2007－BIO－9 (2)
2007／6／14

アレイ比較ゲノムデータ正規化手法 Combfit について

大羽 成征 [1], 富岡 伸元 [2], 大平 美紀 [3], 石井 信 [1]

[1] 奈良先端科学技術大学院大学 情報科学研究科, [2] 北海道大学医学部第一外科, [3] 千葉県がんセンター研究局生化学研究部

概要: マイクロアレイ技術に基く染色体比較の方法としてアレイ CGH 法と呼ばれる手法が近年開発され、がんなどの細胞疾患の原因もしくは結果と目される染色体数変動を調べるのに使われている。通常のアレイ CGH データ解析では、DNA 断片毎の複製数変動を「変動なし／増幅あり／減少あり」の 3 カテゴリーに分類しているが、近年の計測精度の改善によって、複製数変動に関するさらに定量的な解析が可能となってきた。そこで、我々はアレイ CGH データの定量的解釈のための正規化手法として combfit 法を提案する。また、DNA 断片複製数に関して局所的な多様性を含む場合について combfit 法を適用する改善案も提案する。combfit 法によれば、アレイ CGH 法によって計測された蛍光比データのベース値を補正しつつ、染色体断片毎の複製数を $0, \pm 1, \pm 2, \cdots$ のように定量的に推定することができるようになる。また combfit 法は、正常細胞の混入が多い場合や倍数性に関する多様性がある場合も考慮しており適用可能である。

Combfit: A normalization Method for Array CGH data

Shigeyuki Oba [1], Nobumoto Tomioka [2], Miki Ohira [3], and Shin Ishii [1]

[1] Graduate School of Information Science, Nara Institute of Science and Technology, [2] 1st Department of Surgery, Hokkaido University, School of Medicine [3] Division of Biochemistry, Chiba Cancer Center Research Institute,

Abstract: The recently developed array-based comparative genomic hybridization (array CGH) technique measures DNA copy number aberrations which occur as causes or consequences of cell diseases such as cancers. Conventional array CGH analysis classifies DNA copy number aberrations into three categories: no significant change, significant gain, and significant loss. However, recent improvements in microarray measurement precision enable more quantitative analysis of copy number aberrations.
We propose a method, called comb fitting, which extracts quantitative interpretation from array CGH data. We also propose modifications which allow us to apply comb fitting to cases which include heterogeneity of local aberrations in DNA copy numbers. Using comb fitting, we can correct the baseline of the fluorescence ratio data measured by array CGH, and simultaneously, we can translate them into the change amount of copy numbers of each small part of chromosome, such as $0, \pm 1, \pm 2, \cdots$. Comb fitting is applicable even when a considerable amount of contamination by normal cells exists, and when heterogeneity in ploidy number cannot be neglected.

1 Introduction

The recently developed array-based comparative genomic hybridization (array CGH) technique measures DNA copy number aberrations which occur as causes or consequences of cell diseases such as cancers [11]. The segmentation structure of chromosomal aberrations is of major interest, because segmental gains or losses often cause or reflect cell diseases. Fridlyand *et al.* (2004)[4] assumed that measured copy number aberrations could be generated by a hidden Markov model (HMM) with latent segmentation structures, which was estimated by a forward-backward algorithm. Daruwala *et al.* (2004)[2] proposed a similar model to the HMM and calculated the optimum segmentation structure by a dynamic programming algorithm. There are many other researches on segmentation problem such as [10, 7, 6, 9]. Such a sequential segmental structures are used also for noise reduction, and there are many other approaches than segmentation such as simple moving average [1], penalized quantile smoothing [3], wavelet filter [5], and so on. For segmentation and/or noise reduction, there are also researches which compares some alternetive methods [8, 12].

Assigning an appropriate copy number of DNA is the next important issue, because what we can directly observe is fluorescent level of each spot corresponding to each BAC (bacterial artificial chro-

mosome) clone which is complementary to the objective piece of the sample DNA, and the fluorescent level inevitably includes biases and/or variances come from various causes. In many previous studies, copy number aberrations of DNA were classified into categories of no significant change, significant loss, significant gain, and sometimes large amplification. In this study, we present a method that extracts quantitative interpretation from array CGH data, called comb fitting. Using this method, we can correct the baseline of fluorescence ratio data for each clone of each sample measured by array CGH, and simultaneously, transform the data into numerical copy number changes of clones, such as $0, \pm 1, \pm 2, \cdots$. Consequently, it improves analysis of phenomena occurring on chromosomes.

