University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 32

Nonparametric methods for analyzing replication origins in genomewide data

Debashis Ghosh*

*University of Michigan, ghoshd@psu.edu

Nonparametric methods for analyzing replication origins in genomewide data

Debashis Ghosh

Abstract

Due to the advent of high-throughput genomic technology, it has become possible to globally monitor cellular activities on a genomewide basis. With these new methods, scientists can begin to address important biological questions. One such question involves the identification of replication origins, which are regions in chromosomes where DNA replication is initiated. In addition, one hypothesis regarding replication origins is that their locations are non-random throughout the genome. In this article, we develop methods for identification of and cluster inference regarding replication origins involving genomewide expression data. We compare several nonparametric regression methods for the identification of replication origin locations. Testing the hypothesis of randomness of these locations is done using Kolmogorov-Smirnov and scan statistics. The methods are applied to data from a recent study in yeast in which candidate replication origins were profiled using cDNA microarrays.

Nonparametric methods for analyzing replication origins in genomewide data

Debashis Ghosh

Departments of Biostatistics

University of Michigan

Ann Arbor, MI, 48109-2029, USA

ghoshd@umich.edu

Corresponding author:

Debashis Ghosh, Ph.D.

Department of Biostatistics

School of Public Health, University of Michigan

1420 Washington Heights, Room M4057

Ann Arbor, Michigan 48109-2029

Phone: (734) 615-9824

Fax: (734) 763-2215

Email: ghoshd@umich.edu

Abstract

Due to the advent of high-throughput genomic technology, it has become possible to monitor cellular activities on a genomewide basis. With these new methods, scientists can begin to address important biological questions. One such question involves the identification of replication origins, which are regions in the chromosomes where DNA replication is initiated. One hypothesis is that their locations are nonrandom throughout the genome. In this article, we develop methods for identification of and cluster inference regarding replication origins involving genomewide expression data. We compare several nonparametric regression methods for the identification of replication origin locations. Testing the hypothesis of randomness of these locations is done using Kolmogorov-Smirnov and scan statistics. The methods are applied to data from a recent yeast study in which candidate replication origins were profiled using cDNA microarrays.

Keywords: Changepoint, Density Estimation, Derivative Estimation, Gene Expression, Kernel Smoothing, Microarray.

1 Introduction

With the explosion of high-throughput genomic data, scientists are now in the position of having the genetic information available for addressing important biological questions. The types of genomic data that are available range from sequences of complete organisms (Venter et al., 2001) to gene expression profiles from microarray experiments (Tusher et al., 2001) to protein-protein interaction maps (Uetz et al., 2000).

One question involves the existence and location of replication origins. The biology underlying this problem is further detailed in Section 2. A replication origin is the site on the genome where cell replication is initiated; identification of these locations is of great importance to understanding DNA replication. Recently, two global-wide studies attempting to identify replication origins in yeast were reported (Raghuraman et al., 2001; Wyrick et al., 2001). In this paper, we focus on the study of Raghuraman et al. (2001). A major statistical goal is to identify the chromosomal locations of peaks in the expression profiles. An example of such a profile is given in Figure 1.

The statistical analysis of replication origins has been previously considered by Truong et al. (2002), but they were not dealing with the situation of analyzing genomewide data. In addition, they had experimental replicates. In most high-throughput studies, replicates are not available. In addition, while Truong et al. (2002) were interested in finding one replication origin, we are now performing a global search for finding multiple sites of replications. It is generally accepted that in the yeast genome, there are approximately 200-400 sites of replication origins. Statistically, the problem addressed here is that of finding local modes. While there exists a literature on such procedures (Silverman, 1981; Müller and Sawitzki, 1991; Cheng and Hall, 1998; Fisher and Marron, 2001), they tend to deal with the issue of number of modes in contrast to identification of local modes. In addition, most of these methods will not be computationally feasible for finding modes because they would require nonparametric smoothing for multiple values of the smoothing parameter.

In this article, we use nonparametric regression methods to infer the locations of

replication origins and nonparametric clustering techniques to test the hypothesis of clustering of replication origins. Section 2 provides more details on the biology of replication origins and describe the experiment by Raghuraman et al. (2001). A statistical model for the analysis of the expression profiles and methods for identification of replication origins are given in Section 3. This section also describes nonparametric methods for assessing clustering. The proposed methodology is applied to the yeast data of Raghuraman et al. (2001) in Section 4. Finally, we conclude with some discussion in Section 5.

2 Biological Background

We now provide a brief review of DNA replication and origins of replications; more comprehensive discussions can be found in Gilbert (2001), Newlon and Theis (2002), and Bell and Dutta (2002).

Complete and accurate DNA replication is integral to the maintenance of the genetic integrity of all organisms (Bell and Dutta, 2002). In eukaryotic cells, replication begins at chromosomal elements called replication origins. At these locations, multiprotein complexes are assembled that eventually become two bidirectional replication forks. We will focus our discussion here on yeast, as this was the organism studied by Raghuraman et al. (2001). Potential locations for replication origins in yeast contain elements known as autonomously replicating sequences (ARS). These sequences are 100-200 base pairs (bp) in length and contain one or more copies of an 11-bp ARS consensus sequence (ACS) (Newlon and Theis, 2002). The protein that initiates replication is called the origin recognition complex (ORC). The ORC binds *in vivo* to multiple ARS throughout the cell cycle. Then a prereplication complex (pre-RC) forms that regulates replication. The pre-RC formation occurs during the G1 phase of the cell cycle. After it is formed, it then awaits activation by kinases that trigger the replication process. Each of these steps involves an ordered series of assembly of various transcription factors and other proteins.

