Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

SaEP3.2

Statistical Framework for Quantitative Analysis of Array CGH

Shishir Shah, *Member, IEEE*

Abstract— **Over the last several years there has been an explosion of microarray technology in the biosciences, medical sciences, biotechnology, and pharmaceutical industry. The technology has centered on providing a platform for determining the gene expression profiles of hundreds to tens of thousands of genes (or transcript levels of RNA species) in tissue, tumors, cells, or biological fluids in a single experiment. In recent years, this technology has been extended to include the use of microarrays to study genomic DNA for gains and losses of chromosomal regions. This has become possible through the attachment of large genomic fragments such as BACs (Bacterial Artificial Chromosomes). In this paper, we present a methodology to model a CGH (Comparative Genomic Hybridization) profile as a statistical process and solve for distribution parameters to determine genomic changes across the genome, including whole chromosome gains and losses, and focal point variations that are commonly seen in solid tumors and genetic disorders.**

I. Introduction

In their most generic form, microarrays are ordered sets of DNA molecules attached to a solid surface. As can be appreciated, microarrays have been exploited for gene expression studies but other applications can be envisioned and developed. One such application is the use of microarrays to study genomic DNA for gains and losses of chromosomal regions. As our understanding of the sequence, structure and function of the human genome increases, fluctuations in DNA sequence copy number with concomitant microscopic or cryptic chromosomal aberrations are becoming increasingly correlated with phenotypic abnormalities [1-4]. This is particularly important in medicine as many diseases, cancers, and syndromes are caused by deletions, amplifications, and duplication of DNA segments. Classic examples include the deletion of 15q11.2 in Prader-Wili Syndrome [5], trisomy 21 in Down's syndrome [6], and the amplification of erbB2 in breast cancer [7]. In cancer biology, the development of most solid tumors follows a defined series of histopathological stages, involving multiple genetic changes such as translocations, deletions, duplications and alterations in ploidy (chromosomal copy number changes). Constitutional changes in DNA sequence copy number have now been well documented for a number of genetic syndromes, while acquired changes are receiving tremendous attention by virtue of their association with neoplastic transformations. Recently, with the advent of BAC array technology and the ability to screen the entire genome, has re-drawn the attention to the applicability of CGH, albeit array CGH, in the routine cytogenetics laboratory. So-called array or matrix Comparative Genomic Hybridization (CGH) utilizes mapped DNA sequences in a microarray format as an alternative platform for the CGH analyses. In recent years, several reports have described adaptation of microarray technology to the study of genomic alterations [4,8].

The purpose of array-based CGH is to detect and map chromosomal aberrations, on a genomic scale, in a single experiment. Since chromosomal copy numbers cannot be measured directly, two samples of genomic DNA (referred to as the reference and test DNAs) are differentially labelled with fluorescent dyes and competitively hybridized to known mapped sequences (referred to as BACs) that are immobilized on a slide. Subsequently, the ratio of the intensities of the two fluorochromes is computed and a CGH profile is constituted for each chromosome when the $\log_2$ of fluorescence ratios are ranked and plotted according to the physical position of their corresponding BACs on the genome [9]. This process is depicted in figure 1. Different methods have been proposed for the visualization of array CGH data [10,11].

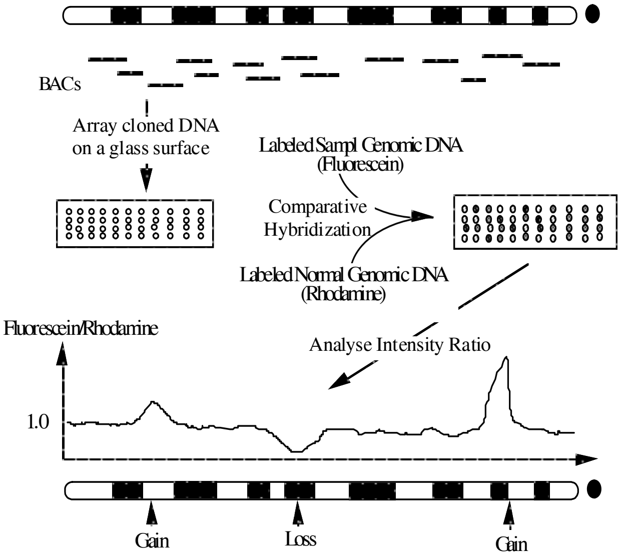

Figure 1. Schematic of array CGH methodology.