1.1 Chromosomal aberration

Each chromosome in somatic cells normally has two DNA copies, while chromosomal aberration in cancer cells sometimes causes aneuploidy, i.e., the total copy number becomes three, four, five or more; they are called triploid, tetraploid, pentaploid, and so on. These ploidy numbers are described as $N_{\mathrm{P}} = 2, 3, 4, 5, \cdots$.

Copy number aberrations of various other types are also known:

- Copy number gains in the whole or a part of a chromosome, which correspond to several or more BAC clones. We call these, +1 gain, +2 gain, and the like.
- Copy number losses in the whole or a part of a chromosome, which correspond to several or more BAC clones. We call these, +1 loss, +2 loss, and the like.
- Copy number gain whose amount is usually larger than ten, in a small part of a chromosome; we call this an amplification.

Each of these events is generally called a local aberration, which denotes the number of local gains or losses of copies, and is expressed by N_{C}.

First, all cancer cells in an objective sample are assumed to have homogeneous genetic aberrations. N_{ij} denotes a copy number of the ith piece of chromosome corresponding to the ith BAC clone in the jth sample; in the following, we call it simply a copy number of the ith clone. The copy number is an integer, $N_{ij} \in \{0, 1, 2, \cdots\}$, and is the sum of the ploidy number and the local aberration, $N_{ij} = N_{\mathrm{P}j} + N_{\mathrm{C}ij}$. In the usual array CGH analysis, we are interested in local aberration, N_{C}, rather than N_{P}, because N_{P} is observed by other conventional methods, and the loci and amounts of local aberrations are believed to have important information about the characteristic of the cancer. The actual measurement involves some degree of noise; thus, obtaining the expected local aberration, Z, which is called mean local aberration, is the major aim of our method.

Next, we consider heterogeneity of the cell characters in an objective sample, which are due to various reasons as listed below.

- (a) Contamination by normal cells
- (b) Presence of multiple types of chromosomal aberration
  - 1. heterogeneity in ploidy number
  - 2. heterogeneity in local aberrations

This heterogeneity presents many difficulties for quantitative analyses of chromosomal aberrations.

If the amount of contamination by normal cells is known, in case (a), then we can correct these effects using the fact that $N_{\mathrm{P}} = 2$ and $N_{\mathrm{C}} = 0$ in the normal cells. When we do not know the amount of contamination, however, the correction is based on an estimation; the estimation is also available in our framework (see section 2.3).

In case (b).1, we need another correction which depends on the mixing ratios of ploidy numbers, but this can be made similarly to case (a) (see section 2.3).

In case (b).2, Z is regarded as the mean of N_{C} over the cells in a sample; we call this mean local aberration. For example, when a tumor sample consists of the same amount of cells with $N_{\mathrm{C}} = -1$ and $N_{\mathrm{C}} = -2$, the mean local aberration becomes -1.5. Our method intends to obtain the real number Z_{ij}, rather than the integer $N_{\mathrm{C}ij}$.

When sample cells are sufficiently homogeneous with respect to N_{C} values, the true mean local aberration Z is an integer for almost every BAC clone. Consequently, when Z is estimated as, for example, $Z = 1.1$, we consider that the real Z is $+1.0$, but includes a noise contribution of 0.1. Note, however, that we cannot eliminate from this observation the possibility of a mixture containing 90% $N_{\mathrm{C}} = 1$ and 10% $N_{\mathrm{C}} = 2$.

1.2 Array CGH measurement

Microarray technology measures fluorescence levels of CY3 ($\mathrm{CY3}_{ij}$), and CY5 ($\mathrm{CY5}_{ij}$), which correspond to the copy numbers of objective and control samples, respectively, at the ith BAC clone in the jth objective sample. The log fluorescence ratio x_{ij} is calculated as

$$x_{ij} = \log_2(\mathrm{CY3}_{ij}/\mathrm{CY5}_{ij}). \qquad (1)$$

Note that equation (1) is conceptual but not necessarily precise, because it always requires correction. Various correction methods are available for various artifacts involved in microarray measurements. The most popular one is to introduce the correction term:

$$f_j(\log_2 \mathrm{CY3}_{ij} + \log_2 \mathrm{CY5}_{ij}), \tag{2}$$

which is a function of total fluorescence intensity.

