{(t-1)}, (\sigma^2)^{(t-1)}\right)$, where $c^{(i)} = \left(c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)}\right)$, $i = 1, 2, \ldots, N_1$, is a change-point vector.

3: For $i = 1, 2, \ldots, N_1$ order $\left(c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)}\right)$ from smallest to biggest.

4: Evaluate the performance of each $c^{(1)}, c^{(2)}, \ldots, c^{(N_1)}$ based on (2).

5: Define the elite sample, which is the best performing combinations of the change-points.

6: Let $N_{elite} = \rho N_1$ be the size of the elite sample.

7: For all $j = 1, 2, \ldots, N$, estimate the parameters $\mu_j^{(t)}$ and $(\sigma_j^2)^{(t)}$ using the elite sample and update the current parameter sets as follows:

$$\mu_j^{(t)} = \frac{\sum\limits_{i \in I} c_j^{(i)}}{N_{elite}}, \qquad \left(\sigma_j^2\right)^{(t)} = \frac{\sum\limits_{i \in I} \left(c_j^{(i)} - \mu_j^{(t)}\right)^2}{N_{elite}},$$

where I is the set of indices of the best performing samples.

8: Stopping criterion is $\max_j (\sigma_j^2)^{(t)} < \varepsilon$.

9: **if** Stopping criterion is met **then**

10: stop the process and identify the combination of the positions of change points $c^{(i)}$ that minimizes the BIC

11: **else**

12: $t = t + 1$;

13: and iterate from step 2.

14: **end if**

Then we obtain the best solution for different models in each of the N situations which minimize the BIC. We can see from Figure 2 that the minimum value of the BIC at $N = 10$, which corresponds to the number of change-points in Table I.

The true profiles of this sequence as well as the estimated profile can be seen in Figures 3, 4. We can see that the estimated and the true plots are very similar to each other. This indicates that the CE method works very well and it properly captures the segments in the binary sequence.

To test the efficiency of the CE method, we have applied this algorithm with various values for the parameter ρ, which is used to obtain the elite sample. We calculate the RMSE for the different algorithms when applied to the synthetic sequence