Numerous biological experiments have studied various properties of replication ori-

gins. However, until recently, no experiments have studied these properties on a genomewide basis. This is due to the fact that the plasmid assay typically used to identify ARS elements is highly labor-intensive (Stinchcomb et al., 1979). In a recent study by Raghuraman et al. (2001), oligonucleotide microarrays were used to identify potential origins of replications. Density transfer experiments were used to measure replication times for each of the probes. Yeast cells were grown for many generations in medium containing two dense isotopes (^{15}N and ^{13}C); they were blocked at the G1-S phase boundary and then released into medium containing the isotopes ^{14}N and ^{12}C. At times $t = 0, 10, 14, 19, 25, 33, 44$ and 60 minutes in the S phase, culture samples were collected. The replicated DNA containing one heavy and one light (HL) strand (for the parent and daughter strands) were separated from the unreplicated DNA (which contained two heavy (HH) strands) was separated for each time point using density gradient centrifugation. Each set of DNA was then separately hybridized to an oligonucleotide microarray, which yielded an intensity measure. On the microarray, each probeset corresponded to a different genomic location in the yeast genome. The genomic locations used were evenly spaced for each chromosome. The intensity measure is then representative of the fraction of each sequence that had replicated at each time point.

The relationship of HL/HH strands as a function of chromosome position is illustrated in Figure 1. Early replicating sequences have higher HL fractions at earlier time points, while later replicating sequences have lower HL fractions. By considering the fraction of HL over all times points to the fraction of both HL and HH across all time points, we have a proxy measure for the time of replication. The microarray data in this yield a value for the HL percentage.

The authors applied a Fourier convolution smoothing algorithm to the microarray data; these are the data that we will consider in this paper. The measurements can then be plotted as a function of chromosome location; we present an example of the data in Figure 2. Based on these data, the goal is to find the local peaks and valleys in the data. Peaks represent replication origins, while valleys represent regions of replication

termination. Here and in the sequel, we will focus only on replication origins.

The authors calculated peaks and valleys using successive differences and then defined robust origins of replications as those origins that survive nine rounds of smoothing. The choice of nine seems relatively ad hoc; our goal is to develop a more formal statistical method for identifying replication origins. In addition, we wish to test the hypothesis of Gilbert (2001) that the location of origins of replication is nonrandom.

3 Experimental Methods

3.1 Data and Model

We observe the data $\{Y_{ij}\}$, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, where i indexes the chromosome, j indexes the location on the ith chromosome, and Y_{ij} is the corresponding expression measurement. We then formulate the following model for Y_{ij} as a function of chromosome location:

$$Y_{ij} = \mu_i(j/n_i) + \epsilon_{ij}, \tag{1}$$

where $\mu_i(t)$ is the mean function for the ith chromosome and ϵ_{ij} is a noise term. We assume that the error terms in (1) are a random sample from a normal distribution with mean zero and variance σ_i^2, $i = 1, \ldots, I$. We will be treating each chromosome separately, so we will suppress dependence on of μ_i and σ^2 on i in the sequel. In addition, we will assume that $n_i = n$. Because of the experimental design of the study by Raghuraman et al. (2001), the points $t_1, \ldots, t_n$, where $t_i = (i-1)/(n-1)$, will be treated as arising from a equispaced, fixed design setting.

Based on Figure 1, replication origins correspond to local peaks in regions of the curve, while replication termination locations are valleys in the profile. Note that peaks and valleys in the curves will be points where the first derivative of the function is zero. Other situations in which the derivatives of a function are of interest have been given by Gasser et al. (1984) and Song et al. (1995).

Our approach will be to use nonparametric smoothing techniques to estimate μ. We will compare three methods: kernel regression, local weighted polynomial smoothing,

and smoothing splines. Define $\mu^{(k)}$ to be the kth derivative of μ. Based on the estimates of $\mu^{(1)}$ and $\mu^{(2)}$, the zero-crossings of $\mu^{(1)}$ where $\mu^{(2)} < 0$ correspond to candidate replication origins. We will then develop a statistic for assessing clustering of replication origin locations.

3.2 Kernel estimation

One method of estimation of μ is the ordinary kernel regression estimator (Jones, Sheather and Marron, 1996)

$$\hat{\mu}(t) \equiv \frac{1}{b}\sum_{i=1}^{n} Y_i g_i(t;b), \tag{2}$$

where

$$g_i(t;b) = \int_{s_{i-1}}^{s_i} W\left(\frac{t-u}{b}\right) du,$$

$s_i = (t_i + t_{i+1})/2$, W is the kernel function and b is the bandwidth. Recall that we are using a fixed design for t; this reflects the experimental design with respect to the chromosome positions in the study of Raghuraman et al. (2001). By simple differentiation, the estimator of the derivatives of μ, $\mu^{(k)}(t)$ is given by

$$\hat{\mu}^{(k)}(t;b) = \frac{1}{b^k}\sum_{i=1}^{n} Y_i g_i^{(k)}(t;b), \tag{3}$$

where $g_i^{(k)}(t;b) \equiv \int_{s_{i-1}}^{s_i} W^{(k)}\{(t-u)/b\}du$ is the kth derivative of $g_i(t;b)$ with respect of t.

A major issue in the construction of kernel estimates for $\mu^{(1)}$ and $\mu^{(2)}$ is the choice of bandwidth, b. As suggested by Jones, Sheather and Marron (1996), we will take a "solve-the-equation plug-in" approach to the problem of estimating b. The idea of this approach is to start with the formula of the asymptotically optimal bandwidth, which is found by minimizing a large-sample approximation to the mean squared error, and to find an estimate of b by iterative methods. Let $\mathrm{MISE}\{\hat{\mu}^{(1)}(\cdot;b)\} = \mathrm{E}[\mathrm{ISE}\{\hat{\mu}^{(1)}(\cdot;b)\}]$, where

$$\mathrm{ISE}\{\hat{\mu}^{(1)}(\cdot;b)\} = \int_0^1 a(t)\{\mu^{(1)}(t) - \hat{\mu}^{(1)}(t;b)\}^2 dt. \tag{4}$$

In (4), function $a(t)$ is assumed to be continuously differentiable with support $[\delta, 1-\delta]$ for some $\delta > 0$, and $a(t) > 0$ for $t \in [\delta, 1-\delta]$. The role of $a(t)$ is to dampen the behavior of the estimate at the boundary; it is only used to help derive the optimal bandwidth. In practice, boundary kernels (Gasser et al., 1991) are used for estimation on the boundary near 0 and 1.

