



## Multiple-Changepoint Testing for an Alternating Segments Model of a Binary Sequence

Aaron L. Halpern

*Biometrics*, Vol. 56, No. 3. (Sep., 2000), pp. 903-908.

Stable URL:

<http://links.jstor.org/sici?sici=0006-341X%28200009%2956%3A3%3C903%3AMTFAAS%3E2.0.CO%3B2-J>

*Biometrics* is currently published by International Biometric Society.

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/ibs.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

## CONSULTANT'S FORUM

# Multiple-Changepoint Testing for an Alternating Segments Model of a Binary Sequence

Aaron L. Halpern

Department of Molecular Genetics and Microbiology, Health Sciences Center,  
University of New Mexico, 915 Camino de Salud, Albuquerque, New Mexico 87131, U.S.A.  
*Current address:* Informatics Research, Celera Genomics,  
45 West Gude Drive, Rockville, MD 20850, U.S.A.  
*email:* aaron.halpern@celera.com

**SUMMARY.** A binary sequence may give the appearance of being composed of alternating segments with relatively high and relatively low probability of success. Determining whether such an alternating pattern is significant is a multiple-changepoint problem where the number of segments and their success probabilities are unknown, with the added constraint of segment alternation. A dynamic programming method for determining the optimal segmentation into a given number of segments is provided. Given this, a variation on the simulation method of Venter and Steel (1996, *Computational Statistics and Data Analysis* **22**, 481–504) may be employed to test the null hypothesis of a homogeneous sequence as well as to estimate the number and location of changepoints. A sample application, the assessment of the possibility of genetic recombination in HIV sequences, is presented.

**KEY WORDS:** Bernoulli random variables; HIV; Monte Carlo; Multiple changepoint; Recombination.

### 1. Introduction

Given a binary sequence (a sequence of zeros and ones) corresponding to a sequence of independent observations, we may wish to evaluate the hypothesis that the sequence is composed of two sorts of segments, alternating with one another, that differ in the probability that a given position is a one; i.e., positions in the sequence may be drawn from two categories having different probabilities of success,  $\theta_1$  and  $\theta_2$ . (These notions will be made more precise below.) The number of segments is not known in advance. Venter and Steel (1996) provided simulation-based methods for solving a related problem, that of a sequence of continuously valued observations that may be composed of some unknown number of segments, each of which has its own mean. It is shown here that related methods provide a solution for the alternating segments problem for binary sequences. Previous approaches to multiple-changepoint estimation are discussed toward the end of the paper.

### 2. Example Application

One statistical problem in the analysis of genetic sequences that has arisen recently is that of determining whether a given strain of a virus is likely to have arisen as the result of recombination between two donor strains. Because of possible implications for pathogenesis and treatment of the virus, this question has been of substantial interest in connection with

the study of diversity of the human immunodeficiency virus (HIV). Although in this paper we will focus on the case of viral recombination, similar issues may arise in examining bacterial or eukaryotic genomes as well.

The genetic material of two genetically distinct viruses may come into physical contact in such a way as to produce a new virus whose genetic sequence consists of alternating segments of DNA (or RNA) that correspond to pieces of the original viruses; this creation of a genetic chimera is known as recombination. Such contact may arise as the result of coinfection by two different strains in a natural infection, as the result of a laboratory artifact, or as the result of an instance of genetic engineering.

Most isolates of HIV-1, the predominant form of the virus found in AIDS patients worldwide, fall into one of several genetically distinct clusters known as subtypes. Since viruses replicate in an asexual fashion, these subtypes for the most part evolve independently. However, recombination between two strains representing different subtypes of HIV-1 has been established in the laboratory as well as in natural infections in cases where the epidemiology of a dual infection can be carefully documented (Diaz et al., 1995; Salminen et al., 1997). Recombination is of significant medical interest because of its relevance to antigenic diversity, vaccine efficacy, (multi-)drug resistance, and epidemiologic issues such as serial infection.

Given access to the sequence of the donor strains and a putative recombinant, determining whether the putative recombinant did or did not exactly match segments from each of the donor strains would be trivial. However, in viruses such as the human immunodeficiency virus (HIV), the rate of accumulation of point mutations makes the isolation of a recombinant virus and copies of its exact parental donors unlikely except in unusual cases where the epidemiology of infection can be documented clearly. (Even then, there may be intervening mutations.) In other viruses, we may be interested in ancient recombination events that again are clouded by accumulated mutations. In such circumstances, it may nonetheless be possible to establish a probable recombinant ancestry for a given isolate if its pattern of genetic relatedness to two or more other isolates forms a sort of patchwork, with some regions most similar to one strain and other regions most similar to another strain.

