CHANGE TREES AND MUTAGRAMS FOR THE VISUALIZATION OF LOCAL CHANGES IN SEQUENCE DATA

Hans-Georg Müller and Newton Wai

Department of Statistics
University of California
One Shields Ave.
Davis, CA 95616
U.S.A.

May 2003

Abstract

The analysis of local changes in sequence data is of interest for various applications such as the segmentation of DNA and other genetic sequences, or financial data sequences. Patterns of change that can be characterized as local jump change or slope change are of special interest. We propose simple graphical tools to visualize such patterns of local change. The concept of mode trees, developed for the visualization of local patterns in densities (Minnotte and Scott, 1993), is adapted to visualize patterns of local change in dependency on a threshold parameter by means of a *Change Tree*. The simultaneous visualization of scale effects, in analogy to SiZer (Chaudhuri and Marron, 1999, 2000) motivates another graphical device, the *Mutagram*. We illustrate these concepts with several sets of sequence data.

Key Words: Biological sequence data, DNA sequence, jump discontinuity, nonparametric segmentation, scaling, slope change.

1. INTRODUCTION

The statistical analysis of abrupt changes in the mean or slope of a regression function has been widely studied. The location where such a sudden change occurs is referred to as a change-point. Change-points have been studied in linear models (Kim, 1994; Bhattacharya, 1994; Jandhyala and MacNeill, 1997), non-linear models (Jandhyala and Al-Saleh, 1997; Rukhin and Vajda, 1997), generalized linear models (Braun, Braun and Müller, 2000) and non-parametric models (Hall and Titterington, 1992; Müller, 1992; Hall, Gijbels, and Kneip, 1999; Darkhovski, 1994; Antoniadis, Gijbels and MacGibbon, 2000). For an overview on some earlier developments in theory and application, we refer to the monograph on Change-point Problems edited by Carlstein, Müller and Siegmund (1994).

In this paper we propose several tools for the graphical detection and display of change-points, based on local quasi-likelihood comparisons. The proposed methods are model free and graphical in nature. A key graph is the *Change Tree*. Change Trees are constructed by obtaining local quasi-likelihood fits with and without change-points at a set of potential locations. The locations are then ranked by evaluating the magnitude of the differences of quasi-deviances between fits with and without a change-point at each location. The resulting ranked change-point locations are then represented in a tree graph. The root sizes indicate the size of the observed difference in quasi-likelihood between the two fits. The longest roots of the tree correspond to locations with the largest change in deviance, corresponding to the difference in quasi-likelihood, the next longest roots indicate points with the next largest changes in deviance, and so on.

A second key graph is the *Mutagram* which supplements Change Trees. It displays the effects of an entire range of window sizes, referring to the window within which the local quasi-likelihoods are computed. The variation in window widths or scaling has an influence on the quasi-likelihood differences. By graphing quasi-likelihood deviance differences simultaneously across a multitude of scales or different bandwidths, the mutagram reveals the dependency of deviance differences on scale and helps to pinpoint scale-independent or "robust" change-point locations. Thus mutagrams help in sorting out deviance differences that occur across many scales, pointing to actual changes, from those that appear only for a few, often the smallest, bandwidths and are more likely to be caused by noise configurations. Mutagrams are constructed in such a way that evidence pointing to a jump change

is combined with evidence for a slope change. They are related in spirit to mode trees (Minnotte and Scott, 1993) and the mode forest (Minnotte, Marchette and Wegman, 1998). Mutagrams are also closely related to the SiZer idea of Chaudhuri and Marron (1999, 2000). While mutagrams are constructed for the analysis of change across various scales, the focus of SiZer is on the effect of scale on smoothing.

These graphical ideas will be developed and illustrated in the following with various sequence data.

2. ONE-SIDED QUASI-LIKELIHOOD

The notion of deviance that is associated with local quasi-likelihood fitting plays a central role for our deliberations. A version of local quasi-likelihood was proposed in Fan, Heckman and Wand (1995), extending the concept of local likelihood (Staniswalis, 1989). Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a sequence of independent random pairs, where $x \in \mathbf{R}^1$, $\mu(x) = E(Y|X = x) = g(\eta(x))$, $\eta = \eta(x) = \beta_0 + \sum_{i=1}^{n} x_i\beta_i$ is the linear predictor and g is a link function. Furthermore, conditional variances $\mathrm{Var}(Y|X = x) = \sigma^2 V(\mu(x))$ are determined by a positive function $V(\cdot)$, the variance function, where σ^2 plays the role of an additional dispersion parameter. The parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ associate η with the predictor variable x, while the link function g relates conditional mean responses μ and linear predictors η.

Quasi-likelihood, first proposed by Wedderburn (1974), is popular in statistical inference as it avoids full specification of a distribution family such as the exponential family that is required in conventional likelihood methods. For quasi-likelihood modelling, it is enough to choose a link function $g(\cdot)$ and a variance function $V(\cdot)$. More precisely, one assumes that the observations (x_i, y_i) of predictors $x_i \in \mathbf{R}^r$, $r \geq 1$ and of responses $y_i \in \mathbf{R}$ are related by a regression function $\mu(x) = E(Y|X = x)$ (McCullagh and Nelder, 1989). The single index assumption then corresponds to

$$\mu(x) = E(Y|X = x) = g(x\beta),$$

where $\beta = (\beta_0, \ldots, \beta_p)^T \in \mathbf{R}^{p+1}$ is a parameter vector and $x = (1, x^T)^T$ is the vector of predictor values, augmented by a 1. The conditional variance is assumed to depend solely on the conditional mean via $\mathrm{Var}(Y|X = x) = \sigma^2 V(\mu(x))$.