Observe that $W^{(k)}$ in the definition of $g_i^{(k)}$ in (3) is also a kernel function. Suppose we make the following assumptions:

(A1) The support of $W^{(k)}$ is $[-1, 1]$;

(A2) $\int W^{(k)}(v)v^j dv = 0$, for $0 \leq j < k$;

(A3) $\int W^{(k)}(v)v^k dv = (-1)^k k!$.

Since we have assumed that μ is $k+2$-times differentiable, by standard mean-squared error arguments, the optimal bandwidth for estimating $\mu^{(k)}$ is

$$b_{opt} = \left(\frac{\sigma^2}{n} \frac{(2k+1)}{2(l-k)} \frac{V}{B} \right)^{1/(2l+1)}, \tag{5}$$

where $l = k + 2$,

$$B = \frac{(-1)^l}{l!} \left\{ \int_0^1 W^{(k)}(v)v^l dv \right\}^2 \int_0^1 a(v) \left\{ \mu^{(k)}(v) \right\}^2 dv,$$

and

$$V = \int_0^1 a(v) \{W^{(k)}(v)\}^2 dv.$$

The numerator of (5) represents the variance of $\hat{\mu}^{(k)}(t; b)$, while the denominator comes from the squared bias term. Note, however, that estimation of (5) requires an estimate of $\hat{\mu}^{(k)}$. We will be using a pilot estimate of $\hat{\mu}^{(k)}$ that gets updated in the iterative algorithm given below. A fast estimator of σ^2, $\hat{\sigma}^2$ can be calculated using the methods of Gasser et al. (1991). Based on the resulting estimate $\hat{\sigma}^2$, we have the following iterative algorithm for estimating the optimal bandwidth for $\mu^{(1)}$, say b_*. Note that $k = 1$ and $l = 3$ here.

1. Set $\hat{b}_0 = 1/n$.

2. For $i = 1, 2, \ldots, 10$, set

$$\hat{b}_i = \left(\frac{\hat{\sigma}^2}{n} \frac{(2k+1)}{2(l-k)} \frac{V}{B_{i-1}} \right)^{1/7},$$

where

$$B_{i-1} = \frac{(-1)^l}{l!} \left\{ \int_0^1 W^{(k)}(v) v^l dv \right\}^2 \int_0^1 a(v) \left\{ \hat{\mu}^{(k)}(v; \hat{b}_{i-1} n^{1/10}) \right\}^2 dv$$

3. Set $\hat{b}_* = \hat{b}_{11}$.

Based on this estimated bandwidth, we obtain estimates of $\mu^{(1)}$ from the formula in (3). A similar algorithm can be developed to estimate the optimal bandwidth for $\mu^{(2)}$ with $k = 2$ and $l = 4$. The iterative algorithm was shown in Gasser et al. (1991) to have desirable asymptotic problems and good finite-sample behavior.

3.3 Locally weighted least squares estimation

Another method of nonparametrically estimating μ in (1) is by using locally weighted polynomial estimation techniques (Fan and Gijbels, 1996). The approach is to approximate $\mu(t)$ locally at a point t_0 by a simple polynomial of order p. Using Taylor series expansions yields that in a neighborhood of t_0,

$$\begin{aligned} \mu(t) &\approx \mu(t_0) + \mu^{(1)}(t_0)(t - t_0) + \frac{\mu^{(2)}(t_0)}{2!}(t - t_0)^2 \\ &\quad + \cdots + \frac{\mu^{(p)}(t_0)}{p!}(t - t_0)^p. \end{aligned} \tag{6}$$

We will take $p = 4$ in this paper so that we can estimate $\mu^{(1)}$ and $\mu^{(2)}$ well. Locally weighted polynomial estimation involves minimizing

$$\sum_{i=1}^{n} \{Y_i - \sum_{j=0}^{p} \beta_j (t_i - t_0)\}^2 K_h(t_i - t_0), \tag{7}$$

where $K_h(\cdot) = h^{-1} K(\cdot)$, K is a kernel weighting function that downweights points that are far away from t_0, and h is the bandwidth of the kernel. The solution that minimizes (7) is given by a weighted least squares solution. In particular, the solution of (7) is

$$\hat{\mu}(t; h, p) = \mathbf{e}_1' (\mathbf{T}_p' \mathbf{W} \mathbf{T}_p)^{-1} \mathbf{T}_p \mathbf{W} \mathbf{Y}, \tag{8}$$

where

$$\mathbf{T}_p = \begin{bmatrix} 1 & t_i - t & \cdots & (t_i - t)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n - t & \cdots & (t_n - t)^p \end{bmatrix},$$

$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{W} = \text{diag}[h^{-1}K\{(t_1 - t)/h\}, \ldots, h^{-1}K\{(t_1 - t)/h\}]$, and $\mathbf{e}_1$ is a column vector with 1 in the first entry and zero otherwise. Advantages of the locally weighted polynomial method include avoidance of the use of boundary kernels as well as appealing minimax properties (Fan and Gijbels, 1996, Sec. 3.2). Differentiation of (8) yields the local pth degree estimate of $\mu(t)$,

$$\hat{\mu}^{(1)}(t; h, p) = \mathbf{e}_2'(\mathbf{T}_p'\mathbf{S}\mathbf{T}_p)^{-1}\mathbf{T}_p\mathbf{S}\mathbf{Y}, \tag{9}$$

where $\mathbf{e}_2$ is a column vector with 1 in the second entry and zero elsewhere.

