The Korean Communications
in Statistics Vol. 12 No. 1, 2005
pp. 139-147

Binary Segmentation Procedure for Detecting Change Points in a DNA Sequence

Tae Young Yang[1] and Jeongjin Kim[2]

Abstract

It is interesting to locate homogeneous segments within a DNA sequence. Suppose that the DNA sequence has segments *within* which the observations follow the same residue frequency distribution, and *between* which observations have different distributions. In this setting, change points correspond to the end points of these segments. This article explores the use of a binary segmentation procedure in detecting the change points in the DNA sequence. The change points are determined using a sequence of nested hypothesis tests of whether a change point exists. At each test, we compare no change-point model with a single change-point model by using the Bayesian information criterion. Thus, the method circumvents the computational complexity one would normally face in problems with an unknown number of change points. We illustrate the procedure by analyzing the genome of the bacteriophage *lambda*.

Keywords : Bayesian information criterion, bacteriophage lambda, binary segmentation procedure

1. Introduction

The deoxyribonucleic acid (DNA) sequences are the basic information carriers of a complete organism. Vast amounts of DNA sequences data are currently available for analysis, primarily as a result of large-scale sequencing projects such as the Human Genome Project. Consequently, there is an increasing need to develop efficient computational and statistical tools to analyze the profusion of biological data. The DNA sequences can be characterized by sequences of 4 alphabets called residues or nucleotides: adenine (A), cytosine (C), thymine (T) and guanine (G). It is interesting to locate homogeneous segments *within* the DNA sequences. Suppose that the DNA sequence has segments *within* which the observations follow the same residue frequency distribution, and *between* which observations have different distributions. In

1) Department of Mathematics, Myongji University, Yongin, Kyunggi, Korea 449-728.
E-mail: tyang@mju.ac.kr
2) Department of Mathematics, Myongji University, Yongin, Kyunggi, Korea 449-728.

this setting, change points correspond to the end points of these segments. More precisely, we represent the observations along the sequence as $Y_1, \cdots, Y_n$, where $Y_i (i=1, \cdots, n)$ represents one of the values of the DNA alphabets (A, C, T or G). The observations $Y_1, \cdots, Y_n$ are taken to be split into $R+1$ contiguous segments by the values $0 = \tau_0 < \cdots < \tau_R < \tau_{R+1} = n$. The observations $Y_{\tau_{r-1}+1}, \cdots, Y_{\tau_r}$ are supposed to be identically distributed. The number of change points R is usually unknown. Then the sequence can be put into the framework of the multiple change-point problems, which can be broken down into the problem of determining how many change points exist in the sequence and of determining the locations of these points.

Given the sequence, the most change points we can detect is n. Therefore, we assume that R is at most n. The interesting question to ask is whether there exist nested sub-models with fewer change points that can fit the sequence reasonably well. The Bayes factor or the Bayesian Information Criterion approximation to the Bayes factor are commonly used for model selection. The computation of them gets more and more complex as R increases, and is essentially infeasible for a large model with many change points. To circumvent the computational difficulties, a binary segmentation procedure is adopted. The procedure converts the model selection problem into several nested hypothesis testing problems; in each step we only need to compare no change-point model to a model with exactly single change point. Therefore, the procedure is easily implemented and circumvents the computational complexity that we would normally face in problems with a variable number of change points.

We explore the binary segmentation procedure. The procedure is a recursive partitioning tool. Roughly speaking, the procedure begins by tentatively dividing data into two parts. The hypothesis of commonality between the two parts is tested and the procedure terminates if commonality exists. If commonality does not exist, then the division takes place and the procedure is continued on each of the two parts. Moreover, when we determine there is no single change point in a part, we don't need to continue testing for the data in that part. This cuts the sample size down quite significantly for locating change points in the remaining part. In the testing step, we identify the two distinct parts for which the likelihood is maximized. Once the parts are identified, testing can be carried out using the Schwarz criterion (Schwarz 1978). If the criterion is positive, then the null hypothesis is rejected and the DNA sequence is split accordingly.