Specification of conditional mean and conditional variance lead to a quasi-likelihood for the obser-

vation y_i, defined as

$$\tilde{Q}(\mu; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V(t)} dt$$

(Wedderburn, 1974). Defining quasi-deviance $\tilde{D} = -2\sigma^2 \tilde{Q}(\mu; y_i)$, the sample deviance is then

$$\tilde{D}(\mu, y) = \sum_{i=1}^{n} \tilde{D}(\mu; y_i).$$

(Fan and Gijbels, 1996). Localizing quasi-likelihood using local polynomial fitting, we obtain the local quasi-likelihood estimates

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)(x) = \arg \max_{\beta \in \Re^{p+1}} \Sigma_{i=1}^{n} Q(\eta_i; y_i) w_i K(\frac{x_i - x}{h}),$$

where $\eta_i = \beta_0 + \beta_1(X_i - x) + \ldots + \beta_p(X_i - x)^p$ (Braun and Müller, 1998). This simply means that the quasi-likelihood is applied only in a small window around the target location x. Here K is a nonnegative kernel function and h is a bandwidth defining the size of the window, while the w_i denote case weights. Usually we will choose $w_i = 1$, $i = 1, \ldots, n$, but in some cases, for instance when the data are averages formed from unequal numbers of observations, case weights may be included. The order of the local polynomial p is usually chosen as $p = 0$ (locally constant) or $p = 1$ (locally linear). Then we set $\hat{\mu}(x) = g(\hat{\beta}_0(x))$, where both link function $g(\cdot)$ and variance function $V(\cdot)$ are assumed to be known. The identity link function, $g \equiv id$, often is a feasible choice for local quasi-likelihood.

We extend the concept of local quasi-likelihood to one-sided local quasi-likelihood, where only data on the left or on the right of the target location enter the quasi-likelihood. Related ideas were considered in Braun and Müller (1998) and Loader (1999); for more recent discussion of such approaches compare Gregoire and Hamrouni (2002) and Huh and Carriere (2002). In the following, we are using a kernel function K with compact support $[-1, 1]$ and fix a target argument x. The windows for the local quasi-likelihoods which we are considering are then $[x - h, x]$ for one-sided left quasi-likelihood and $[x, x + h]$ for one-sided right quasi-likelihood.

Formally, the one-sided quasi-likelihood estimates to the right and to the left are defined as

$$(\hat{\beta}_{0+}, \hat{\beta}_{1+}, \ldots, \hat{\beta}_{p+})(x) = \text{argmax}_{\beta \in \Re^{p+1}} \Sigma_{i=1}^{n} 1_{\{x_i > x\}} Q(\eta_i; y_i) w_i K(\frac{x_i - x}{h})$$

and

$$(\hat{\beta}_{0-}, \hat{\beta}_{1-}, \ldots, \hat{\beta}_{p-})(x) = \text{argmax}_{\beta \in \Re^{p+1}} \Sigma_{i=1}^{n} 1_{\{x_i \leq x\}} Q(\eta_i; y_i) w_i K(\frac{x_i - x}{h}).$$

We then define the maximizers for local and one-sided local quasi-likelihoods,

$$\hat{Q}(x,y;h) = \sum_{i=1}^{n} Q(\hat{\beta}_0(x) + \hat{\beta}_1(x)(x_i - x) + \ldots + \hat{\beta}_p(x)(x_i - x)^p; y_i) w_i K(\frac{x_i - x}{h}),$$

$$\hat{Q}_+(x,y;h) = \sum_{i=1}^{n} 1_{\{x_i > x\}} Q(\hat{\beta}_{0+}(x) + \hat{\beta}_{1+}(x)(x_i - x) + \ldots + \hat{\beta}_{p+}(x)(x_i - x)^p; y_i) w_i K(\frac{x_i - x}{h}),$$

and

$$\hat{Q}_-(x,y;h) = \sum_{i=1}^{n} 1_{\{x_i < x\}} Q(\hat{\beta}_{0-} + \hat{\beta}_{1-}(x)(x_i - x) + \ldots + \hat{\beta}_{p-}(x)(x_i - x)^p; y_i) w_i K(\frac{x_i - x}{h}),$$

along with the corresponding fitted values, $\hat{\mu}(x,h) = g(\hat{\beta}_0(x))$, $\hat{\mu}_\pm(x,h) = g(\hat{\beta}_{0\pm}(x))$ and $\hat{\mu}_i = \hat{\mu}(x_i,h)$, $\hat{\mu}_{i\pm} = \hat{\mu}_\pm(x_i,h)$.

The corresponding quasi-deviances

$$D_\pm(x,y;h) = -2\sigma^2 \hat{Q}_\pm(x,y;h),\ D(x,y;h) = -2\sigma^2 \hat{Q}(x,y;h)$$

measure the deviance of the fitted model. Note that quasi-deviance does not depend on σ^2, as the factor $\frac{1}{\sigma^2}$ appearing in the quasi-likelihood is cancelled.