Taking the second derivative of (8) yields the estimated second derivative of $\mu(t)$,

$$\hat{\mu}^{(2)}(t; h, p) = \mathbf{e}_3'(\mathbf{T}_p'\mathbf{S}\mathbf{T}_p)^{-1}\mathbf{T}_p\mathbf{S}\mathbf{Y}, \tag{10}$$

where $\mathbf{e}_3$ is a column vector with 1 in the third entry and zero elsewhere.

Define the following terms: $\mu_j(K) = \int z^j K(z)dz$, $\mathbf{N}_p$ the $(p+1) \times (p+1)$ matrix with (i, j)th element $\mu_{i+j-2}(K)$ and $\mathbf{M}_p(u)$ the same as $\mathbf{N}_p$ but with the first column replaced by $(1, u, u^2, \ldots, u^p)'$. Because $p = 3$, the bias and variance calculations for $\hat{\mu}^{(1)}$ are given by (Wand and Jones, 1995, p. 137)

$$\begin{aligned} E\{\hat{\mu}^{(1)}(t; h, p) - \mu^{(1)}(t)|t_1, \ldots, t_n\} &= \left[\frac{1}{120}\mu_5(K_{1,3})\mu^{(5)}(t)\right. \\ &\quad \left. + \frac{1}{24}\left\{\mu_4(K_{1,3}) - \mu_4(K_{0,3})\right\}\frac{\mu^{(4)}(t)f^{(1)}(t)}{f(t)}\right]h^4 \\ &\quad + o_P(h^{p-r+2}) \end{aligned} \tag{11}$$

and

$$\text{Var}\{\hat{\mu}^{(1)}(t; h, p)\} = n^{-1}h^{-3}R(K_{1,3})\frac{\sigma^2}{f(t)} + o_P(n^{-1}h^{-3}), \tag{12}$$

where $f(t)$ and $f^{(1)}(t)$ are the density and its derivative of $t_1, \ldots, t_n$, and

$$K_{r,p}(u) = \frac{r!|\mathbf{M}_{r,p}|K(u)}{|\mathbf{N}_p|}.$$

Similar calculations give the bias and variance of $\hat{\mu}^{(2)}$ to be

$$\begin{aligned} E\{\hat{\mu}^{(2)}(t;h,p) - \mu^{(2)}(t)|t_1,\ldots,t_n\} &= \left[\frac{1}{120}\mu_5(K_{1,3})\mu^{(5)}(t)\right. \\ &\quad \left. + \frac{1}{24}\left\{\mu_4(K_{1,3}) - \mu_4(K_{0,3})\right\}\frac{\mu^{(4)}(t)f^{(1)}(t)}{f(t)}\right] h^4 \\ &\quad + o_P(h^{p-r+2}) \end{aligned} \tag{13}$$

and

$$\mathrm{Var}\{\hat{\mu}^{(2)}(t;h,p)\} = n^{-1}h^{-3}R(K_{1,3})\frac{\sigma^2}{f(t)} + o_P(n^{-1}h^{-3}), \tag{14}$$

As with kernel regression estimation, the major issue is selection of the bandwidth. We will use the "solve-the-equation plug-in" method of Ruppert, Sheather, and Wand (1995), which is an adaptation of the ideas of Gasser et al. (1991) to the locally weighted polynomial setting. Ruppert et al. (1995) demonstrate the convergence of the density estimators based on the "solve-the-equation plug-in" method to the MISE-optimal bandwidth.

3.4 Penalized smoothing spline estimation

The last method of nonparametric estimation we will examine is smoothing spline estimation (Green and Silverman, 1994). Smoothing spline methods focus on estimating μ by minimizing the following objective function:

$$\sum_{i=1}^{n}\{Y_i - \mu(t_i)\}^2 + \lambda\int\{m^{(2)}(t)\}^2, \tag{15}$$

where $\lambda > 0$ denotes a smoothing parameter. If $\lambda = 0$, then this corresponds to interpolation in the data, while $\lambda = \infty$ corresponds to fitting a simple linear regression to the data (Y_i, t_i), $i = 1,\ldots,n$. The model complexity in (15) is effectively controlled by λ.

The minimization problem in (15) is given by a cubic spline on the interval $[t_{(1)}, t_{(n)}]$. Denote the solution to (15) as $\tilde{\mu}$. The spline solution $\tilde{\mu}$ has the following properties (Härdle, 1990):

1. $\tilde{\mu}$ is a cubic polynomial between two successive t-values;

2. At the observation points t_i, the curve $\tilde{\mu}$ and its first two derivatives are continuous

3. At the boundary points $t_{(1)}$ and $t_{(n)}$, the second derivative of $\tilde{\mu}(t)$ is zero.

Let D^m denote the m-th differential operator. We will be taking $m = 4$. Then the optimization problem in (15) is equivalent to finding μ to minimize

$$\sum_{i=1}^{n} \{Y_i - \mu(t_i)\}^2 + \lambda \int \{D^m \mu(t)\}^2 dt. \tag{16}$$