Binary segmentation procedures have been considered by various authors. Scott and Knott (1974), and Chen and Gupta (1997) developed methods to split normal data into homogeneous groups. Subsequently, Braun and Müller (1998), Yang and Kuo (2001), Yang (2004), and Yang and Swartz (2005) developed binary segmentation procedures for locating change points with respect to DNA sequencing, homogeneous Poisson processes, sporting performances, spatial intensity and quantal response curves respectively. In Yang (2004) and Yang and Swartz (2005), binomial models are the main focus. This paper considers change points over

multinomial models. Consistency issues related to binary segmentation have been studied by Vostrikova (1981) who proved consistency for locating the number of change points in a multi-dimensional random process under mild conditions. Venkatraman (1992) addressed consistency issues for the procedure when the change points are allowed to approach one another.

We illustrate the procedure by analyzing the genome of the bacteriophage *lambda*, a parasite of the intestinal bacterium *Escherichia coli*. This virus has become a benchmark sequence for the comparison of segmentation algorithms since the experimental segmentation based on gradient centrifugation of its C and G residue by Skalka, Burgi, and Hensley (1968). Its genome is relatively small at 48,502 residues in length, though this is long enough to provide an adequate challenge for the methods in this article. The complete genome sequence is stored in the GenBank sequence database under accession number J02459 and can be obtained from the National Center for Biotechnology Information (NCBI) web site at www.ncbi.nlm.nih.gov. Several statistical techniques have been developed in an attempt to identify these homogeneous DNA segments, many of which are reviewed in Braun and Müller (1998). Other recent works include the Bayesian approach of Liu and Lawrence (1999) and the quasi-likelihood method of Braun, Braun, and Müller (2000), both of which use a multiple change-point framework with the change points delimiting the segments.

An alternative general procedure for classifying data into categories is the method of classification and regression trees (CART); see Breiman et al. (1984). In the tree-based approach, it is generally considered appropriate to first partition the data completely, and second, to prune segmentations based on some cost-complexity measure. However, the binary segmentation procedure terminates the partitioning when the partition process fails to attain a threshold of some target criterion. An advantage of the binary segmentation approach over CART is its simplicity with respect to computation. Yang (2005) provides a tree-based method for grouping multinomial data according to their classification probability vectors. Yang is illustrated on grouping many DNA sequences.

Another general approach which can be used in partitioning problems is mixture modelling. Mixture modelling requires the specification of parametric models whereas the recursive approaches considered in this paper are often described as nonparametric. When the number of components is unknown (which is the case in the problems considered here), mixture modelling becomes more challenging and often requires Markov chain Monte Carlo methods for parameter estimation. An introduction to mixture modelling is given by Titterington, Smith and Makov (1985). Kim and Mallick (2002), van Dyk and Hans (2002) and Shlattmann, Gallinat and Bohning (2002) provide examples of mixture modelling approaches.

2. BIC Criterion

We represent the observations of a DNA sequence as $Y_1, \cdots, Y_n$. Suppose that the data

(n_a, n_c, n_t, n_g) of the residue frequencies have a multinomial distribution with the sample size $n=n_a+n_c+n_t+n_g$ and the classification probability $\boldsymbol{p}=(p_a, p_c, p_t, p_g)$. We let the null hypothesis H_0 denote the constant model with no change points. Under H_0, the likelihood is proportional to

$$L_0(\boldsymbol{p})=\frac{n!}{n_a!\ n_c!\ n_t!\ n_g!}\ p_a^{n_a}p_c^{n_c}p_t^{n_t}p_g^{n_g}$$

which is maximized at

$$(\hat{p}_a,\ \hat{p}_c,\ \hat{p}_t,\ \hat{p}_g)=\left(\frac{n_a}{n},\ \frac{n_c}{n},\ \frac{n_t}{n},\ \frac{n_g}{n}\right).$$