If we align a potential recombinant sequence to two potential parental sequences, we may derive a binary match sequence that indicates those positions at which the potential recombinant matches only one of the parental sequences, excluding other positions as uninformative, as

```

Query:  ACGAGATAGACGATAGGCGATAGACTGGACGATACGATACGATACGA
A       : AGGAGTCAGCGGATGGGAGCCAGACTGGCCATACGATACCATACCA
B       : ACGACAAAGACGATAGGCGATAGATCAGACTGTACAGTATCATATTA
        -1--01---11---1--1-11---000-1--0---00--0---0--
    
```

(1)

In this example, the potential recombinant query matches parent B but not A at nine positions in total, of which eight are clustered to the left portion of the sequence, while it matches A but not B at nine positions, clustered mostly to the right end of the sequence. Such a pattern might lead us to hypothesize that A and B are relatives of parental viruses that recombined to give rise to an ancestor of the query sequence. There are variations on this method of reducing an alignment of sequences to a binary pattern involving the inclusion of additional sequences; by excluding positions that do not clearly support greater similarity to A or to B, variation due to differences in sequence variability along the length of the alignment may be reduced (Robertson, Hahn, and Sharp, 1995a).

In general, the process of recombination may give rise to an offspring that is derived from alternating segments of the parental sequences, ABABA... In evaluating a potential recombinant sequence, one possible test is to ask whether we can reject a null hypothesis of uniformly distributed zeros and ones in our match sequence. Previous work on this question within the HIV literature has been limited to the single-changepoint case, with (invalid) application of single-changepoint statistics to adjacent pairs of segments in sequences where more than one changepoint is suspected.

### 3. Testing for Multiple Changepoints

Let  $X$  be a sequence of independent Bernoulli random variables,  $x_1 \cdots x_l$ , with  $p_i = \Pr(x_i = 1)$ . Our null hypothesis is that the sequence is homogeneous, with  $p_i = \Pr(x_i = 1) = \theta$  for  $1 \leq i \leq l$ . Consider this against the alternative hypothesis that the sequence is composed of between two and  $N$  segments, with the probability of a one at a position in an

odd-numbered segment being  $\theta_1$  and the probability of a one at a position in an even-numbered segment being  $\theta_2$ ,  $\theta_1 \neq \theta_2$ , i.e.,

$H_a: \exists$  an integer  $n, 2 \leq n \leq N$ , and integers  $\tau_1, \dots, \tau_{n-1}$  such that

- (i)  $0 < \tau_1 < \dots < \tau_{n-1} < l$ ,
- (ii)  $p_i = \pi_h$  for  $\tau_{h-1} < i \leq \tau_h$ ,
- (iii)  $\pi_h = \begin{cases} \theta_1 & \text{for } h \text{ odd} \\ \theta_2 & \text{for } h \text{ even;} \end{cases}$  (2)

$\tau_1, \dots, \tau_{n-1}$  are changepoints; we define  $\tau_0 = 0$  and  $\tau_n = l$  for convenience. We speak of segment  $h$  extending from  $[\tau_{h-1} + 1]$  to  $\tau_h$ , inclusive. If  $h$  is odd, we say segment  $h$  is an odd-numbered segment, else it is even numbered.

For an observed binary sequence of length  $l$ , containing a given number of ones, we may ask three questions: Assuming that a sequence is to be cut into  $n$  alternating segments, where should the segment boundaries (changepoints) be located? Should a given segment be divided into 2, 3, ..., or  $N$  segments, possibilities that we denote  $H_2, \dots, H_N$ ? Finally, should we reject  $H_0$ ?