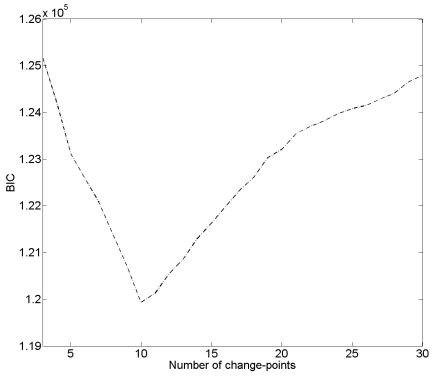

Fig. 2. The scores of the BIC for different N

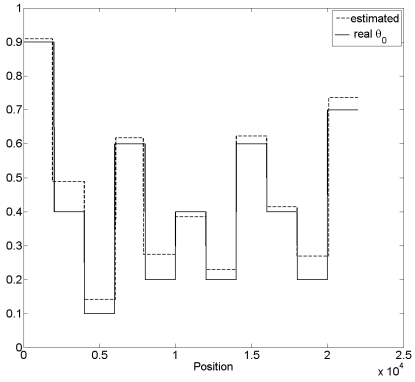

Fig. 3. The profile of θ_0 obtained by the CE algorithm

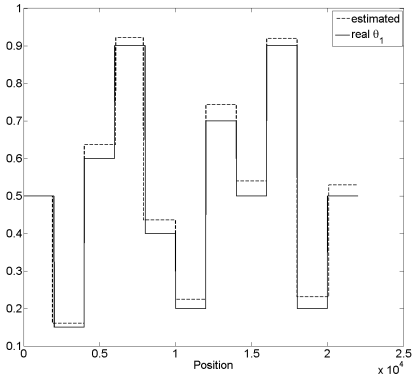

Fig. 4. The profile of θ_1 obtained by the CE algorithm

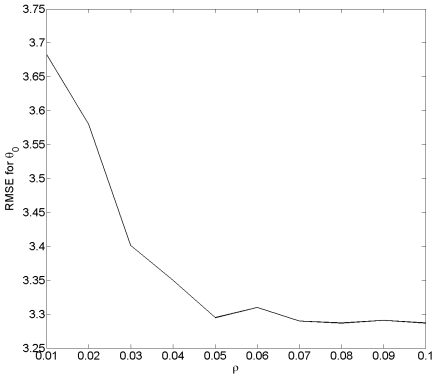

Fig. 5. The values of the RMSE for θ_0 depending on ρ

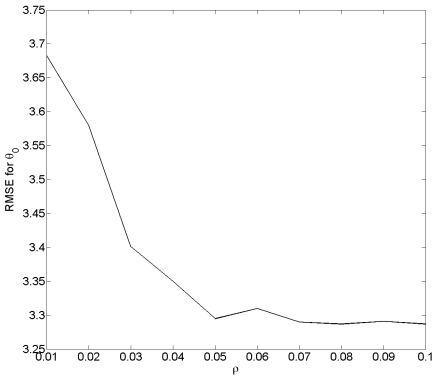

Fig. 6. The values of the RMSE for θ_1 depending on ρ

of 22000 characters

$$\text{RMSE} = \sqrt{\sum_{i=1}^{22000} (t(i) - e(i))^2},$$

where $e(i)$ is estimated value at position i and $t(i)$ is the true parameter value.

The RMSE and CPU time are obtained for ρ values from 0.01 to 0.1 with step of 0.01 for the model when number of change-points is 10. We have obtained the average results based on 50 simulations under each of the ρ values. We can see from Figures 5, 6 that the plots are slowly decreasing, at the same time the plot on Figure 7 is increasing. In this study, we focus on the RMSE, though it would be possible to choose ρ in such a way that will balance the trade-off between the RMSE and the CPU time.

B. Example 2: Real data (Bacteriophage lambda)

We apply the CE with the same parameter specification as above to the genome of the Bacteriophage *lambda*, a virus of the intestinal bacterium *Escherichia coli*. The length of the sequence is 48,502 bases. Boys and Henderson [50] studied this sequence with 4 multinomial outcomes (each base is one of either A, C, G, T) for the comparison of different algorithms

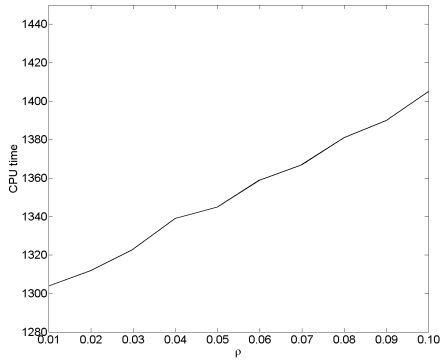

Fig. 7. CPU time for different ρ

TABLE III
OBSERVED FREQUENCIES OF THE 4 POSSIBLE PAIRS OF BASES FOR EXAMPLE 2. EXPECTED FREQUENCIES ASSUMING INDEPENDENCE OF SUCCESSIVE BASES ARE GIVEN IN PARENTHESES

First base	**Second base**	**Second base**	**Total**
	0	1	
0	12544	11776	24320
	(12194.85)	(12125.15)	
1	11776	12405	24181
	(12125.15)	(12055.85)	
Total	24320	24181	48501

under the independence assumption. Table II presents a brief summary of the results obtained in [51].

Table III shows the observed frequencies and the expected frequencies for the Pearson χ^2-test of independence. It can be calculated from the table that the value of the test statistic is 40.22. On comparing with a χ^2-distribution with 1 degree of freedom, we conclude that the hypothesis about independence should be rejected ($p < 10^{-6}$). Therefore, we consider a case with the first-order Markov dependence.

We can calculate the BIC for different number of change-points. From Table II we can see that the authors found 8 change-points based on the use of 4-symbol alphabet. According to our approach we found that 6236 was the minimum value of the BIC at $N = 9$. Next, we check each change-point using the Fisher exact test. We calculate p-values and conclude that there are evidences for change-points at 5806 ($p_1 = 7.82 \cdot 10^{-4}$), 19503 ($p_4 = 0.018$), 22109 ($p_5 = 1.25 \cdot 10^{-11}$), 27660 ($p_6 = 6.18 \cdot 10^{-6}$), 38018 ($p_8 = 0.0045$), and 46259 ($p_9 = 5.19 \cdot 10^{-4}$).