Let H_1 denote the single change-point model with the change point given by the parameter τ. Let $\boldsymbol{m}=(m_a, m_c, m_t, m_g)$ and $\boldsymbol{l}=(l_a, l_c, l_t, l_g)$ denote the residue frequencies of $Y_1, \cdots, Y_\tau$ and $Y_{\tau+1}, \cdots, Y_n$ respectively. The hypothesis H_1 implies $\boldsymbol{p}=\boldsymbol{q}$ if $1\le i\le\tau$; $\boldsymbol{p}=\boldsymbol{r}$ if $\tau+1\le i\le n$, where $\boldsymbol{q}=(q_a, q_c, q_t, q_g)$, $\boldsymbol{r}=(r_a, r_c, r_t, r_g)$, $q_a+q_c+q_t+q_g=1$, and $r_a+r_c+r_t+r_g=1$. Under H_1, the likelihood is proportional to

$$L_1(\boldsymbol{q}, \boldsymbol{r}, \tau)=\frac{\tau!}{m_a!\ m_c!\ m_t!\ m_g!}\ q_a^{m_a}q_c^{m_c}q_t^{m_t}q_g^{m_g}\times\frac{(n-\tau)!}{l_a!\ l_c!\ l_t!\ l_g!}\ r_a^{l_a}r_c^{l_c}r_t^{l_t}r_g^{l_g},$$

which is maximized for fixed $\tau=1, \cdots, n-1$ via

$$\{(\hat{q}_a, \hat{q}_c, \hat{q}_t, \hat{q}_g), (\hat{r}_a, \hat{r}_c, \hat{r}_t, \hat{r}_g)\}=\left\{\left(\frac{m_a}{\tau}, \frac{m_c}{\tau}, \frac{m_t}{\tau}, \frac{m_g}{\tau}\right), \left(\frac{l_a}{n-\tau}, \frac{l_c}{n-\tau}, \frac{l_t}{n-\tau}, \frac{l_g}{n-\tau}\right)\right\}.$$

The fully maximized likelihood under the single change-point model $L_1(\hat{\boldsymbol{q}}, \hat{\boldsymbol{r}}, \hat{\tau})$ is obtained by maximizing $L_1(\hat{\boldsymbol{q}}, \hat{\boldsymbol{r}}, \tau)$ over the finite set $\tau=1, \cdots, n-1$.

Once the parts are identified, testing can be carried out using the Schwarz criterion also known as BIC. Our decision to select the non-null model H_1 over H_0 involves choosing H_1 if

$$\text{BIC}=L_1(\hat{\boldsymbol{q}}, \hat{\boldsymbol{r}}, \hat{\tau})-L_0(\hat{\boldsymbol{p}})-\frac{1}{2}(d_1-d_0)\log(n)\ >\ 0, \qquad (2.1)$$

where the third term in (2.1) is a penalty function which adjusts for the difference in dimensionality between the two models. Note that $d_1=7$ corresponds to the number of unknown parameters (one change point τ, each three classification probabilities for $\boldsymbol{q}=(q_a, q_c, q_t, q_g)$ and $\boldsymbol{r}=(r_a, r_c, r_t, r_g)$) and $d_0=3$ corresponds to $\boldsymbol{p}$. Asymptotically, as the sample size n increases, BIC in (2.1) gives rough approximation to the logarithm of the Bayes factor without having to specify the priors of unknown parameters (Kass and Raftery 1995). Raftery (1995) explores grades of evidence corresponding to BIC; weak evidence for 0-2, positive evidence for 2-6, strong evidence for 6-10, very strong evidence for >10. In this application, if BIC in (2.1) is negative, the decision is to accept the constant model H_0.