We will be using the algorithm suggested by Heckman and Ramsay (2000) for estimation here. If we define the $m \times m$ Wronskian matrix $\mathbf{W}(t)$ to have (i,j)th element $\mathbf{W}(t)_{ij} = D^{(j-1)} b_i(t)$, where $b_1(t), \ldots, b_m(t)$ are the basis functions, then simple differential equation theory yields that the basis functions are $\{1, t, t^2, t^3\}$. We can then define the Green's function $G(s,t)$ as

$$G(s,t) = \begin{cases} \sum_{i=1}^{m} b_i(s) b_i^*(t), & t \leq s \\ 0 \quad \text{otherwise} \end{cases},$$

where $b_i^*(t)$ $(i = 1, \ldots, m)$ is the last row of the inverse of $\mathbf{W}(t)$. We can then define a $n \times n$ kernel matrix $\mathbf{K}$, where the (i,j)th element is $K_{ij} = k(t_i, t_j)$, and

$$k(s,t) = \int_0^1 G(s,u) G(t,u) du.$$

The kernel matrix represents the reproducing kernel function in the reproducing kernel Hilbert space (Wahba, 1990). The theory gives that the minimizer of (16) is of the form

$$\hat{\mu}(t) = \sum_{j=1}^{m} \eta_j b_j(t) + \sum_{j=1}^{n} \gamma_j k(t_j, t).$$

Numerically, it is found by minimizing

$$(\mathbf{Y} - \mathbf{B}\eta - \mathbf{K}\gamma)'(\mathbf{Y} - \mathbf{B}\eta - \mathbf{K}\gamma) + \lambda \gamma' \mathbf{K} \gamma,$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\eta = (\eta_1, \ldots, \eta_m)$, $\gamma = (\gamma_1, \ldots, \gamma_n)$, and $\mathbf{U}$ is an $m \times m$ matrix with (i,j)th entry $u_i(t_j)$. While this will generally lead to a system of linear equations that are solved in $\mathbf{O}(n^3)$ operations, Heckman and Ramsay (2000) propose a $\mathbf{O}(n)$

algorithm for estimation. If we let $\hat{\mu}_\lambda = (\hat{\mu}_\lambda(t_1), \ldots, \hat{\mu}_\lambda(t_n))'$, then for some vector $\mathbf{v} \in R^{n-m}$, $\hat{\mu}_\lambda = \mathbf{Y} - \lambda \mathbf{Q}\mathbf{v}$, where

$$(\mathbf{Q}'\mathbf{K}\mathbf{Q} + \lambda \mathbf{Q}'\mathbf{Q})\mathbf{c} = \mathbf{Q}'\mathbf{Y}, \tag{17}$$

where $\mathbf{Q}$ is any $n \times (n-m)$ matrix of full column rank with $\mathbf{Q}'\mathbf{U} = 0$. The algorithm then involves finding a banded matrix $\mathbf{Q}$ with $\mathbf{Q}'\mathbf{U} = 0$ using the Cholesky decomposition. Once this matrix is generated, the solution $\mathbf{c}$ of (17) can then be found. This then yields a value for $\hat{\mu}_\lambda$. To calculate the values of the derivative, we first solve the equation $\hat{\mu}_\lambda - \mathbf{K}\mathbf{Q}\mathbf{c} = \mathbf{U}\mathbf{d}$ for $\mathbf{d}$. Then the lth derivative of $\hat{\mu}_\lambda$ is given by

$$\hat{\mu}_\lambda^{(l)} = \mathbf{U}^{(l)}\mathbf{d} + \mathbf{K}^{(l)}\mathbf{Q}\mathbf{c},$$

where $\mathbf{U}^{(l)}$ is an $m \times m$ matrix has (i,j)th element $D^l u_j(t_i)$ and $\mathbf{K}^{(l)}$ is an $n \times n$ matrix with (i,j)th element $D^l k(t_j, t_i)$ An important issue in the estimation of $\tilde{\mu}$ is choice of λ. We will use generalized cross-validation (Wahba, 1990) for this.

Once we have an estimator of $\mu^{(1)}$ and $\mu^{(2)}$ using either nonparametric regression estimator, we then define candidate replication origin and replication termination sites as being zero-crossings of the estimated derivative. Since we are using constant bandwidth methods for estimation, the zero-crossings correspond to the values of t where $\hat{\mu}^{(1)}(t-)\hat{\mu}^{(1)}(t+) < 0$. Candidate replication origins from these values of t are those where $\mu^{(2)}(t) < 0$, while candidate replication termination sites are those where $\mu^{(2)}(t) > 0$.

3.5 Clustering methods

As mentioned in Section 2, a hypothesis has been put forward that the location of replication origins is not random throughout the chromosome. An argument for this hypothesis was put forward by Gilbert (2001). He argued that if the positions of replication origins were distributed randomly across the chromosome, then some origins might be physically too far apart for complete replication of DNA to occur within the S phase of the cell cycle.

Based on the data of Raghuraman et al. (2001), we are in a position to test this hypothesis. Given the replication origins found using the methods of the previous sections, we can now test the hypothesis of clustering of replication origins. The null hypothesis, H_0, is that the the locations of the replication origins are uniformly distributed throughout the chromosome, while the alternative hypothesis is that the replication origins cluster.

There are two types of hypotheses involving clustering that we wish to distinguish. The first is that there is no clustering of replication origin locations throughout the chromosome; this will be referred to as a global null hypothesis of clustering. If the global null hypothesis is true, then the replication origin locations found are consistent with that of randomly generated locations. Another type of clustering hypothesis involves determining whether or not a particular cluster is significant, this will be referred to as a local hypothesis of clustering. We will use scan statistics (Glaz et al., 2001) to test local hypotheses of null clustering.

We start by considering the global clustering null hypothesis. The Kolmogorov-Smirnov statistic is used to test this hypothesis. If $F_m(x)$ denotes the empirical cumulative distribution function of the putative replication origins, scaled to the interval $[0, 1]$, the Kolmogorov-Smirnov statistic for testing the global null hypothesis of m random replication origins is

$$D = \sup_x \sqrt{m}(|F_m(x) - x|).$$

While small values of D are consistent with the null hypothesis, large values of D suggest that the locations of the replication origins are not random and will lead to rejection of the null hypothesis. The null distribution of this test statistic is tabulated and p-values can be given by standard statistical software.