We proceed in order. Assuming that a given sequence is composed of  $n$  segments, we wish to estimate  $\tau_1, \dots, \tau_{n-1}$ . First, we define a measure of the fit of a given segmentation. Let  $\sigma_n$  denote a segmentation (a specification of changepoints) involving  $n$  segments. For a given segmentation specifying  $\tau_1, \dots, \tau_{n-1}$ , construct a two-by-two table of counts of ones and zeros in the odd- and even-numbered segments. Letting  $\mathcal{O}(\sigma_n, i) = 1$  if position  $i$  occurs in an odd-numbered segment according to  $\sigma_n$  and zero otherwise, we define  $a(\sigma_n)$ ,  $b(\sigma_n)$ ,  $c(\sigma_n)$ , and  $d(\sigma_n)$ , the four elements of a  $2 \times 2$  table of the number of ones and zeros occurring in the odd- and even-numbered segments of the given sequence and segmentation, i.e.,

	odd	even
1	$a(\sigma_n)$	$c(\sigma_n)$
0	$b(\sigma_n)$	$d(\sigma_n)$

where

$$\begin{aligned}
 a(\sigma_n) &= \sum_{i=1}^l x_i \mathcal{O}(\sigma_n, i) \\
 b(\sigma_n) &= \sum_{i=1}^l (1 - x_i) \mathcal{O}(\sigma_n, i) \\
 c(\sigma_n) &= \sum_{i=1}^l x_i (1 - \mathcal{O}(\sigma_n, i)) \\
 d(\sigma_n) &= \sum_{i=1}^l (1 - x_i) (1 - \mathcal{O}(\sigma_n, i)).
 \end{aligned}$$

On such a table (i.e., for a given segmentation), we may estimate  $\theta_1$  and  $\theta_2$  as

$$\hat{\theta}_1 = a(\sigma_n) / (a(\sigma_n) + c(\sigma_n)) \tag{3}$$

$$\hat{\theta}_2 = b(\sigma_n) / (b(\sigma_n) + d(\sigma_n)). \tag{4}$$

From this table, we may also calculate any of several measures of the degree to which ones and zeros are distributed randomly in the segments defined by the given segmentation. For example, we may determine values for chi-square (equivalent to the mean-squared error measure used by Venter and Steel (1996), as well as an extension of the Anderson–Darling statistic to multiple changepoints), Fisher’s exact test, likelihood ratios under the null hypothesis and under the alternative hypothesis indicated by the given segmentation, or a generalization of the Kolmogorov–Smirnov two-sample statistic. These measures have been employed previously for the problem of detecting a single changepoint in a binary sequence (see Halpern, 1999 for discussion of differences between measures).

For concreteness and comparability to previous literature on recombination, consider the use of the chi-square. For a given sequence  $X$  and segmentation  $\sigma_n$ , let

$$\chi^2(\sigma_n) = \chi^2(a(\sigma_n), b(\sigma_n), c(\sigma_n), d(\sigma_n)), \quad (5)$$

where  $\chi^2(a, b, c, d)$  is the standard chi-square for a  $2 \times 2$  table. We then define an optimal segmentation as well as the maximal chi-square as follows. Let  $\sigma_n^*$  be that segmentation  $\sigma_n$  that maximizes  $\chi^2(\sigma_n)$  and let

$$\chi_n^2 = \chi^2(\sigma_n^*). \quad (6)$$

We choose  $\sigma_n^*$  to be our estimate of the boundaries for the segmentation of our sequence into  $n$  segments, i.e.,  $\sigma_n^*$  is the answer to our first question. (An efficient algorithm for determining  $\sigma_n^*$  is given in the Appendix.) The degree of support for this segmentation compared with the null hypothesis is measured by the value  $\chi_n^2$ .

Now for our second question: How do we choose among  $H_2, \dots, H_N$ ? Define  $P_n$  to be the fraction of random sequences of the same length and composition as the sequence of interest with  $\chi_n^2$  at least as large as that for the sequence of interest. We then choose  $H_{n^*}$  such that  $P_{n^*} = \min\{P_n : 2 \leq n \leq N\}$ ; for notational convenience, let this value be  $P^N$ . The motivation behind this is that  $P_n$  is a measure of how surprising or unlikely a pattern involving  $n$  segments is under  $H_0$ . By choosing  $n^*$  to be the number of segments such that  $P_{n^*} = \min\{P_2, \dots, P_N\}$ , we choose that number of segments that gives the fit hardest to explain as the result of chance.