The objective of our method is to estimate mean local aberration Z_{ij} from the observed log fluorescence ratio x_{ij}. First, we consider the simplest model which assumes homogeneous sample cells. Next, we show some modifications for dealing with various heterogeneities in sample cells.

2 Comb model

2.1 The simplest comb model

Provided that every objective sample consists of homogeneous cells, let N_{ij} and N_0 denote DNA copy number of a clone i in an objective sample j and that of a control sample, respectively. As the control sample, we prepared normal cells, which are all diploid DNA, i.e., $N_0 = 2$.

We obtain the log fluorescence ratio x in a similar fashion to that in the gene expression measurement:

$$\begin{aligned} x_{ij} &= \log_2 \frac{c_j N_{ij}}{c_0 N_0} + \nu_i + \mu'_j + \epsilon_{ij} \\ &= \log_2(N_{\mathrm{P}j} + N_{\mathrm{C}ij}) + \nu_i \\ &+ \mu'_j + \log_2(c_j/c_0 N_0) + \epsilon_{ij}, \\ &= \log_2(\frac{1}{N_{\mathrm{P}j}} N_{\mathrm{C}ij} + 1) + \nu_i + \mu_j + \epsilon_{ij}, \end{aligned} \tag{3}$$

where ν_i and μ'_j are constant biases which denote mean log fluorescence ratios of the BAC clone i and the sample j, respectively. c_j and c_0 are constant factors proportional to the number of cells in corresponding sample, hybridization efficiency, and so on, corresponding to the j-th objective sample and the control sample, respectively. The unknown factors, c_j, c_0, are united to a single bias term, $\mu_j \equiv \mu'_j + \log_2(c_j N_{\mathrm{P}}/c_0 N_0)$. ϵ denotes a residual component which is assumed to obey a normal distribution with mean 0 and variance σ^2. The variance σ^2 corresponds to observation error and its value is assumed to be known before our analysis.

In the following discussion, we ignore the clone-wise bias ν_i, because it is corrected by some common methods, such as equation (2). We also omit j, because our method deals with individual samples. Consequently, equation (3) becomes simply:

$$x_i = \log_2(\frac{1}{N_{\mathrm{P}}} N_{\mathrm{C}i} + 1) + \mu + \epsilon_i. \tag{4}$$

Thus, when the variance σ^2 is small enough, the expected local aberration i can be approximately obtained as:

$$Z_i = E[N_{\mathrm{C}i}] \approx N_{\mathrm{P}}(2^{x_i - \mu} - 1). \tag{5}$$

The unknown bias μ cannot be ignored, because it concerns many aspects of variability in microarray slides, such as inequality between the total amounts of DNA in objective and control samples, asymmetry in the fluorescence of CY3 and CY5, and so on.

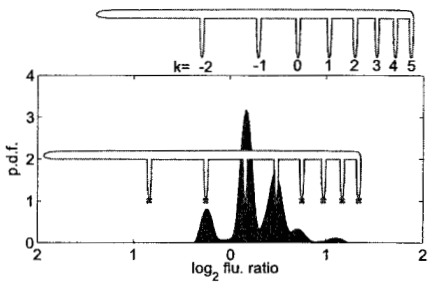

Figure 1: **Conceptual diagram of comb fitting.** The comb, whose teeth are aligned at certain intervals, is fitted into the distribution of log fluorescence ratios. The mean local aberration Z is then obtained.