We note that the change point τ in equation (2.1) is discrete. The asymptotic theories for the Schwarz information criterion do not apply if parameters are discrete (Raftery 1995). However, we have set as it were continuous. Therefore, BIC in equation (2.1) is at best an approximation; see Roeder and Wasserman (1997). BIC in equation (2.1) works well in practice even if it is not theoretically justified. Simulation studies in Yang and Swartz (2005) and Roeder and Wasserman (1997) yield satisfactory results.

Alternative choice to BIC includes Akaike Information Criterion (Akaike 1973) which is given by

$$\text{AIC} = L_1(\hat{q}, \hat{r}, \hat{\tau}) - L_0(\hat{p}) - \frac{1}{2}(d_1 - d_0) \quad > \quad 0. \tag{2.2}$$

As can be seen by comparing BIC in (2.1) and AIC in (2.2), these two criteria differ only in that the coefficient multiplies the number of unknown parameters; in other words, the criteria differ by how strongly they penalize large models. In general, models chosen by BIC will be more parsimonious than those chosen by AIC. The AIC has shown to overestimate the number of parameters in a model (Kadane and Lazar 2004).

Although the BIC or the AIC are convenient, it may sometimes be possible for the experimenter to quantify the significant level of α for each testing. Using standard likelihood ratio test procedures, we reject H_0 at level α when

$$\lambda = L_1(\hat{q}, \hat{r}, \hat{\tau}) - L_0(\hat{p}) > \frac{1}{2}\chi^2_{2,\alpha}$$

where $\hat{\tau}$ is similarly obtained by maximizing $L_1(\hat{q}, \hat{r}, \tau)$ over the finite set $\tau = 1, \cdots, n-1$; $\chi^2_{r,1-\alpha}$ is the $(1-\alpha)100\%$ quantile of the chi-square distribution with r degrees of freedom. This alternative testing procedure requires the experimenter to specify α. Note that making α bigger may increase the number of change points. We may also use the Bonferroni adjustment for multiple testing. The Bonferroni test is conservative in sense that the probability of rejecting at least one hypothesis incorrectly is less than its familywise error rate due to the Bonferroni inequality. This implies that the Bonferroni test does not reject hypotheses as often as it should and therefore lacks power. The Bonferroni-Holm test and the Sidak-Holm test (Holm 1979) allow more rejections, and are therefore less conservative and more powerful than the Bonferroni test.

At this point, the binary segmentation procedure readily presents itself. If H_0 is accepted, the algorithm terminates. However, if H_1 is selected, the data set is divided into the two subsegments given by $(Y_1, \ldots, Y_{\hat{\tau}})$ and $(Y_{\hat{\tau}+1}, \ldots, Y_n)$. The test of a single change point is then carried out on each of the two subsegments. The algorithm continues in this fashion and terminates when no more splitting takes place. We note that the order in which subsegments are divided does not affect the subsequent inference. We also note that the algorithm may be thought of as a forward selection procedure where partitioning continues

until the algorithm terminates. Interestingly, if one applied a backwards step, where given a model, the one with the lowest BIC were chosen, we would always choose the same model from which we came. Hence, a backward elimination procedure or a stepwise procedure based on the same decision criterion would provide the same inferences as our algorithm (Yang and Swartz 2005).