We now turn to the problem involving local inference about the clusters. Let $(X_1, \ldots, X_{m_j})$ be the locations of the replication origins for the jth chromosome; these are the locations estimated using the methods of §3.2 – 3.4. In the sequel, we suppress the dependence of m_j on j. We consider the r-scan statistic (Karlin and Macken, 1991;

Dembo and Karlin, 1992):

$$R_i = \sum_{l=i}^{i+r-1} X_{l+i} - X_l.$$

Note that R_i is the total distance between putative replication origin locations starting from the ith location with a window size of r locations. To assess clustering, we would use m_k^r, the kth smallest R_i. Smaller values of m_k^r correspond to stronger evidence of clustering. If the locations of the replication origins were scattered randomly on the chromosome, then by approximation results in Karlin and Macken (1991),

$$Pr\left(m_k^r < \frac{x}{n^{1+1/r}}\right) \approx 1 - \exp(-\lambda)\left(\sum_{i=0}^{k-1} \frac{\lambda^i}{i!}\right), \tag{18}$$

where $\lambda = x^r/r!$. In (18), x is chosen such that the probability equals 0.01, following previous recommendations (Karlin and Macken, 1991).

4 Yeast Data

We now apply the proposed methodology to the data discussed in §2. Because of numerical error, we define replication origins as locations with estimated first derivative less than 1×10^{-6} in magnitude and second derivative less than -1×10^{-9}. A significance test on the results was done by the following permutation scheme:

1. Gene expression measurements were shuffled within each chromosome.
2. The analysis was repeated and candidate replication origins were determined.
3. Steps 1 and 2 were repeated 10000 times.

The number of replication origins per chromosome is given in Table 1. The corresponding number in parentheses represents the average number of replication origins found, averaged across the 10000 permuted datasets. This represents the expected number of false positives. While the locations found by the three methods do not show perfect concordance, the general locations appear to be consistent. However, the column totals are bigger than the 200-400 replication origins commonly believed. We return to this point in the Discussion.

The next step was to assess the clustering of replication origins on both a global (i.e., chromosomewide) and local basis. Based on the Kolmogorov-Smirnov statistic, there was no evidence of clustering using any of the methods for identifying replication origins based on Table 1. The scan statistic with different choices of r also fails to any statistically significant clusters.

5 Discussion

In this article, we have developed the use of nonparametric regression, Kolmogorov-Smirnov and scan statistics in order to identify replication origins from microarray data and to test a hypothesis put forward by Gilbert (2001) as to whether replication of origins occur randomly in the eukaryotic genome.

Our analysis came up with two relatively surprising conclusions. The first is that the number of predicted replication origins (summarized in Table 1) is much bigger than the 200-400 commonly believed to exist. It should be pointed out that the origins represent computational predictions and would be need to validated in the lab to determine if they are true or not.

The second conclusion is that there is no evidence to suggest that clustering of replication origins occurs on either a chromosomal basis or a more local basis. Potential limitations of the analysis include a lack of experimental replication and experimental-specific artifacts that contribute to additional sources of variation.

There were several key points of note to this analysis. First, based on the microarray data generated, we were interested in studying its pattern as a function of chromosome position. Thus, this analysis presents one method of incorporating biological information with high-throughput genomic data. Second, given the structure of the microarray experiment used, the interest was on the first and second derivatives of the gene expression profile. This is different from many studies in which only the mean structure is considered.

While replication origins are of biological interest, there is very little biological background regarding them that we could have incorporated into the statistical modelling.

This was our main motivation for utilizing nonparametric methods for identification of replication origins. If more biological knowledge regarding replication origins and their patterns of gene expression were available, model-based techniques might be feasible. While the Kolmogorov-Smirnov might have lower power than a model-based approach, its results are relatively robust.

One outstanding issue that remains is assessment of variability of the estimated replication origins. Because the variance of the zero-crossings is related to the variance of the estimated derivatives, it will be relatively wiggly. One approach that we used is the parametric bootstrap to assess the variability in the zero-crossings. We fit the model (1) using one of the estimation methods described in Section 3.2 - 3.4 and then determined optimal bandwidths or smoothing parameters. We then calculated residuals $\hat{e}_i \equiv Y_i - \hat{\mu}(t_i)$ $(i = 1, \ldots, n)$. For each chromosome, we generate a new dataset $(Y_1^b, \ldots, Y_n^b)$, $b = 1, \ldots, B$, where

$$Y_i^b = \tilde{\mu}(t_i) + e_i^*,$$

$e_1^*, \ldots, e_n^*$ is a random sample from a normal distribution with mean 0 and variance $\tilde{\sigma}^2$, where

$$\tilde{\sigma}^2 = 10 \times b.$$

Note that we are oversmoothing the bootstrap, as suggested by Davidson and Hinkley (1997). Changing $\tilde{\sigma}^2$ to other multiples of b did not substantially alter the results. We then apply the estimation procedures to obtain an estimate of the derivative of μ. We then calculate a stability score for the location of an observed zero-crossing as the number of bootstrapped datasets in which it is also a zero-crossing. However, this yielded very small values. This suggests that perhaps even more parametric methods of inference are needed in this area.

The experimental data described in the paper were generated using DNA microarrays; note that the data are quite different than those considered in other papers, which typically two conditions, such as cancer versus noncancer tissue (e.g. Tusher et al., 2001). Increasingly, high-throughput genomic and proteomic studies will become more

commonplace, so it will be important to develop methods for these high-dimensional small sample size settings.