2.2 Comb fitting based on the simplest comb model

Given the observed fluorescence ratio $X = (x_1, \cdots, x_i, \cdots)$ for a sample, the likelihood of the unknown parameter $\theta = \{\mu\}$ is defined as

$$L(\theta|X) = \prod_i \prod_{k \in K} p(x_i|N_{\mathrm{C}i} = k, \mu) P(N_{\mathrm{C}i} = k), \tag{6}$$

$$p(x_i|N_{\mathrm{C}i} = k, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x_i - m_k)^2], \tag{7}$$

where we assume that the residual ϵ_i is an independent normal noise with variance σ^2, θ represents the unknown parameter μ, $K = \{-N_{\mathrm{P}}, \cdots, -1, 0, 1, 2, \cdots\}$ is a set of possible values of local aberrations, and m_k denotes the Gaussian center defined by

$$m_k = \log_2(\frac{1}{N_{\mathrm{P}}} k + 1) + \mu. \tag{8}$$

$p(N_{Ci} = k)$ is the *a priori* probability, in short 'prior', of local aberration being k at the ith clone. In the simplest method, it is set to be equal for all k, but we will discuss some advanced ways to incorporate *a priori* knowledge in section 4.2.

We can determine μ to maximize the likelihood $L(\theta|X)$. The mixture of normal distributions, each of whose Gaussian centers m_k, is aligned in a predetermined manner, and shifted by a single location parameter μ. In the following, we call this mixture model a comb model, and call m_k the kth comb tooth. This maximum likelihood estimation corresponds to fitting the comb model into the data distribution by shifting the location μ of the comb. Figure 1 shows the concept of this comb fitting.

Since the comb teeth, $\log_2(N_P+k)+\mu$, have non-equal intervals, there is a single location at which the set of comb teeth fits best into an ideal data set which obeys a normal mixture distribution with Gaussian centers, which have non-equal intervals, and a homogeneous Gaussian variance σ^2.

2.3 Contamination by normal cells

Let b denote the contamination rate of normal somatic cells whose copy number is $N_0 = 2$ at all clones, and assume that the ploidy number N_P of the tumor cells in the objective sample is homogeneous. When the contamination level is not negligible, the fluorescence ratio x_{ij} becomes

$$x_{ij} = \log_2 \frac{c_j}{c_0} \frac{b_j N_0 + (1-b_j)N_{ij}}{N_0} + \mu'_j + \nu_i + \epsilon_{ij} \tag{9}$$

which leads to a modification of the comb teeth into

$$\begin{aligned} m_k &= \log_2\{Bk+1\} + \mu, \\ B_j &= \Big(\frac{b_j}{1-b_j} N_0 + N_{Pj}\Big)^{-1}. \end{aligned} \tag{10}$$

We can easily find that, when $b_j = 0$ i.e. $B_j = {N_P}_j^{-1}$, it is equivalent to the most simple case (3). And, consider the extreme case of high level of contamination, $b_j \to 1$ i.e. $B_j \to 0$, in this case, the comb teeth become insensitive to the local aberration k.

The two unknown parameters b_j and N_{Pj} are united into B_j and if B_j is obtained the expected local aberration is obtained as

$$Z_i = E[N_{Ci}] \approx B_j^{-1}(2^{x_i-\mu} - 1). \tag{11}$$

Thus, we estimate $\theta = \{B, \mu\}$ by the maximum likelihood estimation, instead of considering b and N_P.

2.4 Heterogeneity in ploidy number

Assume that there is considerable heterogeneity in ploidy number N_P of objective tumor cells; the ratios of $N_P = 2$, $N_P = 3, \cdots$ are given by $\beta^{(2)}$, $\beta^{(3)}, \cdots$ respectively, with the condition $\sum_{N_P} \beta^{(N_P)} = 1$. We assume the contamination ratio b of normal cells is also considerable, but the local aberration N_C is the same for all cells regardless of their ploidy number.

In this case, comb teeth takes also just the same form as eq.(10) except that the definition of the parameter B_j is

$$B = \Big(\frac{b}{1-b} N_0 + \sum_{N_P} \beta^{(N_P)} N_P\Big)^{-1}. \tag{12}$$

Consequently, comb fitting needs only two parameters, μ and B, even when there is heterogeneity in the ploidy number N_P.

Even if there is considerable heterogeneity in local aberrations for some clones, it is negligible when its amount is small relative to the number of total clones in the sample.

Accordingly, the problem becomes to obtain only two parameters, μ and B, in all of the above-mentioned cases. The most ideal case, which is considered first, is equivalent to assuming $B = 0$ in equation (10). How to determine the parameters will be described in the next section.