Each profile can be viewed as a succession of clusters that represent homogeneous regions in the genome whose BACs share the same relative copy number on average. Array CGH data are normalized with a median set to $\log_2(\text{ratio}) = 0$ for regions of no change, segments with positive means

Dr. Shishir Shah is with the University of Houston, Department of Computer Science, Houston, TX 77204 USA (phone: 713-743-3360; fax: 713-743-3335; e-mail: shah@ cs.uh.edu).

1-4244-0033-3/06/$20.00 ©2006 IEEE.

represent gained regions in the test sample genome, and segments with negative means represent deleted regions. Even if the underlying biological process is discrete (counting of relative copy numbers of DNA sequences), the signal under study is viewed as being continuous, because the quantification is based on fluorescence measurements, and because the possible values for chromosomal copy numbers in the test sample may vary considerably, especially in the case of clinical tumor samples that present mixtures of tissues of different natures.

In the existing approaches, two main statistical frameworks have been suggested for the analysis of array CGH data. The first is proposed as a segmentation problem where the task is to identify contiguous segments of biological interest across the ordered sets of clones [10,12-14]. The other approach is based on Hidden Markov Models [15] where the purpose is to cluster individual data points into a finite number of hidden groups. Segmentation methods have provided a framework to handle the spatial coherence of the data on the genome that is specific to array CGH. In this context the signal provided by array CGH data is supposed to be a realization of a Gaussian process whose parameters are affected by an unknown number of abrupt changes at unknown locations on the genome. One of the major limitations has been the ability to identify the number of segments that exist. This problem is theoretically complex, and has lead to *ad hoc* procedures [12-14]. Since the purpose of array CGH experiments is to discover biological events, the estimation of the number of segments remains central.

In this paper we present a statistical modeling approach to automated identification and characterization of copy number changes in a given sample. In order to build a system that can succeed in a realistic environment, certain simplifications and assumptions are made about the problem domain and the noise models associated with the data. The underlying basis for this approach is based on the objective of partitioning the clones on the array into sets with equal copy numbers. It is assumed that based on the biology of genomic rearrangements, gains and losses tend to occur over contiguous regions of the genome, possibly spanning entire or large areas of chromosomes, or alternatively, at focal points across the genome. The problem is posed again in the context of segmentation based on probabilistic clustering. The proposed framework follows the bottom-up approach and uses Bayesian statistics to account for uncertainties in the process. Clone ratios are used to model the class signature as a representation of the segment. The Bayesian implementation of the methodology is presented, in which each class or cluster is characterized by the probability density function of the clone ratios. Individual ratio values are used to identify segments and the distribution of the class region or individual clusters is modeled as a **mixture of Gaussians**. Assuming that each cluster or segment can be modeled by a Gaussian process is too simplistic and fails to accurately capture the noise process across various disparate sources in a microarray experiment. An adaptive Expectation-Maximization (EM) algorithm is used to find the parameters of the mixture distribution.

This paper is organized as follows: Section 2 describes the statistical framework and the modeling of the array CGH data. Section 3 describes experiments and data obtained for validation of the proposed procedure. Results of the developed methodology are presented in section 4. Finally, conclusions and a summary of this study are presented in section 5.

II. The Statistical Framework

To achieve optimum performance from any classification/clustering system, it is essential that its design exploits the specific characteristics of the data. In the case of array CGH, we exploit the properties of the data after normalization. Further, we also account for the ordering of the clones from the *pter* to the *qter* of each chromosome and as such incorporate the spatial location of the clones in the model. Given that, the simplest model would be a two-class discrimination where the class region is easily separable from rest of the classes. In realistic situations, due to the complexity of the sample, simple models would not suffice in classifying the region of interest and identifying all the segments across each chromosome. Further, due to varying fluorescence ratios and the presence of noise, the data characteristics may change drastically.

