Funct Integr Genomics (2005) 5: 28–31
DOI 10.1007/s10142-004-0122-1

ORIGINAL PAPER

Debashis Ghosh

Nonparametric methods for analyzing replication origins in genomewide data

Received: 5 December 2003 / Revised: 13 August 2004 / Accepted: 16 August 2004 / Published online: 9 September 2004

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 analyze data from a recent yeast study in which candidate replication origins were profiled using cDNA microarrays to test this hypothesis. We find no evidence for such clustering.

Keywords Changepoint · Density estimation · Derivative estimation · Gene expression · Kernel smoothing

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.

One question involves the existence and location of replication origins. The biology underlying this problem is further detailed in "Biological background." 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. One such example is given in Fig. 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. Replicates are not available in many genomic studies. Most statistical methods will not be computationally feasible for finding multiple 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. "Biological background" provides more details on the biology of replication origins and describes the experiment of Raghuraman et al. (2001). A statistical model for the analysis of the expression profiles and methods for identification of replication origins are given in "Statistical methods;" this section also describes nonparametric methods for assessing clustering. The proposed methodology is applied to the yeast data of Raghuraman et al. (2001) in "Yeast data." Finally, we conclude with some discussion in "Discussion." Much of the technical detail for this work can be found in Ghosh (2004).

Biological background

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. In a recent study by Raghuraman et al. (2001), oligonucleotide microarrays were used to identify potential origins of replications. Yeast cells were grown for many generations in medium containing two dense isotopes. At times t =0, 10, 14, 19, 25, 33, 44 and 60 min in the S phase, culture samples were collected. The replicated DNA containing one heavy and one light (HL) strand (for the parent and

D. Ghosh (✉)
Departments of Biostatistics, School of Public Health,
University of Michigan,
1420 Washington Heights, Room M4057,
Ann Arbor, MI, 48109-2029, USA
e-mail: ghoshd@umich.edu
Tel.: +1-734-6159824
Fax: +1-734-7632215

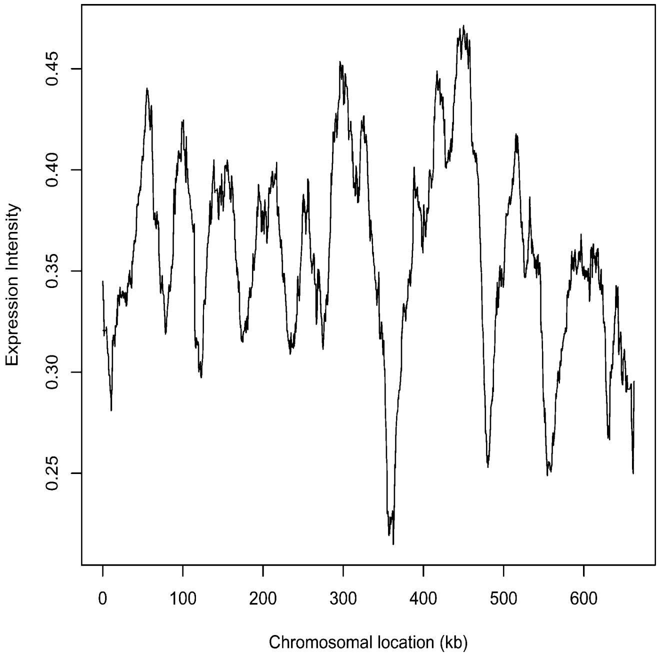

Fig. 1 Gene expression profile of chromosome 11 as a function of location from the microarray experiment by Raghuraman et al. (2001)

daughter strands) were separated from the unreplicated DNA [which contained two heavy (HH) strands] using density gradient centrifugation. The HH and HL DNA were then separately hybridized to an oligonucleotide microarray, which yielded an intensity measure. The intensity measure represents 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 Fig. 2. 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 time 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.

An example of the data we analyze is given in Fig. 1. 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.