3 Maximum likelihood estimation and its problems

Maximum likelihood estimation estimates $\theta = \{\mu, B\}$ to maximize the log likelihood function $L(\theta|X)$.

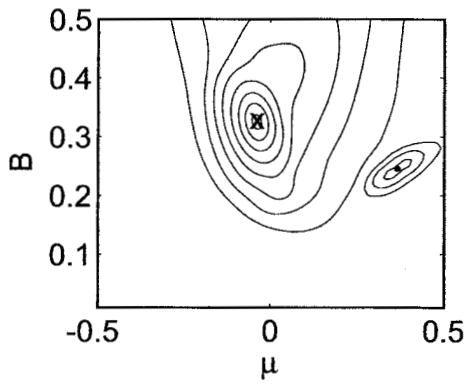

Figure 2: **Contour plot of log-likelihood.** The mark 'x' in the contour denotes the maximum likelihood solution.

Figure 2 shows a contour plot of the functional relationship of the log likelihood with the parameters μ and B, when applied to typical array CGH profile data. The mark 'x' in the contour denotes the maximum likelihood solution, $(\mu_{\mathrm{MAX}}, B_{\mathrm{MAX}})$. Obvious multi-modality can be seen in the log likelihood landscape; such multi-modality is often observed, especially when the heterogeneity is low. To obtain the maximum likelihood solution, we used a mesh search method which searches the mesh over the space of μ and B for the maximum point, because simple gradient-based methods fail to obtain the optimal solution due to the multi-modality.

There are two reasons why such multi-modality appears. The first is that a stepping-stone-like alternative, whose comb teeth correspond to the first, third and fifth teeth of the optimal comb for example, may have comparable likelihood. The second is that as B becomes small the intervals between teeth cannot be distinguished, which makes the likelihood invariant with respect to the shift of the comb position.

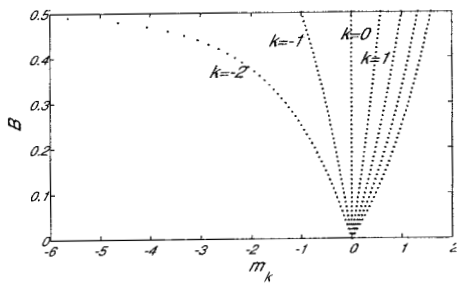

Figure 3: **Comb tooth when B is small**

Figure 3 shows the relationship between B values and comb tooth intervals. For a small B, tooth intervals are narrow and the difference between neighboring intervals becomes small. When the neighboring intervals are similar, unit shift of the comb does not lead to significant difference in fitting performance, which provides the solution ambiguity. There is another difficulty. Because the residual variance σ^2 is fixed, narrower tooth intervals lead to larger overlap between two adjacent Gaussian distributions, which makes it difficult to distinguish them. Accordingly, when B approaches 0, the simple maximum likelihood method to perform comb fitting becomes difficult.

A small B means a high contamination rate. To overcome the difficulties due to high contamination rate, we propose several devices, such as designing data preprocessing, introducing *a priori* knowledge, or brute force by hand-tuning. We explain these devises in the next section.

4 Modifications of comb fitting

4.1 Preprocessing

We applied spatial filtering on each chromosome as a data preprocessing devise. Two types of one-dimensional spatial filter, lowess and block filter, were tried. The lowess filter is based on the assumption that log fluorescence ratio varies continuously along the locus coordinate in a single chromosome. The block filter assumes three block regions in a chromosome, where the copy number is assumed to be identical in a single block. Although there are many segmental or other filtering procedures [4, 2, 10, 7, 6, 9, 1, 3, 5], we did not tried all of them because we only need filtering process in order to obtain a better global profile of histogram of log fluorescence ratio, and the various filtering did not results large difference in the histogram.