Note that our main objective is to identify change-points in GC ratio, not in the model parameters θ_0, θ_1. Therefore, we present our conclusions without profiles of θ_0 and θ_1 and locations of false change-points. The GC profile can be seen on Figure 8. The discordance can be explained by the fact that the results in Table II were obtained using a different model with 4-character alphabet, whereas we used a binary representation. From this comparison we can see that both methods identify

TABLE II
ESTIMATED SEGMENTS AND ESTIMATED PROPORTIONS OF A, C, G, AND T FOR EACH SEGMENT

	A	C	G	T	G+C	A+T
0 – 20091	0.23	0.25	0.32	0.20	0.57	0.43
20092 – 20919	0.29	0.29	0.30	0.11	0.59	0.41
20902 – 22544	0.26	0.24	0.27	0.23	0.51	0.49
22545 – 24117	0.29	0.14	0.16	0.40	0.30	0.70
24118 – 27829	0.29	0.20	0.18	0.33	0.38	0.62
27830 – 33082	0.23	0.26	0.22	0.29	0.48	0.52
33083 – 38029	0.27	0.22	0.21	0.31	0.43	0.57
38030 – 46528	0.30	0.23	0.26	0.22	0.49	0.51
46529 – 48502	0.27	0.18	0.22	0.33	0.40	0.60

TABLE IV
OBSERVED FREQUENCIES OF THE 4 POSSIBLE PAIRS OF BASES FOR EXAMPLE 3

First base	Second base	Second base	Total
	0	1	
0	5344	5345	10689
	(5713.56)	(4975.49)	
1	5346	3964	9310
	(4976.44)	(4333.56)	
Total	10690	9309	19999

the most significant change-points and the proposed method provides a smoother profile of GC ratio.

C. Example 3: Real data (MHC Region)

This example uses a part of the Human Major Histocompatibility Region (MHC) (for further detail, see [52]). Due to this being real DNA, we do not know the true profile (as well as in Example 2). Instead we look for agreement between the CE and two well-known methods: IsoFinder [16], [23], [24] and the BAIS [34], [35]. At first, we repeat the Pearson test of independence. The value of the test statistic from Table IV is 51.35. This means that the hypothesis about independence should also be rejected ($p < 10^{-6}$).

We use the same algorithm parameters as before. We found a change-point vector and checked each position using the exact Fisher test. There are 6 significant change-points in this part of MHC sequence: 953 ($p_1 = 3.67 \cdot 10^{-4}$), 7257 ($p_4 = 0.0078$), 9132 ($p_5 = 7.80 \cdot 10^{-6}$), 13041 ($p_6 = 6.28 \cdot 10^{-12}$), 16114 ($p_7 = 3.19 \cdot 10^{-11}$), and 18954 ($p_8 = 1.05 \cdot 10^{-30}$).

Figure 9 shows the GC profiles for the CE algorithm, the BAIS and the IsoFinder. We use the following simulation parameters: the BAIS algorithm for 1000 iterations and $K = 50$ parallel chains, and IsoFinder with a 0.95 significance level and tract size of 1,000. It is clear that all algorithms can detect the major regions within the MHC sequence. IsoFinder identifies seven major regions while the other methods all identify several smaller regions within these major regions. The agreement between these methods allows for a great deal of confidence in the exactness of the CE method as both the BAIS method and IsoFinder are well established.

VI. CONCLUSION

In this paper, we have developed the Cross-Entropy method for identifying change-points in binary Markov sequences. In order to identify the correct number of change-points we propose to use the BIC. This approach is easy to implement and can also be extended to more general multiple change-point models. We have demonstrated the effectiveness of this technique in examples using both real and synthetic sequences. The method has been shown to be highly effective on synthetic data and real DNA sequences and compete well with existing approaches.

The proposed approach gives results similar to previous outcomes but it is not sufficient for understanding of dependence mechanism in DNA sequences. Our future research will include consideration of Markov dependence of a higher order (the second or more). The proposed method can be implemented using parallel computing, which will significantly decrease the CPU time. For the independent case, this feature was realized in R-package *breakpoint* [53].

ACKNOWLEDGMENT

The first author was supported by ERCIM programme. This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme at the Norwegian University of Science and Technology, Trondheim, Norway. This programme is supported by the Marie Curie Co-funding of Regional, National and International Programmes (COFUND) of the European Commission.