We propose to model the class signatures by using a combination of spatial and $\log_2$(ratio) of each clone. This would aid in developing a practical and realistic technique in a widely varying noise environment. Due to the complex and non-Gaussian distribution of the ratios, we model the data using a **mixture of Gaussians**. Modeling of data is an important consideration in designing statistical clustering techniques. The simplest way to model non-Gaussian data is to use the histograms of the data. However, clustering based on this method does not generalize well over a range of noise processes. The Parzen density estimate [16] is a well established method to establish density estimates for multivariate models. However, the Parzen windows approach is computationally expensive and has problems when the data is large and sparsely distributed. Maximum likelihood estimators [17] compute piecewise estimates of one-dimensional density functions. This approach can be regularized by introducing a penalty term. Such methods are attractive, but rely on a predefined model of the density function. They also do not generalize well in the case of mixture models unless coupled with other optimization techniques.

We use the Expectation-Maximization (EM) algorithm [18] to determine the parameters for the mixture of Gaussians model to estimate the density function. Considering Y to be the data, that is the $\log_2$(ratio), we pose the parameter estimation for various segments/clusters as a maximum likelihood problem. The general form of the density function for the measured feature can be given as:

$$P(Y \mid t) = P(Y \mid \theta) = \sum_{i=1}^{c} p(Y \mid t, \theta_i)\alpha_i$$

where, t is the conditioning variable (class signature), c represents the number of component density functions $p(Y \mid t, \theta_i)$ that make up the mixture, α_i represents the weight associated with each of the density functions (also called mixing parameter), and θ_i represents the parameter vectors for each component density function. θ, α, and c are unknown, and have to be estimated from the data. We assume the component densities to be normal distributed. That is $p(Y \mid t, \theta_i) \approx N(\mu_i, \sigma_i)$, and $\theta_i = (\mu_i, \sigma_i)$, where μ_i and σ_i represent the mean and variance of the normal distribution. To model each cluster, the values of μ_i, σ_i, and α_i have to be estimated. At the start of the process, the number of components densities (c), the density means (μ_i), standard deviation (σ_i), and the mixing weights (α_i) need to be computed. In doing so, we use the K-Means algorithm iteratively with the EM algorithm to determine all the parameters. A stage stagewise K-Means procedure is used, where the initial guess for the cluster centroids is obtained by splitting the centroids resulting from the previous stage. The iterative process is initialized assuming a single segment across the entire dataset. That is, given a set of features Y_m = [$Y_{m,1}$, $Y_{m,2}$, ..., $Y_{m,d}$] for $m=1,...,c$, the number of kernels or components is set to one. The centroid of all data points is computed and a measure of the mean and in-class deviation is computed as:

$$\mu = \frac{1}{M}\sum_{u=1}^{n} Y(u)$$

$$\sigma^2 = \frac{1}{M}\sum_{u=1}^{n} Y^2(u) - \mu^2$$

Now for each cluster, a normalized index is computed as:

$$I = (\frac{1}{M}\sum_{u=1}^{n}(Y(u) - \mu)) / \sigma$$

The normalized index gives a point measure of deviation from the cluster center. The component weights α_i is computed as a ratio of the number of data points in the corresponding component and the total points. Denoting Y_{ik} as the k^{th} sample belonging to a cluster I and using the EM approach, the following equations are obtained for the estimates of μ_i, σ_i, and α_i:

$$\alpha_i = \frac{1}{n}\sum_{k=1}^{n} P(t \mid Y_{ik}, \theta_i)$$

$$\mu_i = \frac{\sum_{k=1}^{n} P(t \mid Y_{ik}, \theta_i) Y_{ik}}{\sum_{k=1}^{n} P(t \mid Y_{ik}, \theta_i)}$$

$$\sigma_i^2 = \frac{\sum_{k=1}^{n} P(t \mid Y_{ik}, \theta_i)(Y_{ik} - \mu_i)^2}{\sum_{k=1}^{n} P(t \mid Y_{ik}, \theta_i)}$$

where,

$$P(t \mid Y_{ik}, \theta_i) = \frac{P(Y_{ik} \mid t, \theta_i)\alpha_i}{\sum_{q=1}^{c} P(Y_{qk} \mid t, \theta_i)\alpha_q}$$

If the normalized index is greater than a set threshold, a new mean is initialized and the nearest neighbor partition is computed. This results in the formation of a new cluster representing a new segment in the array CGH data. A new estimate of the means, variances, and the distortion are computed. These equations are iteratively solved until convergence of the parameter values is achieved. The convergence leads to the identification of optimal number of segments in the data as well as the appropriate groupings.