For the three most common types of data the quasi-deviances are as follows (see, e.g., McCullagh and Nelder, 1989): The quasi-normal deviance is $D(x,y;h) = \sum (y_i - \hat{\mu}_i)^2$, the quasi-Poisson deviance $D(x,y;h) = \sum -2(y_i \log \hat{\mu}_i - \hat{\mu}_i)$ and the quasi-binomial deviance $D(x,y;h) = \sum (y_i \log(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}) + \log(1 - \hat{\mu}_i))$.

3. LOCAL CHANGE-POINT STATISTICS

Two basic change-point models are useful to describe rapid changes that are observed in data. On one hand, a change can be so abrupt that it is best modelled by assuming a discontinuity. On the other hand, in many cases one observes a rapid change in a trend that is best modelled as a slope change or discontinuity in the first derivative. We note that in our applications the identity link generally performed well, and we use this link in our examples. Other link functions could be equally chosen.

Local Jump Model: The polynomial that is fitted locally is a constant with

$$D(x, y; h) = \sum_{i=1}^{n} D(\hat{\beta}_0; y_i) w_i K(\frac{x_i - x}{h})$$

and $\hat{\mu} = \bar{y}$. Fitting constants in the one-sided windows $[x - h, x]$ and $[x, x + h]$ provides a local approximation to a discontinuity at x, and the resulting fit is compared with fitting a constant over the entire window, i.e., the local fit without change-point at x.

Local Slope Change Model: A slope change is approximated by a function corresponding to a two-stick regression line, with a change-point located at the midpoint x, given by $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - x) x_{i2}$, where $x_{i1} = x_i$ and $x_{i2} = 1_{\{x_i > x\}}$. The quasi-likelihood in this case equals

$$D^*(x, y; h) = \sum_{i=1}^{n} D(\beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - x) x_{i2}; y_i) w_i K(\frac{x_i - x}{h}).$$

This model reflects a slope change occurring at x, while the function itself is continuous. This model fit is compared with a fit that is obtained for an unbroken line fitted to the data over the entire window, i.e., the local linear fit without change-point.

We fit these models with local and local one-sided quasi-likelihood as described above. A comparative illustration of jump and slope change models is provided in Fig. 1, using the Dow Jones Industrial Average data and the Bacteriophage λ data ($n = 485$) series of binned DNA base frequencies from Skalka et al. (1968). The Dow Jones Industrial Average data is a record of the log-transformed weekly closings of the Dow Jones Industrial Average from 1970 to 1999, and for this analysis we use the quasi-normal deviance. The Bacteriophage λ data correspond to Guanine and Cytosine (G+C) proportions which have been aggregated into bins from the DNA sequence of this phage, and accordingly, the quasi-binomial deviance is used for these data. Fig. 1 demonstrates that while for the Bacteriophage λ data the jump model is seemingly more appropriate, the slope change model appears to provide a better fit for the Dow Jones data, especially when choosing large bandwidths.

Using locally weighted maximum quasi-likelihood estimation, we obtain deviances $D_+(x, y; h)$, $D_-(x, y; h)$ and $D(x, y; h)$ of locally one-sided and local model fits. We then define the evidence function $\Delta(x, h)$ for the local jump model as

$$\Delta(x, h) = |(D_+(x, y; h) + D_-(x, y; h)) - D(x, y; h)|.$$

For the local slope change model, we define analogously

$$\Delta^*(x, h) = |D^*(x, y; h) - D_*(x, y; h)|,$$

where D^* is defined as above and is the deviance for the model with slope change, fitted to the data in the entire window, while D_* is the deviance for an unbroken line without slope change fitted to the data in the entire window,

$$D_*(x, y; h) = \sum_{i=1}^{n} D(\beta_0 + \beta_1(x_i - x); y_i) w_i K(\frac{x_i - x}{h}).$$

Note that functions Δ, Δ^* depend on a bandwidth h and quantify the degree of evidence that separate left- and right-sided local models provide a better fit than an overall two-sided local fit. The improvement is measured as the amount of decline in the deviance that one observes for the more flexible one-sided fits. In the case of fitting constants (Local Jump Model), we compare the deviances between a model that allows for separately fitted constants in left and right half-windows with a model that fits one global constant to the data in the entire window by maximum quasi-likelihood.

We note that local statistical evidence can be provided for the local changes in deviance by selecting a level α, then testing at level α for a change at each location x for a grid of locations. Since the models are nested within each other, testing for no change at x in the local jump model corresponds to testing $H_0 : \beta_{0+} = \beta_{0-}$, and in the local slope change model to testing $H_0 : \beta_2 = 0$. We assume that the local approximation to the change that actually occurs which is provided by a local jump model or local slope change model is sufficiently accurate under suitable regularity conditions (compare Müller and Song, 1997, and Gijbels et al., 1999, for a detailed discussion of examples for such conditions). Then an approximate asymptotic p-value and inference at level α can be obtained for each location x by referring to the asymptotic χ^2_1-distribution of the test statistic under the null hypothesis H_0 as shown above.