In practice, one aspect of this approach is not always feasible. For sequences for which  $H_0$  is clearly false, two or more of  $P_2, \dots, P_N$  may be too small to be practically estimated via simulations; we are able to say that they are small but are not able to say which is the smallest. This does not substantially interfere with the estimation of  $P^N$ , but it does interfere with the selection of  $n^*$ . Although this is essentially a model-choice problem, standard model-choice methods such as the Akaike Information Criterion (AIC) are not directly applicable because of the optimization involved in the choice of changepoint locations (Auger and Lawrence, 1989) and because of the alternating segments character of the model. As illustrated in Figure 1,  $E(\chi_n^2) - E(\chi_{n-1}^2)$  decreases progressively but is considerably uneven if the number of ones is much smaller or larger than the number of zeros. A similar plot of log likelihoods of sequences given optimal segmentations and MLE parameter values could be constructed, illustrating the difficulty with the AIC. However, a variation on the AIC

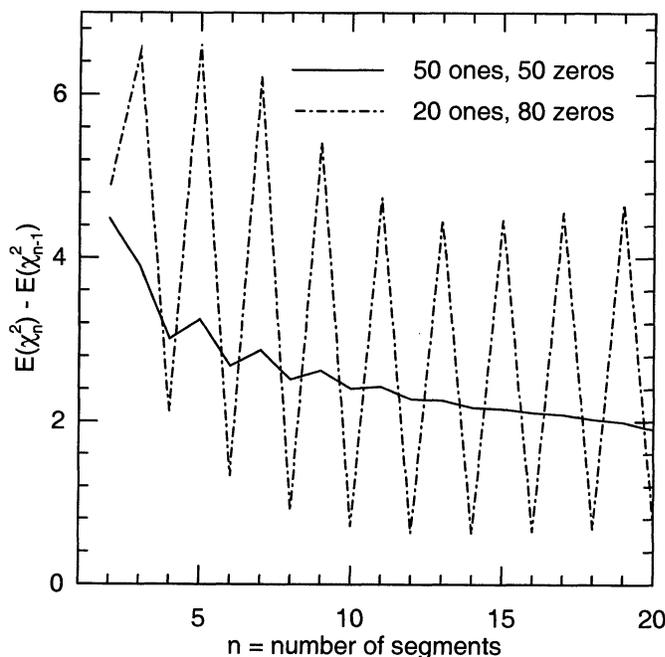


Figure 1. Expected changes in  $\chi_n^2$  for increasing  $n$ .

involving an empirical fit to the model may be proposed. Estimating  $E(\chi_n^2)$  by the mean of  $\chi_n^2$  in the simulations conducted to estimate  $P_n$ , let

$$D_n = \chi_n^2 - E(\chi_n^2) \quad (7)$$

and choose  $n^* = i$  such that  $D_i = \max_n D_n$ .

Finally, our third question: Should we reject  $H_0$ ? As discussed by Venter and Steel (1996), the Bonferroni argument gives  $(N-1)P^N$  as a conservative estimate of the probability of obtaining at least as small a value of  $P^N$  as that of our sequence of interest by chance. For the current implementation, concerned with binary sequences with an alternating pattern of segment types, the bound might be acceptable for  $N$  less than four or five but becomes excessively conservative as  $N$  rises. In simulations not presented here, this bound gave as much as a fourfold overestimation of the  $p$ -value for  $N = 10$  and up to a ninefold overestimation for  $N = 20$ .

A more accurate estimate of significance may be obtained by the double-simulation method described in Appendix B of Venter and Steel (1996). Briefly, two sets of sequences are simulated, with the first providing estimates of the null distributions of each of  $\chi_2^2, \dots, \chi_N^2$ ; from this, we determine  $P_2, \dots, P_N$ , and thus  $P^N$ , for each of the sequences in the second set, as well as for the sequence of interest. The fraction of sequences in the second set with  $P^N \leq P^*$ , where  $P^*$  is the value of  $P^N$  for the sequence of interest, is our final estimate of a  $p$ -value.

The methods just described depend on being able to determine optimal segmentations for a sufficiently large number of sequences simulated under the null hypothesis. An efficient way of determining  $\sigma_n^*$  is presented in the Appendix.

**4. Illustration: Recombination in HIV**

Several studies have proposed that various isolates of HIV-1 and HIV-2 may be recombinants between the major groups of genetic variation, or subtypes. The possibility of multiple changepoints may have a large impact on assessment of significance, but several examples survive the closer scrutiny given here.