III. Data and Experiments

We demonstrate our approach on commercial cell lines from Vysis, Inc., as well as publicly available Coriel cell lines. In addition, we also compare our approach to the results of known karyotypes and classical CGH. In the various comparisons, we were able to analyze data having chromosomes with partial changes as well as chromosomal monosomies and trisomies.

All experiments were conducted using BAC arrays with approximately 1Mb coverage across the genome. Cyanine dyes (Cy3 and Cy5) were used to label the sample and reference genomic DNA using random priming. Protocols published by Spectral Genomics, Inc. were followed for labeling and hybridization. After hybridizing at 60°C for 12 hours, imaging was performed using the GenePix 4000A laser scanner from Axon, Inc., and the images were analyzed using the SpectralWare system (Spectral Genomics, Inc.). In the microarrays used, each BAC is spotted in duplicate. Data obtained after image analysis was globally normalized such that the summed Cy3 signal equals the summed Cy5 signal.

IV. Results

The tumor cell line MPE-600 (Vysis, Inc.), with verified cytogenetically- and CGH-detected chromosomal aberrations, was used to validate our approach. In MPE-600, at least five major chromosomal aberrations have been found by CGH, and include a deletion of 1pter, gain of 1q, loss of 9p, gain of 14p, and loss of 16q. Figure 2 shows the pseudo-color image of the BAC array, with MPE-600 labeled with Cy3 and normal genomic DNA labeled with Cy5. The green spots represent gains, the red spots represent deletions, and yellow represents no change. Figure 3a (BAC array) show the result of our procedure while figure 3b show analysis of chromosome 1 by classical CGH. The y-axis represents the fluorescent ratio compared to the control, while the x-axis shows the chromosomal position. This analysis of array CGH exquisitely reproduces and corresponds with the classical CGH data. Interestingly, the classical CGH data suggests that there might be an amplification event close to the centromere on 1q; the BAC array data clearly demonstrates this amplification event. Figure 4 shows the results obtained across all the chromosomes. We also performed analysis of the H526 cell line (lung carcinoma) and were successful in identifying all the known segments across the genome. The results were confirmed by comparison to classical CGH profiles.

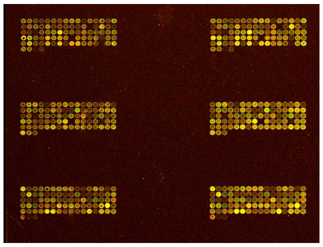

Figure 2. Subimage of MPE-600 hybridized on BAC array.

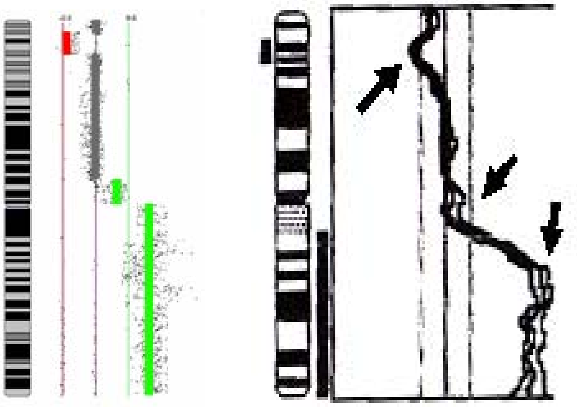

Figure 3. Left profile (3a) shows the results of identified segments from array CGH data and right profile (3b) shows the results obtained from classical CGH.

V. Summary and Conclusion

The main issue of array CGH data analysis lies in the robust estimation of the number of segments across the genome. In this paper we have presented a statistical methodology for analysis of array CGH data for automatic identification of gains and losses across the genome. This method identifies segments based on clusters, each based on a varying noise model estimated by a mixture of Gaussians. Computing the number of clusters in any model is theoretically complex and our approach converges to the final number in an iterative algorithm. Results obtained on both commercial and public cell lines is compared to classical CGH profiles and the effectiveness of our method demonstrated.