REFERENCES

[1] G. Bernardi, *Structural and evolutionary genomics. Natural Selection in Genome Evolution*. Amsterdam: Elsevier, 2004.
[2] M. Costantini, O. Clay, F. Auletta, and G. Bernardi, "An isochore map of human chromosomes," *Genome Res.*, vol. 16, pp. 536–541, 2006, http://dx.doi.org/10.1101/gr.4910606
[3] L. Ren, G. Gao, D. Zhao, M. Ding, J. Luo, and H. Deng, "Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprint by models of stem cell differentiation," *Genome Biol.*, vol. 8, p. R35, 2007, http://dx.doi.org/10.1186/gb-2007-8-3-r35
[4] M. Semon, D. Mouchiroud, and L. Duret, "Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance," *Hum. Mol. Genet.*, vol. 14, pp. 421–427, 2005, http://dx.doi.org/10.1093/hmg/ddi038

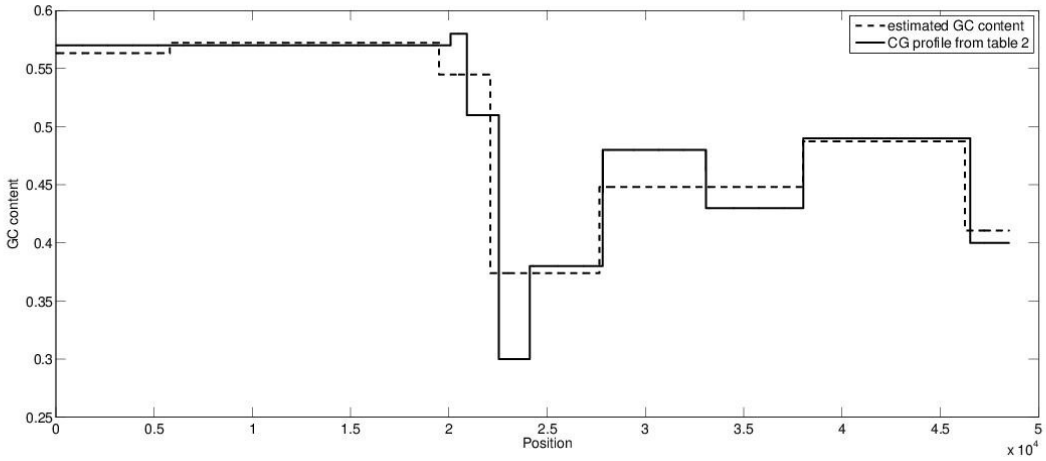

Fig. 8. Bacteriophage *lambda*.

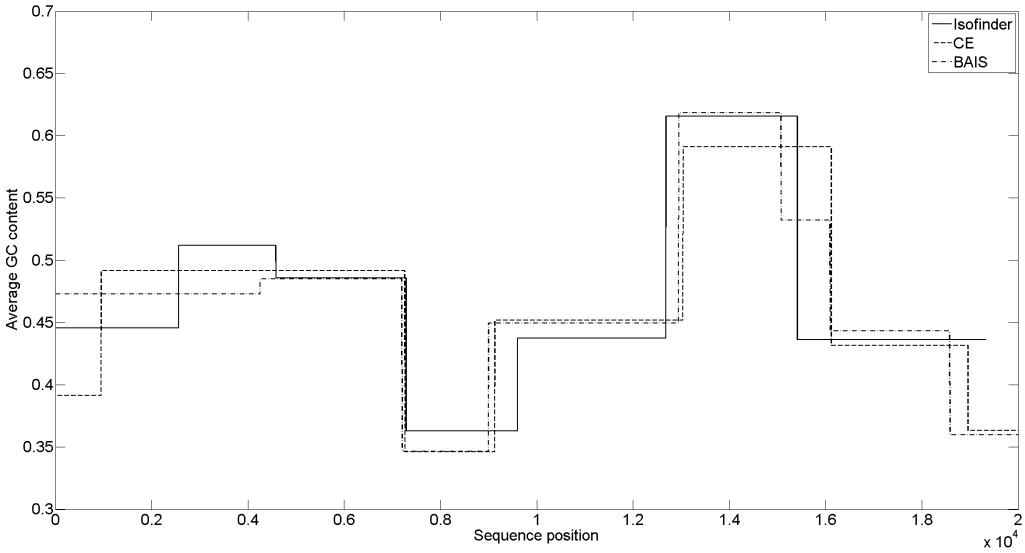

Fig. 9. GC profiles for the CE, IsoFinder and BAIS methods on the MHC sequence.

