Statistical issues in the analysis of the array CGH data

Jane Fridlyand Antoine Snijders Dan Pinkel Donna Albertson
Ajay Jain
UCSF Comprehensive Cancer Center
2340 Sutter Str. N412
San Francisco CA 94143-0128
janef@cc.ucsf.edu

Abstract

The development of solid tumors is associated with acquisition of complex genetic alterations, indicating that failures in the mechanisms that maintain the integrity of the genome contribute to tumor evolution. Thus, one expects that the particular types of genomic derangement seen in tumors reflect underlying failures in maintenance of genetic stability, as well as selection for changes that provide growth advantage. In order to investigate genomic alterations we are using microarray-based comparative genomic hybridization (array CGH). The computational task is to map and characterize the number and types of copy number alterations present in the tumors, and so define copy number phenotypes as well as to associate them with known biological markers.

To utilize the spatial coherence between nearby clones, we use unsupervised Hidden Markov Models approach. The clones are partitioned into the states which represent underlying copy number of the group of clones. The method is demonstrated on the two cell line datasets with known copy number alterations for one of them. The biological conclusions drawn from the analyses are discussed.

1. Introduction

In this article we present an automated method for identifying and characterizing copy number changes in a given tumor. We distinguish 4 types of genomic changes: *transitions*, *whole chromosomal gains and losses*, *focal aberrations* and *high-level focal amplifications*. While this taxonomy is not novel, the manual process is time-consuming, prone to human error and non-reproducible.

1.1. Array CGH

Microarray-based comparative genomic hybridizations (*aCGH*) provides a means to qualitatively measure DNA copy-number aberrations and to map them directly onto genomic sequence. The arrays comprised of large-insert genomic clones such as BACs provide reliable copy number measurements on individual clones and have shown to be useful for for research and clinical applications in medical genetics and cancer. The relative copy number of these spotted DNA sequences is measured by monitoring the differential hybridization of the two samples to the sequences on the array.

2. Methods

For a given genomic profile, the goal is to partition the clones into sets with equal copy number. The biological model underlying this approach is that genomic rearrangements lead to gains or losses of sizable contiguous parts of the genome, possibly spanning entire chromosomes, or, alternatively, to focal high-level amplifications. In particular, it is desirable to make use of the physical dependence of the nearby clones.

2.1. Unsupervised HMM partitioning

The observed $\log_2$ ratio for a given clone, y, is determined by the true copy number of the clone in a tumor cell, c_T, ploidy of the sample, c_N, normal cell admixture, a_N and fraction of the tumor cells which have not acquired a given aberration, t_N. Then, proportion of the cells with a given aberration is $p_{aber} = (1 - a_N)(1 - t_N)$ and

$$y = \log_2 \frac{c_T p_{aber} + c_N(1 - p_{aber})}{c_N} + \epsilon, \epsilon \text{ is } N(0, \sigma^2).$$

The HMM approach is a natural framework for the task at hand as the hidden states represent underlying copy number

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

of the clones and there exist probabilistic transitions among different states. We fit HMM to individual chromosomes for each sample. For each chromosome we need to determine the number of states and allocate the clones to the derived states.

Algorithm:

- For $k = 1 \ldots K$ states:
  1. Specify initial state parameters (mean and variance) and state emission probabilities and the transition probabilities between the states.
  2. Fit k-state HMM
  3. Assign observations to the states
  4. Calculate penalized negative log-likelihood $\psi(k)$
- Choose the model corresponding to the k with the smallest $\psi(k)$, $k^\star$
- For models with more than one state, merge the states those median are within a $threshold$ of each other.

The parameters are the maximum size of an HMM model, the model selection criteria and the threshold for the state merge. We use $K = 5$ and AIC criterion. The threshold to merge the states is dependent on the problem at hand. We have allowed the threshold to be as low as 0.15 for the tumor data and as high as 0.35 for particularly pure cell lines.

2.2. Characterizing genomic aberrations

We characterize the genomic profiles using 4 types of genomic changes: *transitions*, *whole chromosomal gains and losses*, *focal aberrations* and *high-level focal amplifications*. Figure 1 gives examples of transition and focal aberrations.

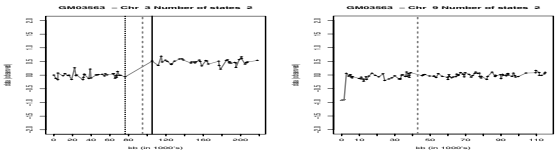

Figure 1. Example of the method application to 4 different Coriel chromosomes. The solid black line indicate the start of the region and the dotted light line shows the end of region. Light dots indicate focal aberrations. Dotted light line shows centromere.

3. Data

We demonstrate our approach on the two publicly available datasets. The first dataset was was featured in [2]. and consisted of single experiments on 15 fibroplast cell lines containing cytogenetically mapped partial or whole-chromosome aneuploidi. The second dataset was presented in [1] and consisted of aCGH profiles of 10 MMR deficient and 10 proficient cell lines. Cytogenetic analyses have shown that tumors with defects in MMR have fewer chromosomal changes than most solid tumors.

4. Results

4.1. Coriel Cell Lines

To assess the performance of our algorithm on the known aberrations of the Coriel cell lines we used the table of the agreement between known karyotypes and manual segmentation of the aCGH profiles published in [2]. There were 15 chromosomes with partial changes and 8 whole chromosomal monosomies and trisomies. We were able to detect all of the known aberrations. In addition, we have found several single clone aberrations (average of 4 per sample) which may be real or may be due to mismapped clones.

We counted the number of gains or losses of whole chromosomes, which might be expected to occur following failures of karyokinesis or cytokinesis, and the number of copy number transitions within a chromosome, which are likely to reflect DNA strand breakage that led to non-reciprocal translocations. We found that MMR deficient cells showed significantly fewer aberrations than MMR proficient cells in accord with earlier observations, although we observed a substantial number of aberrations in some MMR deficient lines. We also found a dependency of aberration type on the specific MMR defect. Cells deficient in MLH1 had a higher frequency of transitions and focal aberrations than MSH2 deficient cells.

References

[1] A. M. Snijders, J. Fridlyand, D. Mans, R. Segraves, T. Pualson, G. Wahl, A. N. Jain, D. Pinkel, and D. G. Albertson. Shaping of tumors and drug-resitant genomes by instability and selection. *Oncogene*, 2003.

[2] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics*, 29, November 2001.