Consider the case of the HIV-1 isolate known as CI32. Robertson et al. (1995b) suggest segmenting the sequence of the *gag* gene from this virus into three pieces, with the first and last being (hypothesized to be) inherited from a virus of subtype A and the middle inherited from a virus of subtype D. Comparison of the sequence of CI32 to reference sequences from these two subtypes, plus an outgroup sequence, according to the methods similar to those illustrated in (1) above, yields the following: The numbers of positions supporting assignment to subtypes A and D, respectively, were 10 and 4 for the first segment, 0 and 4 for the second segment, and 21 and 7 for the third segment.

Robertson et al. (1995b) used simulations on adjacent pairs of segments to estimate significance, following Maynard Smith (1992) in using the maximum chi-square measure for single changepoints to compare the observed patterns to the best segmentations of random sequences. Considering the first and second segments, a *p*-value of 0.081 was estimated; for the second and third segments,  $p \approx 0.037$ . Both tests appear to be of borderline significance, and the interpretation that is naively tempting is that two borderline tests taken together indicate a significant pattern. However, as discussed further in Halpern (1999), this ignores the fact that not only the internal changepoint but also one of the ends of each pair of segments was optimized.

The methods described here may be used to take this into account. Based on the segmentation described above, we have  $\chi_3^2 = 9.0540$ . Evaluation of 100,000 random binary sequences of length 46 containing 31 zeros gives  $\Pr\{\chi_3^2 \geq 9.054\} \approx 0.325$ . Suppose we retrospectively and, admittedly arbitrarily, decide to allow segmentations involving between two and five segments: Comparison of 100,000 additional sequences to the initial 100,000 indicates that  $\Pr\{P^5 \leq 0.325\} \approx 0.57$ —hardly a significant result!

Should this case make us reluctant to accept any proposed examples of recombination that involve multiple changepoints? The remaining potential recombinants that Robertson et al. (1995b) suggest may involve two or more changepoints are summarized in Table 1, with significance values assigned by the method presented in this paper. Segmentations involving up to five segments were considered. For these examples and others in the HIV literature (not shown; see Robertson et al., 1995a; Salminen et al., 1995; Siepel et al., 1995; Cornelissen et al., 1996), the hypotheses of multiple-changepoint recombination are strongly supported by the data.

**5. Comparison to Alternative Methods**

Several methods for multiple-changepoint estimation have previously been proposed. For assessing recombination as proposed here, what is needed is a model involving alternation between two categories of segments whose parameters (here,  $\theta_1$  and  $\theta_2$ ) are estimated from the data; optimal, unbiased estimation of the best segmentation, both in number of changepoints and in their locations; and a test of the null hypothesis. None of the methods discussed below inherently provides a

test of the null hypothesis; to the author’s knowledge, there is no distributional theory for any of the methods. Monte Carlo simulations such as discussed above may be employed.

Auger and Lawrence (1989) introduced a dynamic programming method that can estimate multiple changepoints; however, it does not solve the problem of a model with alternating segment types and unknown segment parameters. Fu and Curnow (1990) proposed a related method for identifying the boundaries of alternating segments but required that the parameters  $\theta_1$  and  $\theta_2$  (their  $p_0$  and  $p_1$ ) be specified in advance. Both proposals offer heuristic approaches to the choice of the best number of segments. Indeed, Auger and Lawrence (1989) have a nice discussion of why the problem at hand does not readily fit methods designed for model selection in other contexts (e.g., the AIC, *F*-ratios, etc; see also Figure 1 here).

Green (1995) introduced a method that includes model choice in Markov chain Monte Carlo (MCMC) estimation. His sample application involved determining the number and locations of multiple changepoints. The method does not easily extend to an alternating segments model since the introduction or deletion of a segment in such a model reverses the category of all following segments.

Perhaps the clearest alternative to the current method is the hidden Markov model (HMM) approach described in Churchill (1989), which allows inference of multiple changepoints under an alternating segments model as well as estimation of segment parameters. The HMM method implicitly considers all possible numbers of changepoints and identifies a best number of segments. The method is very computationally efficient and can be applied to sequences of almost any length. In addition, it can be extended to deal with multiple segment types and larger alphabets. However, there are also certain disadvantages that make the dynamic programming method proposed here a useful complement. Although it generally works well in practice, the HMM method does not provide a guaranteed optimal segmentation due to the use of expectation maximization (EM) in estimating model parameters; convergence, both in terms of iterations and optimality, may especially suffer on sequences generated under the null hypothesis. The HMM method does not easily allow explicit comparison of different numbers of changepoints. For short sequences (length under 100), at least as implemented by the current author, it is slower than the method introduced here. Finally, the architecture of the HMM imposes certain biases on the estimation of changepoint number and location. One manifestation of these biases is that the HMM favors a single changepoint over multiple changepoints, as illustrated in the following example:

```
Input: 0001000100010001110111011100010001000111011101110111
HMM:   AAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
DP:    AAAAAAAAAAAAAABBBBBBBBBBAAAAAAAAAAAAABBBBBBBBBBBBBBB
rHMM:  AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBB
```