[5] C. Spencer, P. Deloukas, S. Hunt, J. Mullikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, and G. McVean, "The influence of recombination on human genetic diversity," *PLoS Genet*, vol. 2, p. e148, 2006, http://dx.doi.org/10.1371/journal.pgen.0020148

[6] A. Vinogradov, "Isochores and tissue-specificity," *Nucleic Acids Res.*, vol. 31, pp. 5212–5220, 2003, http://dx.doi.org/10.1093/nar/gkg699

[7] A. Vinogradov, "Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth," *Trends Genet*, vol. 21, pp. 639–643, 2005, http://dx.doi.org/10.1016/j.tig.2005.09.002

[8] L. Hurst, C. Brunton, and N. Smith, "Small introns tend to occur in GC-rich regions in some but not all vertebrates," *Trends Genet*, vol. 15, pp. 437–439, 1999, http://dx.doi.org/10.1016/S0168-9525(99)01832-6

[9] T. Ikemura and K. Wada, "Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data," *Nucleic Acids Res.*, vol. 19, pp. 4333–4339, 1991.

[10] T. Takano-Shimizu, "Local changes in GC/AT substitution biases and in crossover frequencies on drosophila chromosomes," *Mol. Biol. Evol.*, vol. 18, pp. 606–619, 2001, http://dx.doi.org/10.1093/oxfordjournals.molbev.a003841

[11] E. Willams and L. Hurst, "The proteins of linked genes evolve at similar rates," *Nature*, vol. 407, pp. 900–903, 2000, http://dx.doi.org/10.1038/35038066

[12] P. Avery and D. Henderson, "Fitting Markov chain models to discrete state series such as DNA sequences," *Appl. Statist.*, vol. 48, no. 1, pp. 53–61, 1999, http://dx.doi.org/10.1111/1467-9876.00139

[13] G. E. Evans, G. Y. Sofronov, J. M. Keith, and D. P. Kroese, "Estimating change-points in biological sequences via the cross-entropy method," *Ann. Oper. Res.*, vol. 189, no. 1, pp. 155–165, 2011, http://dx.doi.org/10.1007/s10479-010-0687-0

[14] J. M. Keith, "Segmenting eukaryotic genomes with the generalized Gibbs sampler," *J. Comp. Biol.*, vol. 13, no. 7, pp. 1369–1383, 2006, http://dx.doi.org/10.1089/cmb.2006.13.1369

[15] J. Krauth, "Multiple change points and alternating segments in binary trials with dependence," in *Innovations in Classification, Data Science,*

and Information Systems, D. Baier and K. Wernecke, Eds. Springer, Berlin, 2004, pp. 154–164, http://dx.doi.org/10.1007/3-540-26981-9_19

[16] J. Oliver, P. Bernaola-Galvan, P. Carpena, and R. Roman-Roldan, "Isochore chromosome maps of eukaryotic genomes," *Gene*, vol. 276, pp. 47–56, 2001, http://dx.doi.org/10.1016/S0378-1119(01)00641-2

[17] I. López, M. Gámez, J. Garay, T. Standovár, and Z. Varga, "Application of change-point problem to the detection of plant patches," *Acta Biotheoretica*, vol. 58, pp. 51–63, 2010, http://dx.doi.org/10.1007/s10441-009-9093-x

[18] J. R. Thomson, W. J. Kimmerer, L. R. Brown, K. B. Newman, R. Mac Nally, W. A. Bennett, F. Feyrer, and E. Fleishman, "Bayesian change point analysis of abundance trends for pelagic fishes in the upper San Francisco estuary," *Ecological Applications*, vol. 20, no. 5, pp. 1431–1448, 2010, http://dx.doi.org/10.1890/09-0998.1

[19] M. Priyadarshana, G. Sofronov, "A modified cross-entropy method for detecting change-points in the Sri-Lankan stock market," In: *The IASTED International Conference on Engineering and Applied Science (EAS2012)*, Chen, B. M.; Khan, M. T. and Tan, K-K. (Eds.), 2012, pp. 321-326, http://dx.doi.org/10.2316/P.2012.785-041

