Hierarchical Bayes Approach to Array CGH Data Normalization

Shigeyuki Oba
shige-o@is.naist.jp

Shin Ishii
ishii@is.naist.jp

Graduate School of Information Science, Nara Institute of Science and Technology, 8916-6
Takayama-chou, Ikoma-shi, Nara 630-0192, Japan

Keywords: array CGH, chromosomal aberration in cancer cells, type II maximum likelihood

1 Introduction

Array CGH is a high-throughput technology for measuring DNA copy number aberrations. Various types of DNA copy number aberrations, such as gains, losses, large amplifications, and so on, over various sizes of segments in chromosomes are observed from the array CGH measurements and they are thought as causes and/or consequences of cell diseases like cancers. However, the measurements based on array CGH can include various sorts of biases and variances, and detecting aberration events are not straightforward.

Copy number aberration at each segment of a chromosome is denoted as an integer number $K = -2, -1, 0, +1, +2, ...$, which is the objective of an array CGH measurement. On the other hand, the direct observation of an array CGH measurement is the real number of log-fluorescent ratio, $x$, which is similar to expression fluorescent ratio between control and objective samples in microarray measurements. We define chromosomal aberration score $Z$, which is a continuous and monotonic function of $x$ as approximating the integer $K$.

In this work, we propose a method to determine a map from $x$ to $Z$, which can correct biases and scale variability contained in array CGH measurements.

![Figure 1: Segmentation and normalization of array CGH data. Horizontal axes denote chromosomal coordinates, and vertical axes denote fluorescent ratio $x$ and corresponding aberration score $Z$ in the upper left and the lower left panels, respectively. The upper right panel shows the histogram of mean log-fluorescent ratio in each segment and the corresponding aberration scores.](image-url)
2 Method and Results

2.1 Bayesian Exact Segmentation
We developed a segmentation method for array CGH data based on an exact Bayesian estimation, which provides posterior probability of each pair of probes being a breakpoint. It also provides straightforwardly a posterior probability distribution on the number of breakpoints on a chromosome. In addition, we incorporated outliers and missing values in the estimation, which enabled robust segmentation.

2.2 “Combfit” Model and “Combfitting Algorithm”
In order to assign a score $Z$ to each of the segments, we applied a probabilistic model which we call “Combfit model” [1]. The Combfit model assumes normal noise in the log-fluorescent ratio $x$,

$$p(x | Z, b, \mu) = \text{Normal}(x; \log(bZ + 1) + \mu, \sigma^2),$$

where $b$ and $\mu$ denotes a bias and a scale of the log-fluorescent ratio in each array CGH measurement, respectively, and $\sigma^2$ denotes the variance of the normal distribution. DNA copy numbers in tumor samples may suffer from contamination of normal cells and/or heterogeneity of tumor cells, which causes variability of scales. Namely, difference magnitude of log-fluorescent ratio with respect to a single copy alteration of DNA may change according to the scale parameter, $b$.

Since we can assume integer values of copy number aberrations $Z \in \{-2, -1, 0, +1, \ldots\} = K$ for almost all of the segments, likelihood function of the pair of parameters $(b, \mu)$ is given as,

$$L(b, \mu) = \prod_i \sum_{Z=K} p(x_i | Z, b, \mu) P(Z | \text{chr}_i),$$

where $i$ is an index of each segment, $\text{chr}_i$ denotes a chromosome to which the $i$th segment belongs, and $P(Z | \text{chr}_i)$ is a prior probability that a segment in the chromosome $\text{chr}_i$ has copy number aberration $Z$. The maximum likelihood estimation of the parameters $(b, \mu)$ is equivalent to a fitting of an appropriate scale to a histogram (Fig. 1 upper right), which we call “Combfitting” algorithm. After obtaining $(b, \mu)$, a mapping from $x$ to $Z$ is straightforward.

2.3 Hierarchical Bayes Approach
The prior $P(Z | \text{chr}_i)$ in the Combfit model plays a crucial role, because, if the prior is uniform, an arbitrary shift of the bias $\mu$ corresponding to an integer number of DNA copies will not lead to any difference of the likelihood.

A simple prior belief is that there is no aberration in many of the segments, and that there are seldom segments with two copies losses because two copies loss in a diploid chromosome means complete loss of corresponding genes, which may cause a cell death. Such a belief can construct initial setting of the prior, which can be a constraint to the Combfitting result. However, it may not lead objective enough results and cannot incorporate specific knowledge on a specific cancer.

Type II maximum likelihood estimation based on a set of samples in a same group of cancers can deal with this problem, which leads to a cancer specific knowledge and possibly better normalization of array CGH data.

References

[1] Oba, S., Tomioka, N., Ohira, M., and Ishii, S.,
Combfit: A normalization method for array CGH data,
IPSJ Transactions on Bioinformatics, (to appear).