The relevant test statistic is the quasi-likelihood ratio which corresponds to the deviance difference between the models with and without change-point at x and is given by $\Delta(x, h)$ for the jump change model and $\Delta^*(x, h)$ for the slope change model. The asymptotic χ^2_1-distribution of these test statistics under the null hypothesis follows from basic properties of quasi-likelihood (Wedderburn, 1974). Simultaneous inference procedures are needed if one were to formally test for the presence of change-points with unknown location x. While such a formal test is not within the scope of our proposed procedure, we demonstrate below that in some cases the construction of mutagrams for very small levels α provides evidence for overall significance of observed changes, simply by applying a Bonferroni correction.

4. CHANGE TREES

A change tree is a graph that provides a visual summary of the location of possible change-points in a data sequence. It is related in spirit to the mode tree (Minnotte and Scott, 1993; Marchette and Wegman, 1997). For a fixed point x, the evidence that a change-point is located at x is based on the evidence functions $\Delta(x,h), \Delta^*(x,h)$ defined above. These evidence functions depend on a bandwidth, bandwidth, link function, variance function and, last not least, linear predictor function - in our case local constants, local two-stick regressions or local linear regressions.

In order to construct change trees for given bandwidth h, we obtain the functions Δ, Δ^* on a grid of n points in the domain $[0,T]$ of the predictor variable, $0 < x_1 < x_2 < \ldots < x_n < T$, which correspond to the support points of the sequence where measurements are available, for example gene location in the DNA sequence examples. We then find the concomitant locations of the ordered maxima of functions Δ, Δ^* over this finite grid of points. For a small $h_0 \ll h$, define $S_0 = [0,T]\backslash\{[0,h_0) \cup (T-h_0,T]\}$, where $\tau_1 = \text{argmax}_{x \in S_0} \Delta(x,h)$. Then define $S_{j+1} = S_j \backslash (\tau_j - h_0, \tau_j + h_0)$, where $\tau_j = \text{argmax}_{x \in S_{j-1}} \Delta(x,h)$ for $j = 1,2,\ldots,n$. Depending on the choice of h_0, reflecting an assumed minimum distance between two change-points, the sequence of maxima will include a number of $M \leq n$ such maxima after which no further maxima can be found.

Let $(\tau_1, \ldots, \tau_M)$ be the M locations identified as ordered maxima locations and thus possible jump locations. The ordered evidence values $\Delta(\tau_1,h) \geqslant \ldots \geqslant \Delta(\tau_M,h)$ then form the basis for the construction of change trees. Here and in the following, we describe this development for jump changes and the function Δ, keeping in mind that the development for slope changes is completely analogous, replacing Δ by Δ^*.

Once $(\tau_j, \Delta(\tau_j,h))$, $j = 1,2,\ldots,M$, have been determined, we construct vertical line segments defining the change tree. Vertical line segments are drawn from the location points τ_j, $j = 1,\ldots,M$ to the associated values of observed change, $\Delta(\tau_j,h)$. Horizontal segments join a vertical segment to the nearest vertical segment with a higher evidence value for change. That is, we join the vertical segment of τ_j with that of τ_{j+1} and τ_{j+2} for $j = 1$ to $M-2$. Thereby we construct a change tree in analogy to the construction of the mode tree of Minnotte and Scott (1993).

The values of Δ can also be thresholded in order to "vertically" prune the tree, in the sense that a

location where $\Delta(\tau_j, h) < \Delta_0$ for a threshold Δ_0 is discarded from further consideration. As a rule of thumb, we choose $\Delta_0 = 0.1\max_{1\le j\le M}\Delta(\tau_j, h)$, and h_0 (see definition of S_0 at the end of the previous chapter), which governs "horizontal pruning", was chosen as $h_0 = 0.04n$. For change trees in the jump model, locations where $\hat{\beta}_{0+} > \hat{\beta}_{0-}$, i.e., where changes from lower to larger level are observed, are indicated by a solid red vertical line, whereas locations where $\hat{\beta}_{0+} < \hat{\beta}_{0-}$ are denoted by a dotted red vertical line. In change trees for the slope change model, locations where the slope increases are indicated with a solid blue line, and locations where it decreases, with a dashed blue vertical line.

As an example, consider a biological sequence, the Bacteriophage λ sequence (Skalka et al. 1968; $n = 485$) that is shown in Fig. 1. The data correspond to Guanine and Cytosine (G+C) proportions which have been aggregated from the DNA sequence of this bacteriophage. Braun and Müller (1998) used split local polynomial fits for the Bacteriophage λ data and provided evidence for a change-point at 22.6Kbp and 33.2Kbp (kilo base pairs). A global step function fit with quasi-likelihood was described in Braun, Braun and Müller (2000) and applied to these data, confirming the same change-point locations. Global segmentation methods are numerically tedious and local methods such as those considered here provide far more flexibility and faster computation.

The jump and slope change trees for these data (Fig. 2) were constructed using Binomial quasi-likelihood and $h_0 = 1.9$Kbp. The changes that are highlighted in these trees are in agreement with earlier findings, and the two change trees viewed together provide a useful visual summary of these changes. Note the remarkable agreement in this example between Local Jump and Local Slope Change model.

5. MUTAGRAMS

Change trees are scale-dependent and are affected by the specific bandwidth choice that is made. By viewing one tree one cannot fully appreciate the effect that different bandwidths and scaling might have on sequence segmentation. A graphical approach to address the scaling issue is to display a whole range of deviance changes simultaneously for different bandwidths, for each support point x. This leads to a graphical device that we call *Mutagram* (inspired by the Latin mutare, to change).