Given the input sequence, the Churchill HMM method gives the segmentation labeled HMM with two segments, while the current method gives the segmentation labeled DP with four segments ( $P_2 \approx 0.066$  and  $P_4 \approx 0.004$ ). When a single changepoint is inferred, this allows the second state of the HMM (that corresponding to the second segment) to have zero probability of a transition back to the first state. This example also

**Table 1**  
Evaluation of some putative HIV-1 recombinants

Isolate <sup>a</sup>	Region <sup>b</sup>	$n^c$	Table <sup>d</sup>		$\chi_n^2$ <sup>e</sup>	$\hat{P}_n$ <sup>f</sup>	$\Pr(P^5 \leq \hat{P}_n)$ <sup>g</sup>	
KE124	<i>gag+env</i>	5	Subtype		100.06	0/100,000	3/100,000	
			Segments	A				D
			Odd	72				8
			Even	4				60
MAL	<i>gag+env</i>	4	Subtype		60.92	0/100,000	3/100,000	
			Segments	A				D
			Odd	39				7
			Even	13				73
ZM184	<i>env</i>	4	Subtype		38.76	0/100,000	10/100,000	
			Segments	A				C
			Odd	1				19
			Even	46				9

<sup>a</sup> Common name of HIV-1 isolate suspected of being a recombinant.

<sup>b</sup> Gene(s) sequenced.

<sup>c</sup> Number of segments suspected.

<sup>d</sup> Table of counts of positions in odd and even segments matching one but not the other of the two hypothesized parental HIV-1 subtypes for optimal  $n$ -way segmentation.

<sup>e</sup> Chi-square for the optimized table of d.

<sup>f</sup>  $\hat{P}_n$ , the fraction of randomizations with at least as large a value of  $\chi_n^2$  as that in the preceding column. A value of zero indicates that no simulation achieved as large a value.

<sup>g</sup> Fraction of randomizations with  $P^5 \leq \hat{P}_n$ .

illustrates another point: The HMM method has a bias, in single changepoint estimates, toward having the changepoint occur early rather than late. Here, input is a sort of palindrome in which  $x_{l-i} = 1 - x_i$ . Consequently, the DP method will score two segmentations the same if the locations of the changepoints of the first segmentation ( $\tau_1 \cdots \tau_n$ ) and the second ( $\tau'_1 \cdots \tau'_n$ ) are related by  $\tau'_{n-i} = l - \tau_i + 1$ , as in the HMM and rHMM patterns above. However, the HMM gives substantially different likelihoods to these segmentations. While complications of the HMM (e.g., requiring transition probabilities  $[[1 - \alpha, \alpha], [\alpha, 1 - \alpha]]$ ) could provide solutions to these cases, such solutions introduce their own biases.

#### ACKNOWLEDGEMENTS

I wish to thank Carla Wofsy and Ed Bedrick for discussion and the anonymous reviewers for drawing my attention to several relevant papers on previous methods. All errors, of course, remain my own. Funding was provided by NIH grant 5P20-RR11830-02 as well as the Albuquerque High Performance Computing Center (AHPCC). Computational support was provided by the AHPCC.

#### RÉSUMÉ

Une séquence binaire peut donner l'apparence de segments avec alternativement une grande ou une faible probabilité de succès. Déterminer si une telle structure d'alternance est

significative constitue un problème de changement multiple où le nombre de segments et leur probabilité de succès sont inconnus, avec la contrainte supplémentaire d'alternance de segment. Une méthode de programmation dynamique pour déterminer la segmentation optimale en un nombre donné de segments est fournie. Etant donné ce résultat, une variation de la méthode de simulation de Venter et Steel (1996, *Computational Statistics and Data Analysis* **22**, 481–504) peut être employée pour faire le test de l'hypothèse nulle d'une séquence homogène, aussi bien que pour estimer le nombre et la position des changements. Comme application, nous présentons la détermination de la possibilité de recombinaison génétique dans des séquences du virus du SIDA.