[20] G. Sofronov, T. Polushina, and M. Priyadarshana, "Sequential change-point detection via the cross-entropy method," in *The 11th Symposium on Neural Network Applications in Electrical Engineering (NEUREL2012)*, B. Reljin and S. Stankovic, (Eds.), 2012, pp. 185–188, http://dx.doi.org/10.1109/NEUREL.2012.6420004

[21] C. Sonesson and D. Bock, "A review and discussion of prospective statistical surveillance in public health," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 166, no. 1, pp. 5–21, 2003, http://dx.doi.org/10.1111/1467-985X.00256

[22] J. Whittaker and S. Frühwirth-Schnatter, "A dynamic changepoint model for detecting the onset of growth in bacteriological infections," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 43, no. 4, pp. 625–640, 1994, http://dx.doi.org/10.2307/2986261

[23] J. Oliver, P. Carpena, R. Roman-Roldan, T. Mata-Balaguer, A. Mejias-Romero, M. Hackenberg, and P. Bernaola-Galvan, "Isochore chromosome maps of the human genome," *Gene*, vol. 300, pp. 117–127, 2002, http://dx.doi.org/10.1016/S0378-1119(02)01034-X

[24] J. Oliver, R. Roman-Roldan, J. Perez, and P. Bernaola-Galvan, "Segment: identifying compositional domains in DNA sequences," *Bioinformatics*, vol. 15, pp. 974–979, 1999, http://dx.doi.org/10.1093/bioinformatics/15.12.974

[25] G. Y. Sofronov, G. E.Evans, J. M. Keith, and D. P. Kroese, "Identifying change-points in biological sequences via sequential importance sampling," *Environmental Modeling and Assessment*, vol. 14, no. 5, pp. 577–584, 2009, http://dx.doi.org/10.1007/s10666-008-9160-8

[26] T. Polushina and G. Sofronov, "Change-point detection in biological sequences via genetic algorithm," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC'2011)*, 2011, pp. 1966–1971, http://dx.doi.org/10.1109/CEC.2011.5949856

[27] M. Priyadarshana and G. Sofronov, "The Cross-Entropy method and multiple change-points detection in zero-inflated DNA read count data," in *The 4th International Conference on Computational Methods (ICCM2012)*, Y. T. Gu, S. C. Saha (Eds.), 2012 , pp. 1–8.

[28] M. Priyadarshana and G. Sofronov, "GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data," in *Proceedings of the 28th International Workshop on Statistical Modelling*, Muggeo VMR, Capursi V, Boscaino G, Lovison G (Eds.), 2013, vol.1, pp. 453–457.

[29] T. V. Polushina and G. Y. Sofronov, "A hybrid genetic algorithm for change-point detection in binary biomolecular sequences," in *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2013)*, 2013, pp. 1–8, http://dx.doi.org/10.2316/P.2013.793-026

[30] M. Priyadarshana, T. Polushina, and G. Sofronov, "A hybrid algorithm for multiple change-point detection in continuous measurements," in *International Symposium on Computational Models for Life Sciences, AIP Conference Proceedings*, vol. 1559, 2013, pp. 108–117, http://dx.doi.org/10.1063/1.4825002

[31] M. Priyadarshana, T. Polushina, and G. Sofronov, "Hybrid algorithms for multiple change-point detection in biological sequences," in Signal and Image Analysis for Biomedical and Life Sciences (Advances in Experimental Medicine and Biology), Sun, C., Bednarz, T., Pham, T. D., Vallotton, P., Wang, D. (Eds.), Springer, 2014, in press.

[32] J. M. Keith, P. Adams, S. Stephen, and J. S. Mattick, "Delineating slowly and rapidly evolving fractions of the drosophila genome," *J. Comp. Biol.*, vol. 15, no. 4, pp. 407–430, 2008, http://dx.doi.org/10.1089/cmb.2007.0173

[33] J. M. Keith, D. P. Kroese, and D. Bryant, "A generalized Markov sampler," *Methodology and Computing in Applied Probability*, vol. 6, no. 1, pp. 29–53, 2004, http://dx.doi.org/10.1023/B:MCAP.0000012414.14405.15

[34] J. Keith, D. Kroese, and G. Sofronov, "Adaptive independence samplers," *Statistics and Computing*, vol. 18, no. 4, pp. 409–420, 2008, http://dx.doi.org/10.1007/s11222-008-9070-2

[35] G. Sofronov, "Change-point modelling in biological sequences via the bayesian adaptive independent sampler," *International Proceedings of Computer Science and Information Technology*, vol. 5, pp. 122–126, 2011.