The idea of the mutagram is related to the mode forest (Minnotte, Marchette and Wegman, 1998). It is essentially a graphical representation of a 2-dimensional matrix. For given coordinate points x

and h, define $M_{xh} = \Delta(x, h)$ (analogously for Δ^*), with evidence functions Δ, Δ^* as defined in Section 3. The matrix entries are based on a grid of log-equidistant bandwidths $\log(h_i)$ (in descending order), $i = 1, \ldots, m$, where we choose $m = 20$, and locations x_j, $j = 1, \ldots, n$. The functions $\Delta(x, h)$ and $\Delta^*(x, h)$ for both jump and slope change model are calculated and the significance of a jump or slope change at x is determined according to a pre-specified level α (see section 3).

Areas in the graph that are associated with significant upwards change are colored red, and areas associated with a significant decrease are colored blue. The color is chosen as more intense in areas where both jump and slope change model indicate a significant change in the same direction, and darker or lighter where only one of the models is significant. Significance is with respect to a pre-specified level α. Areas that have no significant changes for the given level α or for which jump and slope changes point to different directions of change are indicated with shades of purple. Details regarding the color scheme that is used for displaying the mutagrams can be found in the legend of Fig. 3.

This figure displays examples of mutagrams for the Bacteriophage λ data for two different levels α. The mutagram for $\alpha = 0.05$ picks up considerably more changes at small scales as compared to the mutagram for the smaller level $\alpha = 0.001$. This is typical and indicates that change-points detected at small bandwidths and at the usual levels for α are not very reliable. It is advisable to choose a relatively small level for α that may allow for simultaneous inference, if necessary even with the conservative Bonferroni criterion. Change-points that are present at all levels of scaling and at small levels α clearly possess more relevance.

We note that the same major change occurring at 22Kbp is identified in both mutagrams in Fig. 3, and this location has been identified as a major change-point in previous analyses of this sequence. Additional changes occur to the right of this major change-point. The presence of such additional change-points is consistent with the earlier analysis of these data of Braun et al. (2000).

Mutagrams enable the simultaneous study of the effects of location and bandwidth or scale. For a in-depth discussion of scale effects and their significance, we refer to SiZer (Chaudhuri and Marron, 1999, 2000). Mutagrams highlight change-points, narrowly pin-pointing them at small scales and broadening their footprint at larger scales for major change-points. This is due to the fact that a major change-point will exercise its impact more at large scales as compared to small scales, leading to increased visibility in the mutagrams in the areas of larger bandwidths. At small scales, random

noise configurations can cause the corresponding change configurations to fluctuate wildly (similar small scale fluctuations are observed in SiZer). Large scale significance is thus useful to corroborate the existence, significance and impact of change-points on the sequence, while analysis at smaller scales is better suited to pinpoint locations. Mutagrams therefore are a useful complement to change trees and a valuable tool to assess the nature of changes that occur in sequence data simultaneously over many scales.

As another illustration, we consider the S. Cerevisiae III ($n = 526$) sequence of brewer's yeast gene III. The data correspond to binned relative frequencies of the Guanine and Cytosine (G+C) proportion in the DNA sequence (Oliver et al., 1992). The data set is available at the Genbank database (http://www.ncbi.nlm.nih.gov/Genbank/). DNA segmentation for these data has been studied by Braun et al. (2000), Chechetkin and Lobzin (1998) and Liö et al. (1996). The Binomial quasi-likelihood is used in our analysis, as the G+C proportion is a binary outcome. The change trees for both local jump model as well as local slope change model and when using different bandwidths are remarkably similar (Fig. 4), indicating a satisfactory level of stability. The major change occurs at around 344 Kbp where the G+P proportion rapidly rises above previous levels.

The corresponding mutagram constructed for a small level α of $\alpha = 10^{-8}$ in Fig. 5 pinpoints essentially the same locations as the two change trees, and also identifies the change at 344Kbp as a major change, attaching significance to it. The number of individual locations (due to the aggregation into 526 bins) and scales (20 bandwidth levels) combined is small enough so that local significance as indicated by the mutagram translates into global significance at the usual 5% level, when applying the conservative Bonferroni correction.

6. CONCLUSIONS

Change trees and mutagrams are simple graphical tools that are useful to enhance the detection, description and visualization of changes in sequence data. Such data originate from diverse fields. As the local likelihoods are determined using quasi-deviance, these methods easily adapt to data with non-normal distributions and only require specification of the variance-mean relationship. Change trees are basic graphical representations of local evidence for changes in a given sequence. Both change trees and mutagrams are flexible graphical devices that can easily be adapted to various local change-point

statistics that depend on location and scale, and their construction is not limited to the specific local change-point statistics that we propose in Section 3.

In our illustrations, mutagrams enable the user to gauge the stability of a "change feature" across different scales, taking into account a whole range of scales and window widths simultaneously. Local and to some extent also global inference for the presence of a change-point at a fixed location is a consequence. The two local change models that are simultaneously included in this graph, the local jump model and the local slope change model, complement each other. Their combined evidence that results in the mutagram is especially useful. Mutagrams are close in spirit to the philosophy of SiZer and also show similar characteristics. They combine information on jump changes, slope changes and inference, all in dependence on location and scale. For a single graphical device, they convey a wealth of complex information in an informative manner.