#### REFERENCES

- Auger, I. E. and C. E. Lawrence. (1989). Algorithms for the optimal identification of segment boundaries. *Bulletin of Mathematical Biology* **51**, 39–54.
- Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**, 79–94.
- Cornelissen, M., Kampinga, G., Zorgdrager, F., Goudsmit, J., and The UNAIDS Network for HIV Isolation and Characterization. (1996). Human immunodeficiency virus type 1 subtypes defined by *env* show high frequency of recombinant *gag* genes. *Journal of Virology* **70**, 8209–8212.
- Diaz, R., Sabino, E., Mayer, A., Mosley, J., Busch, M., and

The Transfusion Safety Study Group. (1995). Dual human immunodeficiency virus type 1 infection and recombination in a dually exposed transfusion recipient. *Journal of Virology* **69**, 3273–3281.

Fu, Y.-X. and R. N. Curnow. (1990). Maximum likelihood estimation of multiple change points. *Biometrika* **77**, 563–573.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Halpern, A. L. (1999). Minimally selected  $p$  and other tests for a single abrupt changepoint in a binary sequence. *Biometrics* **55**, 1044–1050.

Maynard Smith, J. (1992). Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34**, 126–129.

Pearson, W. R. and Miller, W. (1992). Dynamic programming algorithms for biological sequence comparison. *Methods in Enzymology* **210**, 575–601.

Robertson, D., Hahn, B., and Sharp, P. (1995a). Recombination in AIDS viruses. *Journal of Molecular Evolution* **40**, 249–259.

Robertson, D., Sharp, P., McCutchan, F., and Hahn, B. (1995b). Recombination in HIV-1. *Nature* **374**, 124–126.

Salminen, M., Carr, J., Burke, D., and McCutchan, F. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Research and Human Retroviruses* **11**, 1423–1425.

Salminen, M., Carr, J., Robertson, D., Hegerich, P., Gotte, D., Koch, C., Sanders-Buell, E., Gao, F., Sharp, P., Hahn, B., Burke, D., and McCutchan, F. (1997). Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. *Journal of Virology* **71**, 2647–2655.

Siepel, A., Halpern, A., Macken, C., and Korber, B. (1995). A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Research and Human Retroviruses* **11**, 1413–1416.

Venter, J. and Steel, S. (1996). Finding multiple abrupt change points. *Computational Statistics and Data Analysis* **22**, 481–504.

Received February 1999. Revised October 1999.  
Accepted January 2000.

APPENDIX

Determining Optimal Segmentations

To enable the methods described above, a rapid way of determining  $\sigma_n^*$  is needed. The following calculation determines the optimal  $n$ -segment model for a given sequence via recurrence relationships. Programs implementing these analyses, written in ANSI C, are available from the author.

Conceptually, we subclassify segmentations involving  $n$  segments according to  $k$ , the number of positions they assign to odd segments. For given  $n$  and  $k$ , we determine, via recurrence relationships, segmentations that give the extreme (largest

and smallest) values of  $a$ , the number of ones assigned to odd segments. By evaluating the fit of these extreme segmentations for the possible values of  $k$ , the best overall segmentation into  $n$  segments may be determined.

Let  $S_{mn} = \{(\tau_1, \dots, \tau_{n-1})\}$  be the set of partial segmentations dividing the subsequent  $x_1, \dots, x_m$  into  $n$  segments with changepoints  $0 < \tau_1 < \dots < \tau_{n-1} < m$ . Let  $S_{mn}^k = \{s \in S_{mn} \mid a(s) + b(s) = k\}$ , where  $a(s)$  and  $b(s)$  are the numbers of ones and zeros in odd segments, respectively, according to  $s$ . From  $S_{mn}^k$ , we desire one segmentation  $t_{mn}^k$  such that  $a(t_{mn}^k) = \max\{a(s) \mid s \in S_{mn}^k\}$  and another segmentation  $u_{mn}^k$  such that  $a(u_{mn}^k) = \min\{a(s) \mid s \in S_{mn}^k\}$ . The degenerate case of a segmentation involving only the first position leads to the following initial conditions:

$$t_{1,1}^1 = u_{1,1}^1 = \phi. \tag{8}$$

Let  $o(n) = 1$  if  $n$  is odd and zero otherwise and let  $\wedge$  be the string concatenation operator. We may now write the following recurrence relationship:

$$t_{m,n}^k = \begin{cases} t_{m-1,n}^{k-o(n)} & \\ \text{if } a(t_{m-1,n-1}^{k-o(n)}) < a(t_{m-1,n}^{k-o(n)}) & \\ t_{m-1,n-1}^{k-o(n)} \wedge ([m-1]) & \\ \text{if } a(t_{m-1,n-1}^{k-o(n)}) \geq a(t_{m-1,n}^{k-o(n)}) & \end{cases} \tag{9}$$

If  $a(t_{m-1,n-1}^{k-o(n)}) = a(t_{m-1,n}^{k-o(n)})$ , more than one segmentation gives the same maximal number of ones in odd segments for given  $m$ ,  $n$ , and  $k$ ; the recurrence relationship above arbitrarily chooses the second clause in this case. The treatment of impossible conditions in which  $k < n/2$  or  $n > m$  is ignored here; the exclusion of such conditions from calculations is straightforward. An analogous equation specifies recurrence relations for  $u_{mn}^k$ .

Given these relationships, for a given sequence, we may determine  $t_{ln}^k$ ,  $u_{ln}^k$ ,  $a(t_{ln}^k)$ , and  $a(u_{ln}^k)$  for  $2 \leq n \leq N$  and  $n/2 \leq k \leq m - (n-1)/2$ .

For all of the measures of fit of a segmentation discussed above, it can easily be proved that the best segmentation into a given number of segments,  $\sigma_n^*$ , will be drawn from among the sets  $\{t_{ln}^k\}$  and  $\{u_{ln}^k\}$ ; i.e., we choose  $\sigma_n^*$  such that

$$\chi_n^2 = \chi^2(\sigma_n^*) = \max \left\{ \chi^2(s) \mid s \in \left\{ t_{ln}^k \right\} \cup \left\{ u_{ln}^k \right\} \right. \\ \left. \text{for } n/2 \leq k \leq m - (n-1)/2 \right\}. \tag{10}$$

This final evaluation will involve no more than  $2Nl$  evaluations of  $\chi^2$  (or other measure of fit). An implementation of the calculation of  $\{t_{ln}^k\}$  and  $\{u_{ln}^k\}$  into a simple dynamic programming scheme will require time and memory whose orders are quadratic in  $l$ , the sequence length, and linear in  $N$ , the maximum number of segments. As described elsewhere (Pearson and Miller, 1992), memory requirements can be made linear in  $l$  and in  $N$  at the expense of approximately doubling the run time.

## LINKED CITATIONS

- Page 1 of 1 -



You have printed the following article:

### **Multiple-Changepoint Testing for an Alternating Segments Model of a Binary Sequence**

Aaron L. Halpern

*Biometrics*, Vol. 56, No. 3. (Sep., 2000), pp. 903-908.

Stable URL:

<http://links.jstor.org/sici?sici=0006-341X%28200009%2956%3A3%3C903%3AMTFAAS%3E2.0.CO%3B2-J>

---

*This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.*

## References

### **Maximum Likelihood Estimation of Multiple Change Points**

Yun-Xin Fu; R. N. Curnow

*Biometrika*, Vol. 77, No. 3. (Sep., 1990), pp. 563-573.

Stable URL:

<http://links.jstor.org/sici?sici=0006-3444%28199009%2977%3A3%3C563%3AMLEOMC%3E2.0.CO%3B2-2>

### **Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination**

Peter J. Green

*Biometrika*, Vol. 82, No. 4. (Dec., 1995), pp. 711-732.

Stable URL:

<http://links.jstor.org/sici?sici=0006-3444%28199512%2982%3A4%3C711%3ARJMCMC%3E2.0.CO%3B2-F>

### **Minimally Selected $p$ and Other Tests for a Single Abrupt Changepoint in a Binary Sequence**

Aaron L. Halpern

*Biometrics*, Vol. 55, No. 4. (Dec., 1999), pp. 1044-1050.

Stable URL:

<http://links.jstor.org/sici?sici=0006-341X%28199912%2955%3A4%3C1044%3AMSPAOT%3E2.0.CO%3B2-B>