3. Numerical Example

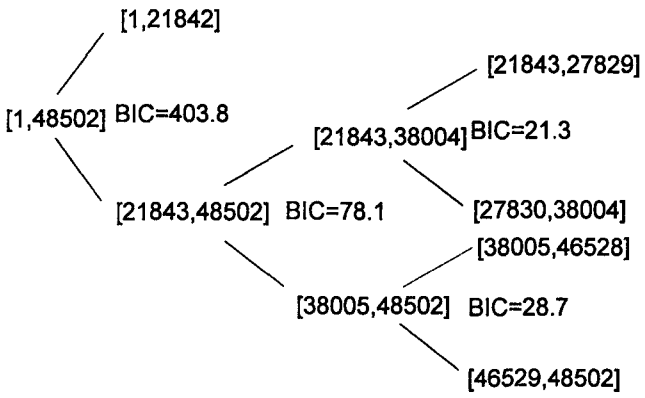
We apply the proposed binary segmentation procedure to the DNA sequence of 48,502 residues for the bacteriophage *lambda*. Figure 1 presents the step by step result of the procedure using the BIC. The procedure begins in step 1 by identifying the first candidate change point. The value is the 21,842th position and this tentatively divides the full sequence [1,48502] into two subsegments of [1,21842] and [21843,48502]. The calculated BIC for this split is 403.8, since this is positive, the split is accepted. In step 2, the first subsegment [1,21842] is further divided according to the new change point. This time, the BIC is negative and the split is rejected. In step 3, the second subsegment [21843,48502] is further divided into the candidate change point of 38,004. The corresponding BIC is 78.1 and the split is accepted. The data has now been divided according to [1,21842], [21843,38004] and [38005,48502]. We continue in this fashion until no more splits are accepted. At the completion of the procedure, the groupings are [1,21842], [21843,27829], [27830,38004], [38005,46528], and [46529,48502]. The corresponding estimated residue probabilities are given in Table 1.

<Table 1> Estimated base probabilities of $\boldsymbol{p}=(p_a, p_c, p_t, p_g)$ on each segmentation by using the BIC criterion.

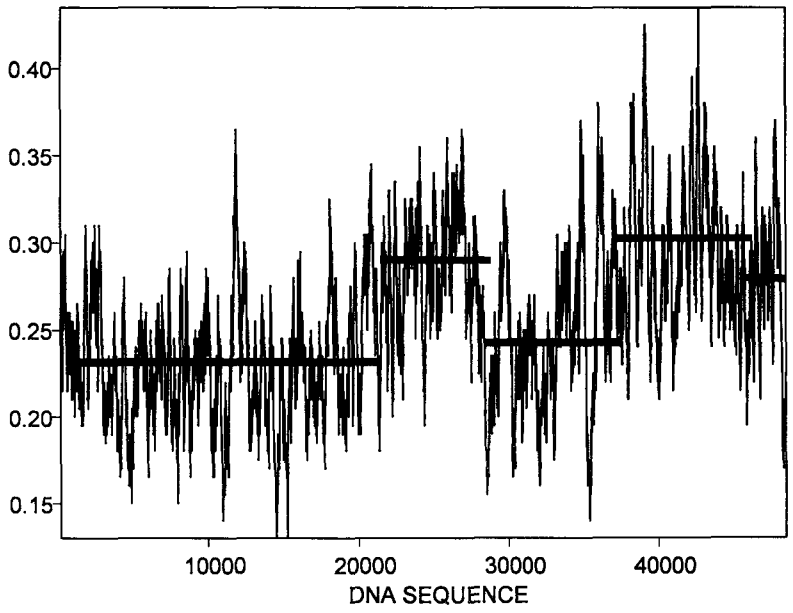
	[1,21842]	[21843,27829]	[27830,38004]	[38005,46528]	[46529,48502]
$\hat{p}_a$	0.230	0.289	0.248	0.296	0.270
$\hat{p}_c$	0.254	0.186	0.237	0.227	0.181
$\hat{p}_t$	0.201	0.338	0.301	0.217	0.331
$\hat{p}_g$	0.315	0.187	0.214	0.260	0.218

It is interesting to compare the results of the binary segmentation based on the BIC with the results from the Skalka, Burgi and Harshley (1968) segmentation model for the bacteriophage *lambda*; see Braun and Müller (1998). Two methods give the similar change points. Figure 2 provides a scan analysis of the bacteriophage *lambda*. A moving average of residue A with a 1000-base window centered at the current position shows similar pattern with the groupings from the procedure. In Figure 2, we add the straight lines denoting $\hat{p}_a$

according to Table 1.



[Figure 1] Step by step results of the binary segmentation procedure using the BIC criterion for splitting the DNA sequence of the bacteriophage lambda.