Questions that might be addressed in future investigations are specific asymptotic considerations that might allow for improved simultaneous inference. Theoretical investigations of the behavior of the change statistics over the entire range of scale factors would also be of interest, for example through the investigation of the properties of a stochastic process that is defined as a random function of the scale parameter. Finally, the extension to models which allow the explicit incorporation of dependency structure in observed sequences is also a topic for further research.

ACKNOWLEDGMENTS

We wish to thank two referees and an Associate Editor for helpful comments. The comments of one referee especially led to signficant improvements in the construction of mutagrams and change trees. This research was supported in part by NSF grants DMS-99-71602 and DMS-02-04869.

REFERENCES

Antoniadis, A., Gijbels, I. and Macgibbon, B. (2000). Non-parametric estimation for the location of a change-point in an otherwise smooth hazard function under random censoring. *Scandinavian Journal of Statistics*, **27**, 501–519.

Bhattacharya, P.K. (1994). Some aspects of change-point analysis. In: Carlstein, E., Müller, H.-G. and Siegmund, D. (eds). *Change-point Problems.* Institute of Mathematical Statistics, Hayward, California, Lecture Notes and Monograph Series, **23**, 28–56.

Braun, J.V., Braun, R.K. and Müller, H.G. (2000). Multiple change-point fitting via quasi-likelihood, with application to DNA sequence segmentation. *Biometrika* **87**, 301–314.

Braun, J.V. and Müller, H.G. (1998). Statistical methods for DNA segmentation. *Statistical Science* **13**, 142–162.

Carlstein, E., Müller, H.-G. and Siegmund, D. (eds.) (1994). *Change-point Problems.* Institute of Mathematical Statistics, Hayward, California, Lecture Notes and Monograph Series, Vol 28.

Chaudhuri, P and Marron, J.S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association* **94**, 807–823.

Chaudhuri, P and Marron, J.S. (2000). Scale space view of curve estimation. *Annals of Statistics* **28**, 408–428.

Chechetkin, V.R. and Lobzin, V.V. (1998). Study of correlations in segmented DNA sequences: Applications to structural coupling between exons and introns. *Journal of Theoretical Biology* **190**, 69–83.

Darkhovski, B.S. (1994). Nonparametric methods in change-point problems: a general approach and some concrete algorithms. In: Carlstein, E., Müller, H.-G. and Siegmund, D. (eds). *Change-point Problems.* Institute of Mathematical Statistics, Hayward, California, Lecture Notes and Monograph Series, **23**, 99–107.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall.

Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**, 141–150.

Gijbels, I., Hall, P. and Kneip, A. (1999). On the estimation of jump points in smooth curves. *Annals of the Institute of Statistical Mathematics* **51**, 231–251.

Gregoire, G. and Hamrouni, Z. (2002). Change point estimation by local linear smoothing. *Journal of Multivariate Analysis* **83**, 56–83.

Huh, J. and Carriere, K.C. (2002). Estimation of regression functions with a discontinuity in a derivative with local polynomial fits. *Statistics and Probability Letters* **56**, 329–343

Hall, P., Titterington, D.M. (1992). Edge-preserving and peak-preserving smoothing. *Technometrics* **34**, 429–440.

Jandhyala, V.K. and Al-Saleh, J.A. (1999). Parameter changes at unknown times in non-linear regression. *Environmetrics* **10**, 711-724

Jandhyala, V.K. and MacNeill, I.B. (1997). Iterated partial sum sequences of regression residuals and tests for change-points with continuity constraints. *Journal of the Royal Statistical Society Series B*, **59**, 147–156

Kim, H. (1994). Tests for a change-point in linear regression. In: Carlstein, E., Müller, H.-G. and Siegmund, D. (eds). *Change-point Problems.* Institute of Mathematical Statistics, Hayward, California, Lecture Notes and Monograph Series, **23**, 170–176.

Liö, P., Politi, A., Ruffo, S. and Buiatti, M. (1996). Analysis of genomic patchiness of haemophilus influenzae and Saccharomyces chromosomes. *Journal of Theoretical Biology* **183**, 455–469.

Loader, C. (1999). *Local Regression and Likelihood.* Springer.

Marchette, D. and Wegman, E. (1997). The filtered Mode Tree. *Journal of Computational and Graphical Statistics* **6**, 143–159.

McCullagh, P. and Nelder, A.J. (1989). *Generalized Linear Models.* Chapman and Hall.

Minnotte, M.C. and Scott, D.W.(1993). The Mode Tree: a tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics* **2**, 51–68.

Minnotte, M.C., Marchette, D. and Wegman, E. (1998). The bumpy road to the mode forest. *Journal of Computational and Graphical Statistics* **7**, 239–251.

Müller, H.G. (1992). Change-points in nonparametric regression analysis. *Annals of Statistics*, **20**, 737–761.

Müller, H.G. and Song, K.S. (1997). Two-stage change-point estimators in smooth regression models. *Statistics and Probability Letters* **34**, 323–335.

Oliver, S.G., van der Aart, Q.J.M., Agostoni-Carbone, M.L., Aigle, M., Alberghina L., Alexandraki D., Antoine G., Anwar R., Ballesta, J.P.G., and Benit P. (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46.

