CCIT Report #687 February 2008

Gene Expression

A Fast and Flexible Method for the Segmentation of aCGH Data

Erez Ben-Yaacov[1], Yonina Eldar[1,*]

[1]Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa Israel.

Received on February 11, 2008; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Array Comparative Genomic Hybridization (aCGH) is used to scan the entire genome for variations in DNA copy number. A central task in the analysis of aCGH data is the segmentation into groups of probes sharing the same DNA copy number. Some well known segmentation methods suffer from very long running times, preventing interactive data analysis.

Results: We suggest a new segmentation method based on wavelet decomposition and thresholding, which detects significant breakpoints in the data. Our algorithm is over 1,000 times faster than leading approaches, with similar performance. Another key advantage of the proposed method is its simplicity and flexibility. Due to its intuitive structure it can be easily generalized to incorporate several types of side information. Here we consider two extensions which include side information indicating the reliability of each measurement, and compensating for a changing variability in the measurement noise. The resulting algorithm outperforms existing methods, both in terms of speed and performance, when applied on real high density CGH data.

Availability: Implementation is available under software tab at: http://www.ee.technion.ac.il/Sites/People/YoninaEldar/

Contact: yonina@ee.technion.ac.il

1 INTRODUCTION

Array Comparative Genomic Hybridization (aCGH) is used to scan the entire genome for variations in DNA copy number. DNA from a test and reference cell populations is differentially labeled and hybridized on the array, and the log ratio between the two hybridization results is used to detect copy number variations. High density aCGH, spanning hundreds of thousands of probes, is a powerful tool in the research of cancer (Barrett *et al.,* 2004, Pinkel and Albertson 2005) and copy number polymorphisms (Conard *et al.,* 2006, Redon *et al.,* 2006, Perry *et al.,* 2008). A central task in the analysis of aCGH is the segmentation of the data into groups of probes that share the same DNA copy number.

Various segmentation methods have been proposed over the last years. Olsen *et al.* (2004) suggested a circular binary segmentation (CBS) algorithm, based on recursively applying a statistical test to detect significant breakpoints in the data. Picard *et al.* (2005) developed a dynamic programming procedure to segment the data when the number of segments is known in advance, which is referred to as CGHseg. The actual number of segments in real data is determined by maximizing a penalized likelihood function. While other segmentation methods exist, such as Lipson *et al.* (2005, 2006), a comparison study (Lai *et al.,* 2005) which tested 11 segmentation methods, concluded that CBS and CGHseg tend to have the best performance under various conditions. Another comparison study (Willenbrock *et al.,* 2005) compared 3 methods and proclaimed CBS as the method with the best results.

While presenting good segmentation performance, CBS is not sensitive to short segments, and often fails to detect them. On the other hand, CGHseg is sensitive to outliers in the data, leading to short segments corresponding to noise. A common drawback of both CBS and CGHseg is the long running time required to segment real high density arrays. Furthermore, it is not clear how to extend these methods to support side information.

In this paper we present *HaarSeg*, a new segmentation method, based on well known wavelet denoising principles. HaarSeg identifies statistically significant breakpoints in the data, using the maxima of the Haar wavelet transform, and segments accordingly. HaarSeg is a fast method, over 1,000 times faster than CBS and CGHseg, enabling interactive data analysis, with a slight compromise in performance. Due to its simple and intuitive structure, it is also a flexible method, and therefore easy to extend. We show how HaarSeg can be generalized to use quality of measurement data, additional information which exists in some platforms, indicating the reliability of each measurement. The use of quality of measurement was first suggested in Lipson *et al.* (2005), and it is currently used in ADM2, a segmentation algorithm based on Lipson *et al.* (2005, 2006) and used for example in de Smith *et al.* (2007) and in Perry *et al.* (2008). Since ADM2 does not have a freely available implementation we did not compare our performance to this segmentation algorithm. We also suggest an extension to compensate for the large variance in the log ratio measurements which occurs when one of the raw measurements has a very low value. Using these two generalizations, we show that HaarSeg outperforms existing methods, while remaining much faster.

The use of the Haar wavelet for microarray analysis is not new. Hsu *et al.* (2005) suggested applying standard wavelet denoising on microarrays, using the Haar wavelet. HaarSeg is different from that approach as it performs segmentation rather than smoothing of the data. To emphasize this difference we compare our results to Hsu *et al.*, and show that HaarSeg outperforms this method as well.

The rest of the paper is organized as follows: The basic HaarSeg algorithm is discussed in Section 2.1.Generalizations including quality of measurement data and adaptation to non-stationary variance are presented in Section 2.2 In Sections 3.1 and 3.2 we pro-

*To whom correspondence should be addressed.