[Figure 2] Moving probabilities of residue A against estimated base probabilities of $\hat{p}_a$ from the BIC criterion.

References

[1] Akaike, H. (1973). Information measures and model selection, *Bulletin of the International Statistical Institute,* Vol. 50, 277-290.

[2] Braun, J.V. and Müller, H. (1998). Statistical methods for DNA sequence segmentation, *Statistical Science,* Vol. 13, 142-162.

[3] Braun, J.V., Braun, P.K. and Müller, H. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation, *Biometrika,* Vol 87, 301-314.

[4] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees,* Wadworth and Brooks/Cole, Monterey.

[5] Chen, J. and Gupta, A. (1997). Testing and locating variance change points with applications to stock prices, *Journal of the American Statistical Association,* Vol. 92, 739-747.

[6] Holm, S. (1979). A simple sequentially rejective Bonferroni test procedure, *Scandinavian Journal of Statistics.*Vol. 6, 65-70.

[7] Kadane, J.B. and Lazar, N.A. (2004). Methods and criteria for model selection, *Journal of the American Statistical Society,* Vol. 99 279-290.

[8] Kass, R.E. and Raftery, A.E. (1995). Bayes factor, *Journal of the American Statistical Association,* Vol. 90, 773-795.

[9] Kim, H. and Mallick, B.K. (2002). *Analyzing spatial data using skew-Gaussian processes, In Spatial Cluster Modelling*, A. Lawson and D. Denison (editors). Chapman and Hall, London, 163-173.

[10] Liu, J.S. and Lawrence, C.E. (1999). Bayesian inference on bipolymer models, *Bioinformatics,* Vol. 15, 38-52.

[11] Raftery, A. (1995). Bayesian model selection in social research, In *Sociological Methodology*, Marsden P(ed). Blackwells, Cambridge, 111-196.

[12] Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals, *Journal of the American Statistical Association,* Vol. 92, 894-902.

[13] Schlattmann, P., Gallinat, J. and Bohning, D. (2002). Spatio-temporal partition modelling: an example from neurophysiology, *In Spatial Cluster Modelling*, A. Lawson and D. Denison (editors). Chapman and Hall, London, 227-234.

[14] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, Vol. 6, 461-464.

[15] Scott, A. and Knott, M. (1974). Cluster analysis method for grouping means in the analysis of variance, *Biometrics,* Vol. 30, 507-512.

[16] Skalka, A., Burge, E. and Hershey, A.D. (1968). Segmental distribution of nucleotides in the DNA of bacteriophage lambda, *Journal of Molecular Biology*, Vol. 34, 1-16.

[17] Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions,* Wiley, New York.

[18] van Dyk, D.A. and Hans, C.M. (2002). Accounting for absorption lines in images obtained with the Chandra X-ray Observatory, *In Spatial Cluster Modelling*, A. Lawson and D. Denison (editors). Chapman and Hall, London, 175-198.

[19] Venkatraman, E.S. (1992). *Consistency results in multiple change-point situations*, Unpublished PhD Thesis, Department of Statistics, Stanford University.

[20] Vostrikova, L.J. (1981). *Detecting 'disorder' in multidimensional random processes*, Soviet Mathematics Doklady, Vol. 24, 55-59.

[21] Yang, T.Y. and Kuo, L. (2001). Bayesian binary segmentation procedure for a Poisson process with multiple changepoints, *Journal of Computational and Graphical Statistics*, Vol. 10, 772-785.

[22] Yang, T.Y. (2004). Bayesian binary segmentation procedure for detecting streakiness in sports, *Journal of the Royal Statistical Society Series A*, Vol. 167, 627-637.

[23] Yang, T.Y. (2005). A tree-based model for homogeneous groupings of multinomials, *Statistics in Medicine,* in press.

[24] Yang, T.Y. and Swartz, T. (2005). Applications of binary segmentation to the estimation of quantal response curves and spatial intensity. *Biometrical Journal*, in press.

[Received December 2004, Accepted February 2005]