Acknowledgments

The research of the author is supported by grant GM72007 from the Joint DMS/DBS/NIGMS Biological Mathematics Program. He thanks Tom Braun for very helpful comments on a draft manuscript.

References

Bell SP, Dutta A (2002) DNA replication in eukaryotic cells. Ann Rev Biochem 71: 333 – 374

Cheng MY, Hall P (1998) Calibrating the excess mass and dip tests of modality. J R Stat Soc Ser B 60: 579 – 589

Davison A, Hinkley D (1997) Bootstrap Methods and their Application. Cambrige University Press, Cambridge

Dembo A, Karlin S (1992) Poisson approximations for r-scan processes. Ann Appl Prob 2: 329 – 357

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

Fisher NI, Marron JS (2001) Mode testing via excess mass estimate. Biometrika 88: 499 – 517

Gasser T, Kneip A, Köhler W (1991) A fast and flexible method for automatic smoothing. J Am Stat Assoc 86: 643 – 652

Gasser T, Müller HG, Köhler W, Molinari L, Prader A (1984) Nonparametric regression analysis of growth curves. Ann Statist 12: 210 – 224

Gilbert DM (2001) Making sense of eukaryotic DNA replication origins. Science 2001: 96 – 100

Glaz J, Naus J, Wallenstein S. (2001) Scan Statistics. Springer-Verlag: New York

Green P, Silverman BW (1994) Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach. Chapman and Hall: London

Härdle W (1990) Applied Nonparametric Regression. Cambridge University Press: Cambridge

Heckman NE, Ramsay JO (2000) Penalized regression with model-based penalties. Canad J Stat 28: 241 – 258

Jones MC, Marron JS, Sheather SJ (1996) A brief survey of bandwidth selection for density estimation. J Am Stat Assoc 91: 401 – 407

Karlin S, Macken C (1991) Assessment of inhomogeneities in an *E. coli* physical map. Nucl Acids Res 19: 4241 – 4246

Müller DW, Sawitzki G (1991) Excess mass estimates and tests for multimodality. J Am Stat Assoc 86: 738–746

Newlon CS, Theis JF (2002) DNA replication joins the revolution: whole-genome views of DNA replication in budding yeast. BioEssays 24: 300 – 304

Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, Conway A, Lockhart DJ, Davis RW, Brewer BJ, Fangman WL (2001) Replication dynamics in the yeast genome. Science 294: 115 – 121

Ruppert D, Sheather SJ, Wand MP (1995) An effective bandwidth selector for local least squares regression. J Am Stat Assoc 90: 1257 – 1270

Silverman BW (1981) Using kernel density estimates to investigate multimodality. J R Stat Soc Ser B 43: 97 – 99

Song KS, Müller HG, Clifford AJ, Furr HC, Olson JA (1995) Estimating derivatives of pharmacokinetic response curves with varying bandwidths. Biometrics 51: 12 – 20

Stinchcomb DT, Struhl K, Davis RW (1979) Isolation and characterisation of a yeast chromosomal replicator. Nature 282: 39 – 43.

Truong YK, Scott RS, Vos JMH (2002) The origin of DNA replication and Fieller's problem. Stat Med 21: 3571 – 3582

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to ionization radiation response. Proc Natl Acad Sci 98: 5116–5121

Uetz P et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerivisiae. Nature 403: 623 – 627

Venter JC et al (2001) The sequence of the human genome. Science 291: 1304 – 1351

Wahba G (1990) *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics: Philadelphia

Wyrick JJ, Aparicio JG, Chen T, Barnett JD, Jennings EG, Young RA, Bell SP, Aparicio OM (2001) Genome-wide of ORC and MCM proteins in *S. cerivisiae.* Science 294: 2357 – 2360