[36] A. Polansky, "Detecting change-points in Markov chains," *Computational Statistics and Data Analysis*, vol. 51, pp. 6013–6026, 2007, http://dx.doi.org/10.1016/j.csda.2006.11.040

[37] N. Zhang and D. Siegmund, "A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, vol. 3, pp. 22–32, 2007, http://dx.doi.org/10.1111/j.1541-0420.2006.00662.x

[38] P. Avery and D. Henderson, "Detecting a changed segment in DNA sequences," *Appl. Statist.*, vol. 48, no. 4, pp. 489–503, 1999, http://dx.doi.org/10.1111/1467-9876.00167

[39] A. Pettitt, "A non-parametric approach to the change-point problem," *Appl. Statist.*, vol. 28, pp. 126–135, 1979, http://dx.doi.org/10.2307/2346729

[40] J. Krauth, "Tests for multiple change points in binary Markov sequences," in *From Data and Information Analysis to Knowledge Engineering*. Springer, Berlin, 2006, pp. 670–677, http://dx.doi.org/10.1007/3-540-31314-1_82

[41] R. Thurman, N. Day, W. Noble, and J. Stamatoyannopoulos, "Identification of higher-order functional domains in the human ENCODE regions," *Genome Res.*, vol. 17, pp. 917–927, 2007, http://dx.doi.org/10.1101/gr.6081407

[42] H. Xu, C. Wei, F. Lin, and W. Sung, "An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data," *Bioinformatics*, vol. 24, pp. 2344–2349, 2008, http://dx.doi.org/10.1093/bioinformatics/btn402

[43] B. Zeitouni, V. Boeva, I. Janoueix-Lerosey, O. Delattre, A. Nicolas, and E. Barillot, "SVDetect - a bioinformatic tool to identify genomic structural variations from paired-end next-generation sequencing data," *Bioinformatics*, vol. 26, pp. 1895–1896, 2010, http://dx.doi.org/10.1093/bioinformatics/btq293

[44] Z. I. Botev, D. Kroese, and T. Taimre, "Generalized cross-entropy methods with applications to rare-event simulation and optimization," *Simulation*, vol. 83, no. 11, pp. 785–806, 2007, http://dx.doi.org/10.1177/0037549707087067

[45] R. Rubinstein and D. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York: Springer-Verlag, 2004.

[46] R. Rubinstein and D. Kroese, *Simulation and the Monte Carlo Method*. John Wiley & Sons, 2007.

[47] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978, http://dx.doi.org/10.1214/aos/1176344136

[48] Y. Yao, "Estimating the number of change-points via Schwarz criterion," *Statistics and Probability Letters*, vol. 6, pp. 181–189, 1988, http://dx.doi.org/10.1016/0167-7152(88)90118-6

[49] A. Costa, O. Jones, and D. Kroese, "Convergence properties of the cross- entropy method for discrete optimization," *Operations Research Letters*, vol. 35, no. 5, pp. 573–580, 2007, http://dx.doi.org/10.1016/j.orl.2006.11.005

[50] R. Boys and D. Henderson, "A Bayesian approach to DNA sequence segmentation," *Biometrics*, vol. 60, pp. 573–588, 2004, http://dx.doi.org/10.1111/j.0006-341X.2005.040701_1.x

[51] J. Braun, R. Braun, and H. Müller, "Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation," *Biometrika*, vol. 87, no. 2, pp. 301–314, 2000, http://dx.doi.org/10.1093/biomet/87.2.301

[52] The MHC Sequencing Consortium, "Complete sequence and gene map of a human major histocompatibility complex," *Nature*, vol. 401, no. 6756, pp. 921–923, 1999, http://dx.doi.org/10.1038/44853

[53] M. Priyadarshana, and G. Sofronov, "Breakpoint (R-package);" software available at http://cran.r-project.org/web/packages/breakpoint.