Rukhin, A.L. and Vajda, I. (1997). Change-point estimation as a nonlinear regression problem. *Statistics* **30**, 181–200.

Skalka, A., Burgi, E. and Hershey, A.D. (1968). Segmental distribution of nucleotides in the DNA of bacteriophage lambda. *Journal of Molecular Biology* **34**, 1–16.

Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* **84**, 276–283.

Wedderburn, R.M.W. (1974). Quasi-Likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–447.

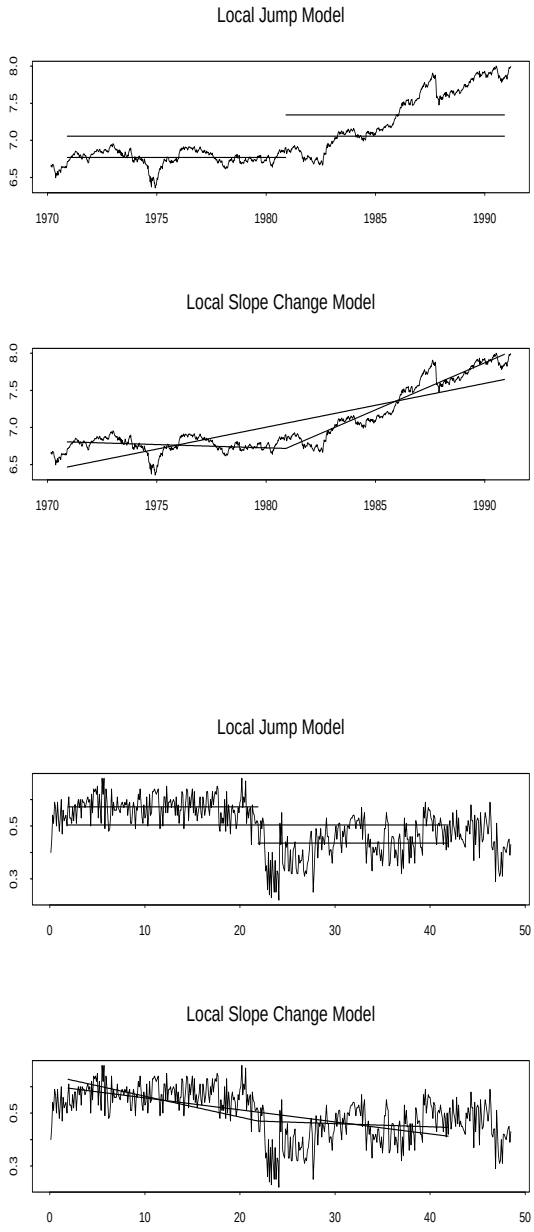

Fig. 1. Displaying fits of the *Local Jump Model* and the *Local Slope Change Model* for the Dow Jones Industrial data, centered at the week ending on Nov., 29, 1980, with bandwidth 10 years (upper panel) and the Bacteriophage λ data, centered at 27.9 kilo base pairs (Kbp), with bandwidth 20Kbp (lower panel).

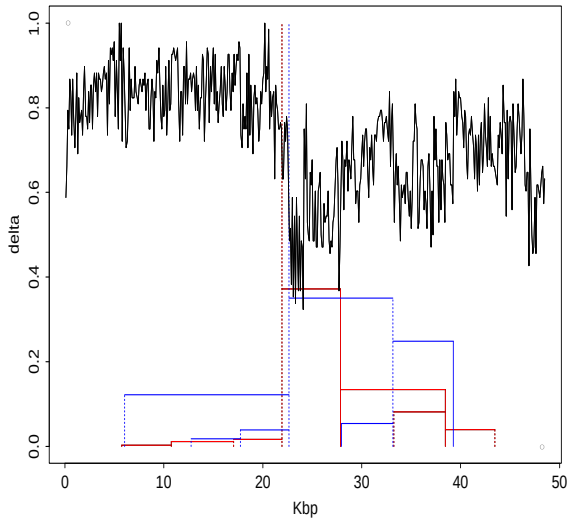

Fig 2. Change Trees for the Bacteriophage lambda sequence, with bandwidth 5 Kbp, for local jump model (red) and local slope change model (blue). Changes from lower to higher levels are indicated by solid lines, and changes to lower levels by dashed lines.