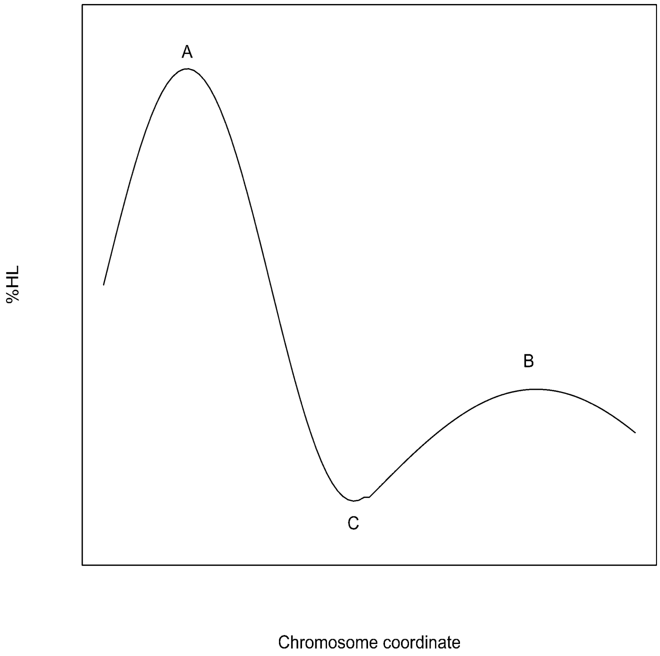

Fig. 2 Schematic of replication origins and termination in a chromosome. The *horizontal* is position on a chromosome, while the *vertical axis* is the percentage of *HL* (replicated DNA containing one heavy and one light strand) relative to total. Point *A* represents an early replication origin, while *B* indicates a late replication origin. Point *C* (valley) is a replication terminus

Statistical methods

We observe the data $\{Y_{ij}\}$, I =1,⋯, I, j =1,⋯, 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

$$Y_{ij} = \mu_I(j/n_I) + \varepsilon_{ij}, \qquad (1)$$

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 Eq. 1 are a random sample from a normal distribution with mean zero and variance σ_I^2, I =1,⋯, I. We will be treating each chromosome separately, so we will suppress dependence on of μ_I and σ_I^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,⋯, t_n, where $t_I = (I-1)/(n-1)$, will be treated as arising from a equispaced, fixed design setting.

Based on Fig. 1, the 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 using locally weighted polynomial smoothing (Fan and Gijbels 1996). 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 use the "solve-the-equation plug-in" method of Ruppert et al. (1995) to estimate the variance and bandwidth.

Given the replication origins found using the methods of the previous sections, we can now test the hypothesis that they cluster. The null hypothesis, H_0, is that 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. 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 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.

We now turn to the problem involving local inference about the clusters. Let $\left(X_1, \ldots, X_{m_j}\right)$ be the locations of the replication origins for the jth chromosome; these are the locations estimated using the methods outlined above. 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; Glaz et al. 2001)

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

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)

$$\Pr\left(m_k^r < xn^{1+1/r}\right) \approx 1 - \exp(-\lambda)\left(\sum_{i=0}^{k-1} \lambda^i i!\right), \qquad (2)$$

where $\lambda = x^r/r!$. In Eq. 2, x is chosen such that the probability equals 0.01, following previous recommendations (Karlin and Macken 1991).

Yeast data

We now apply the proposed methodology to the data discussed in "Biological background." Because of numerical error, we define replication origins as locations with an 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 10,000 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 10,000 permuted datasets. This represents the expected number of false positives. The estimated proportion of false positives based on the permutation scheme appears to be in the order of 10–20%. 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 (r =4−20) also fails to identify any statistically significant clusters at a significance level of 0.1.

Table 1 Analysis of replication origins using yeast data of Raghuraman et al. (2001) based on methods developed in the "Statistical methods" section. *Numbers* denote estimated number of replication origins, while those *in parentheses* represent estimated number across 10,000 permuted datasets

Chromosome	Locally weighted LS
1	47 (9)
2	140 (21)
3	47 (6)
4	290 (58)
5	102 (11)
6	47 (7)
7	189 (34)
8	98 (14)
9	80 (14)
10	142 (28)
11	130 (18)
12	158 (22)
13	162 (28)
14	153 (31)
15	239 (33)
16	164 (28)

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 need to be validated in the lab to determine if they are true or not. Even subtracting out the estimated number of false positives based on the permutation scheme still yields more than 400 replication origins.

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.

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

References

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

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 & Hall, New York

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

Ghosh D (2004) Nonparametric methods for analyzing replication origins in genomewide data. Technical report. Department of Biostatistics, University of Michigan. http://www.bepress.com/umichbiostat/paper32

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

Glaz J, Naus J, Wallenstein S (2001) Scan statistics. Springer, New York Berlin Heidelberg

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

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

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

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

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