References

[1] D. Albertson, B. Ylstra, R. Segraves, C. Collins, S. Dairke, D. Kowbel, W. Kuo, J. Gray, D. Pinkel, "Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene," in *Nature Genetics*, vol. 25, pp. 144-146, 2000.

[2] M. Heiskanen, J. Kononen, M. Barlund, J. Torhorst, G. Sauter, A. Kallioniemi, O. Kallioniemi, „CGH, cDNA and tissue microarray analyses implicate FGFR2 amplification in a small subset of breast tumors," in *Anal Cell Pathol*, vol. 22, pp. 229-234, 2001

[3] J. Pollac, C. Perou, A. Alizadeh, M. Eisen, A. Pergamenschikov, C. Williams, S. Jeffrey, D. Botstein, P. Brown, "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," in *Nature Genetics*, vol. 23, pp. 41-46, 1999.

[4] S. Takeo, H. Arai, N. Kusano, T. Harada, T. Furuya, S. Kawauchi, A. Oga, T. Hirano, T. Yoshida, K. Okita, K. Sasaki, "Examination of oncogene amplification by genomic DNA microarray in hepatocellular carcinomas, comparison with comparative genomic hybridization analysis," in *Cancer Genet Cytogenet*, vol. 130, pp. 127-132, 2001.

[5] C. Browne, N. Dennis, E. Maher, F. Long, J. Nicholson, J. Sillibourne, J. Barber, "Inherited interstitial duplications of proximal 15q: genotype-phenotype correlations." in *Am J Hum Genetics*, vol. 61(6), pp. 1342-52, 1997.

[6] G. Capone, "Down syndrome: advances in molecular biology and the neurosciences," in *J Dev Behav Pediatr*, vol. 22(1), pp. 40-59, 2001.

[7] D. Slamon, G. Clark, S. Wong, W. Levin, A. Ullrich, W. McGuire, "Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene," in *Science*, vol. 235(4785), pp. 177-82, 1987.

[8] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. Kuo, C. Chen, Y. Zhai, S. Dairkee, B. Ljung, J. Gray, D. Albertson, "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," in *Nature Genetics*, vol. 20, pp.207-211, 1998.

[9] A. Snijders, N. Nowak, R. Segraves, S. Blakwood, N. Brown, J. Conroy, G. Hamilton, A. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. Yue, J. Gray, A. Jain, D. Pinkel, D. Albertson, "Assembly of microarrays for genome-wide measurement of DNA copy number," in *Nature Genetics*, vol. 29, pp. 263-264, 2001.

[10] R. Autio, S. Hautaniemi, P. Kauraniemi, O. Yli-Harja, J. Astola, M. Wolf, A. Kallioniemi, "CGH-plotter: MATLAB toolbox for CGH-data analysis," in *Bioinformatics*, vol. 13, pp. 1714-1715, 2003.

[11] P. Eilers, R. Menezes, "Quantile smoothing of array CGH data," in *Bioinformatics*, vol. 21, pp. 1146-1153, 2005.

[12] K. Jong, E. Marchiori, A. van der Vaart, B. Ylstra, M. Weiss, G. Meijer, "Applications of Evolutionary Computing," in *Proceedings of EvoWorkshops*, *Springer-Verlag Heidelberg*, vol. 2611, pp. 54-65, 2003.

[13] A. Olshen, E. Venkatraman, R. Lucito, M. Wigler, "Circular Binary segmentation for the analysis of array-based DNA copy number data," in *Biostatistics, vol.* 5(4), pp. 557-572, 2004.

[14] P. Hupe, N. Stransky, J. Thiery, F. Radvanyi, E. Barillot, "Analysis of array CGH data: from signal ratio to gain and loss of DNA regions," in *Bioinformatics*, vol. 20(18), pp. 3413-3422, 2004.

[15] J. Fridlyand, A. Snijders, D. Pinkel, D. Albertson, A. Jain, "Hidden Markov Models approach to the analysis of array CGH data," in *Journal of Multivariate Analysis*, vol. 90, pp. 132-1533, 2004.

[16] K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, New York, 1972.

[17] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," in *Journal of the Royal Statistical Society*, 39-B:1–38, 1977.