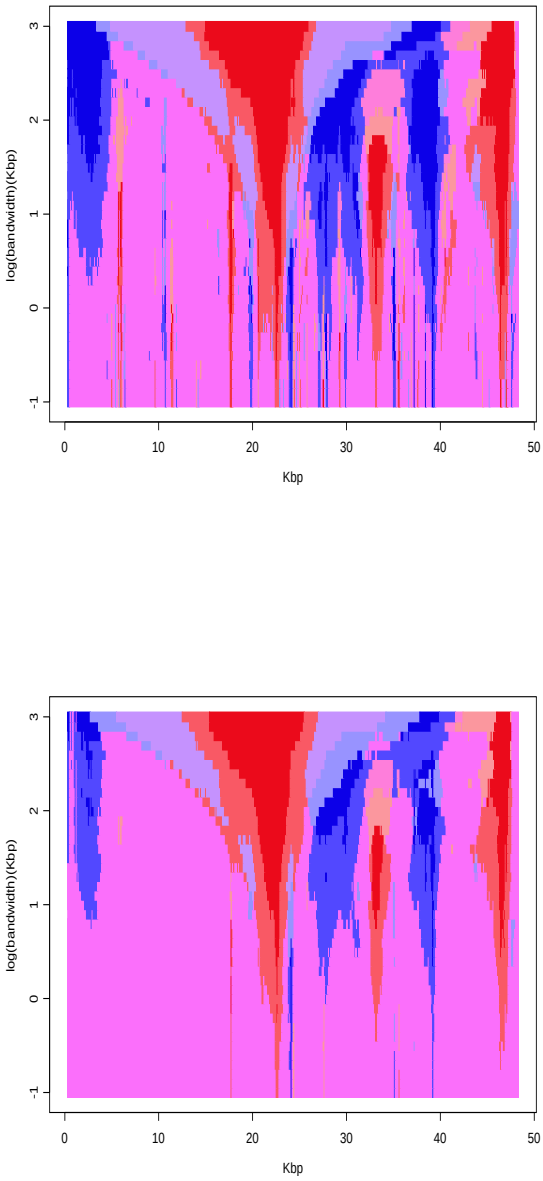

Fig 3. Mutagrams for the Bacteriophage λ sequence for levels $\alpha = 0.05$ (upper panel) and $\alpha = 0.001$ (lower panel). For these levels, areas in the graph associated with significant upward changes are colored blue, those associated with downward changes are colored red, and areas with insignificant or diverging changes under the two models are colored purple. The detailed color code is as follows, where changes up or down refer to changes that are significant under the given level α: Intensive blue – up in both jump and slope; Dark blue – up in jump only; Light blue – up in slope only; Intensive

red – down in both jump and slope; Dark red – down in jump only; Light red – down in slope only; Purple – not significant; Dark purple – up in jump and down in slope; Light purple – down in jump and up in slope.

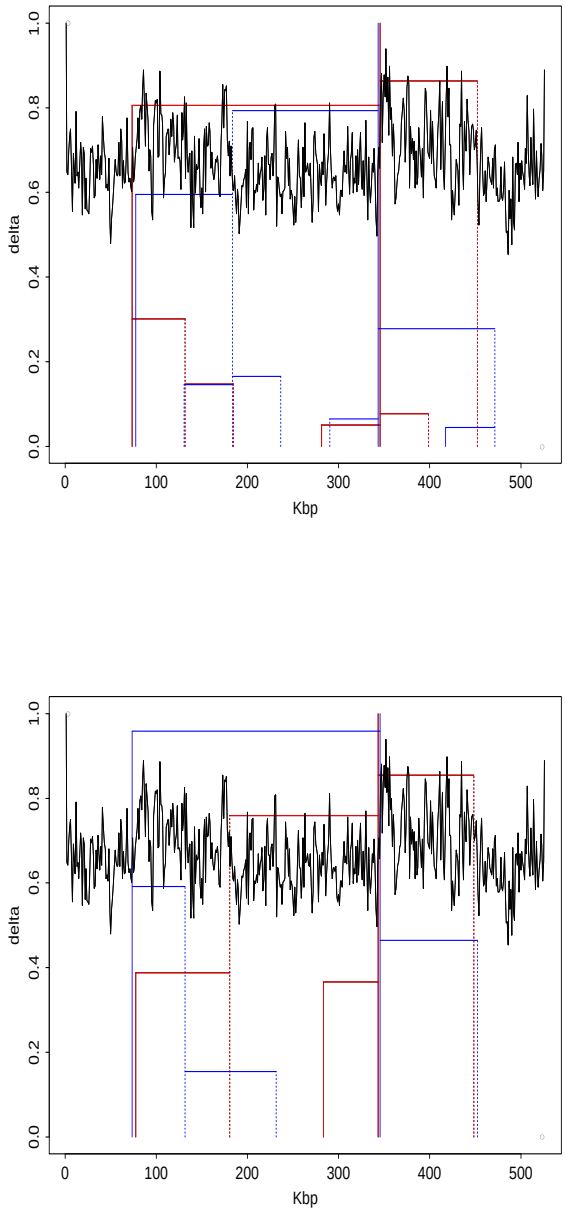

Fig 4. Change Trees for the S. Cerevisiae III sequence, using bandwidths 50 Kbp (upper panel) and 100 Kbp (lower panel).

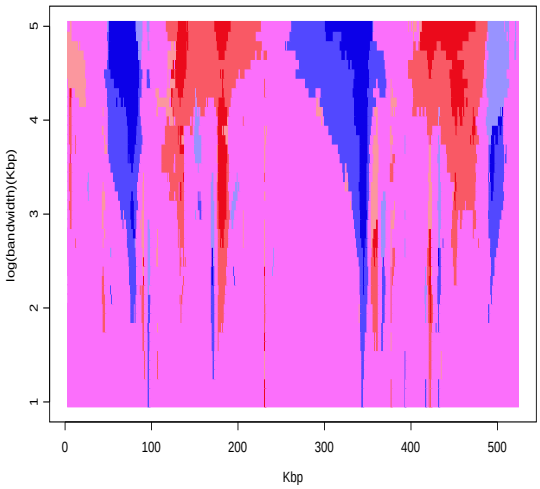

Fig 5. Mutagram for the S. Cerevisiae III sequence, at level $\alpha = 10^{-8}$.