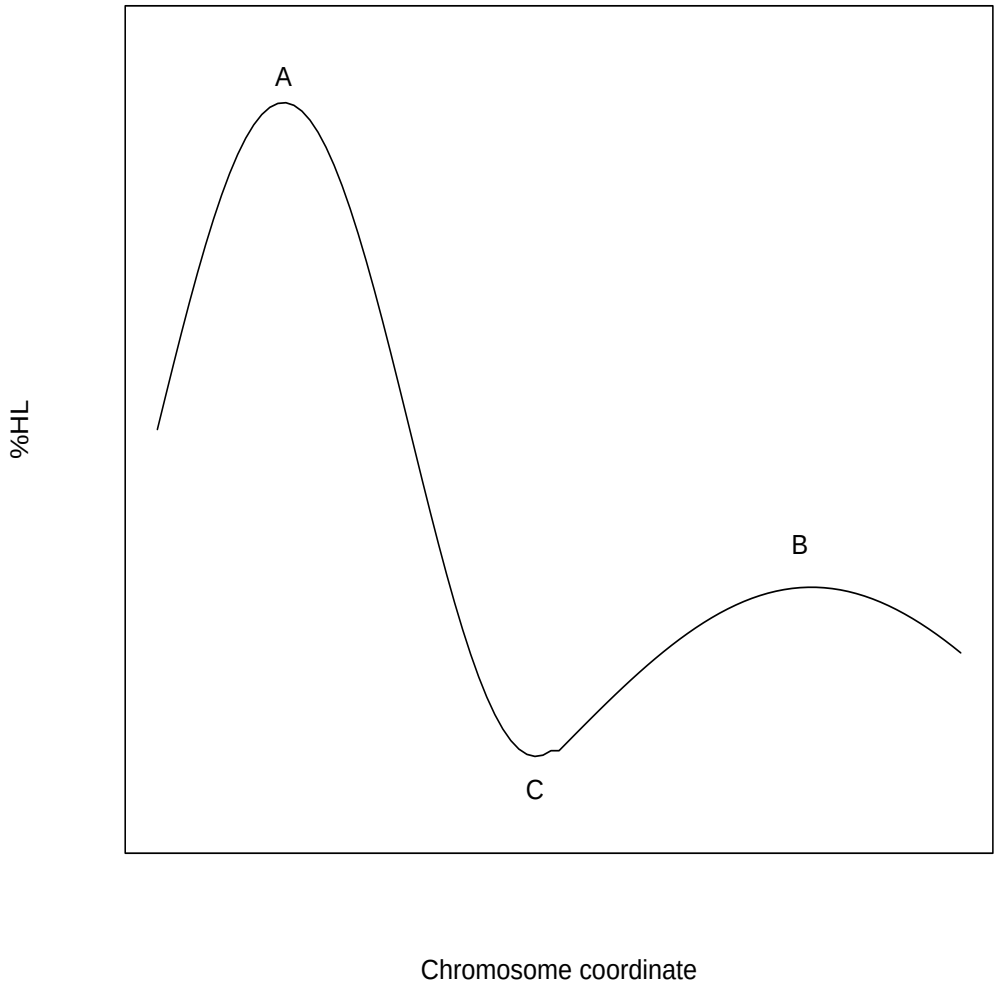

Figure 1: Schematic of replication origins and termination in a chromosome. The horizontal is position on a chromosome, while the vertical axis is the percentage of HL relative to total. Point A represents an early replication origin, while B indicates a late replication origin. Point C (valley) is a replication terminus.

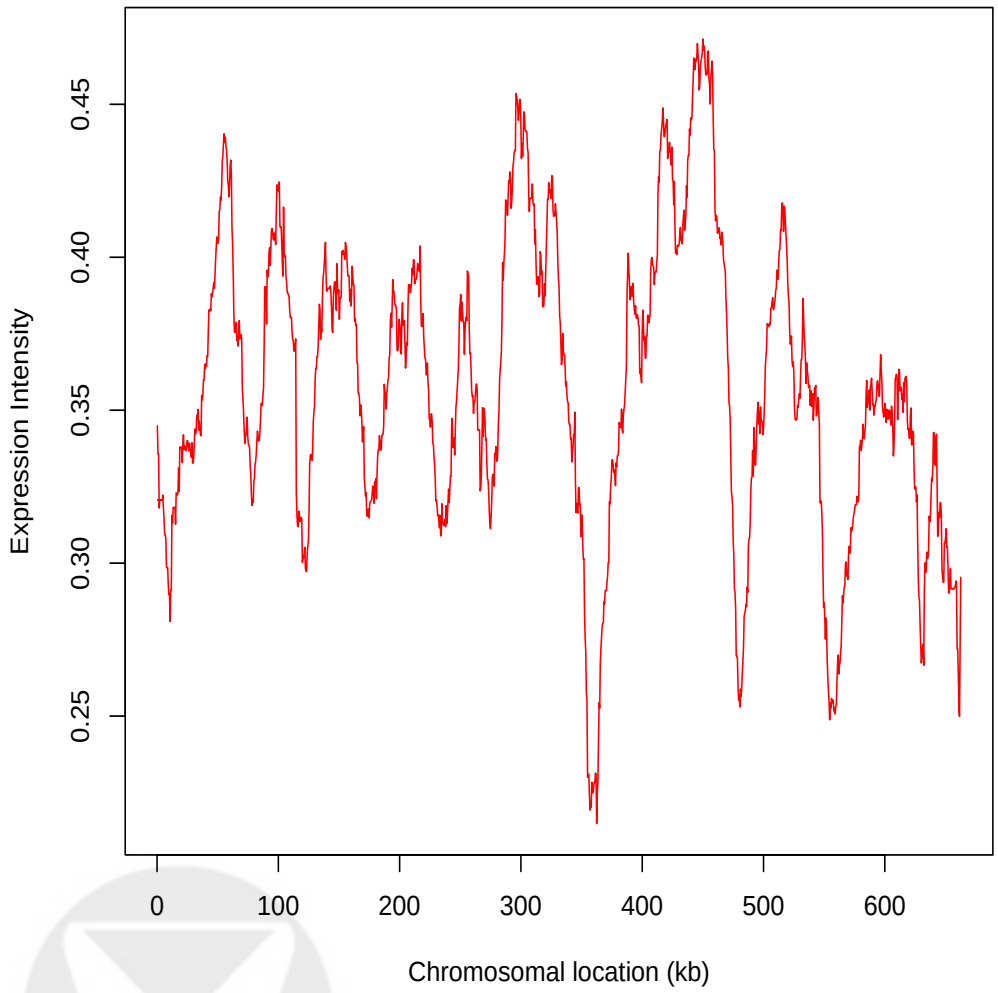

Figure 2: Gene expression profile of Chromosome 11 as a function of location from microarray experiment by Raghuraman et al. (2001).

Table 1: Number of candidate replication origins found for each chromosome in yeast data of Raghuraman et al. (2001) using methods developed in §3.2 – §3.4.

Chromosome	Kernel estimation	Locally weighted LS	Smoothing spline
1	38 (8)	47 (9)	40 (6)
2	130 (14)	140 (21)	158 (29)
3	42 (6)	47 (6)	60 (11)
4	280 (50)	290 (58)	319 (54)
5	90 (10)	102 (11)	99 (16)
6	36 (5)	47 (7)	44 (9)
7	170 (32)	189 (34)	221 (27)
8	90 (12)	98 (14)	113 (15)
9	85 (9)	80 (14)	91 (14)
10	120 (16)	142 (28)	133 (14)
11	108 (12)	130 (18)	118 (20)
12	132 (25)	158 (22)	196 (39)
13	120 (16)	162 (28)	207 (32)
14	134 (21)	153 (31)	188 (19)
15	176 (18)	239 (33)	256 (48)
16	152 (24)	164 (28)	205 (36)