The upper left panel of Figure 1 shows fluorescence ratios of clones aligned along the locus coordinate. The log fluorescence ratios of the majority of clones in a chromosome have the same value. In each of chromosomes 1, 6 and 17, one chromosomal region shows gain or loss of a unit copy number, and the residuals are considered as noise. Figures

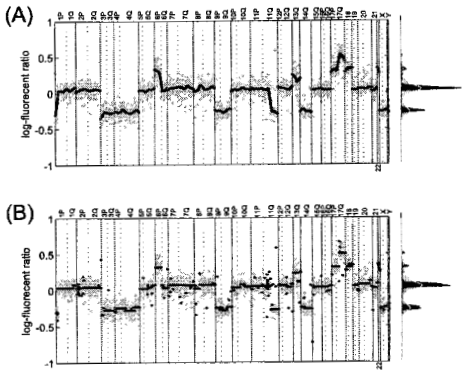

Figure 4: **Smoothing filters as data preprocessing.** Gray and black spots (histograms) denote original and filtered data (histograms), respectively.

4(A) and 4(B) show the results after applying the lowess filter and the block filter, respectively. With either filter, the comb tooth structure was clarified, as can be seen in the histogram shown to the right of the corresponding panel. Note that the final results of the comb fitting are comparable whichever

filter was used because the two histograms are similar.

DNA copy number aberrations sometimes seem to have a block structure, i.e., a certain region of a P-arm or Q-arm exhibits fixed copy number gain or loss. To obtain such a block structure, the block filter splits the chromosome into three blocks, each with a unique copy number. For determining break points for the three blocks, we define the distortion measure based on the comb:

$$D_r = \sum_{i \in \text{chromosome } r} (x_i - m(i))^2, \qquad (13)$$

where r denotes chromosome index, and $m(i)$ denotes mean log fluorescence ratio of the block to which the ith clone belongs. Block filtering determines two break points for each chromosome r to minimize D_r. Pre-defined outlier clones are omitted from the filtering. Figure 4(B) shows an example result of the block filtering.

4.2 Using *a priori* knowledge

If we have biological knowledge about copy number aberrations, it can be used as *a priori* knowledge for improving the estimation made by comb fitting.

Most local aberrations of chromosomes are at most a single gain or loss. If comb fitting suggests that a large area of a chromosome in diploid sample cells has lost its two copies, this result is not natural, because such a loss would cause serious damage even to tumor cells. Because our comb model is formulated as a probabilistic model, such *a priori* knowledge can be incorporated as the prior distribution. The prior $p(N_{\mathrm{C}i} = k)$ represents the probability that a local aberration of the ith clone is k. There are three possible ways of preparing the prior: subjective tuning, empirical tuning and recursive tuning.

Subjective tuning is based on subjective belief about the frequency of copy number aberrations. As a standard setting for diploid tumor cells, we used the following prior:

$$\begin{aligned} p(N_{\mathrm{C}} = -2) &= \varepsilon C \\ p(N_{\mathrm{C}} = -1) &= 5C \\ p(N_{\mathrm{C}} = 0) &= 10C \\ p(N_{\mathrm{C}} = 1) &= 2C \\ p(N_{\mathrm{C}} = 2) &= C \\ p(N_{\mathrm{C}} = 3) &= C, \end{aligned}$$

where ε is a small number ($= 0.01$) so as to represent the rareness of the event $N_{\mathrm{C}} = -2$; however, $\varepsilon = 0$ incurs too large a penalty in the case that Z becomes -2, possibly due to occasional noise. C is set from the normalization condition $\sum_{k=-2}^{3} p(N_{\mathrm{C}} = k) = 1$.

When much information is available about occurrence rates of local aberrations, empirical tuning is advantageous. Namely, we set the prior probability $p(N_{\mathrm{C}} = k)$ at the empirical ratio of copy numbers $N_{\mathrm{C}} = -2, -1, 0, 1, 2, 3, \cdots$ directly. Note that setting $p(N_{\mathrm{C}} = k) = \varepsilon > 0$ will be better even when $N_{\mathrm{C}} = k$ has not occurred empirically.

We may consider recursive tuning of prior, when we have insufficient background information but have a fairly large amount of array CGH data. Namely, frequencies of copy number aberrations estimated using a subjectively tuned prior are used as new background information for the next empirical tuning.

When it is known, dependence on ploidy and/or chromosome numbers should be used in each tuning method.

4.3 Hand tuning

Either when B is close to 0 or when there is significant heterogeneity of local aberrations in sample cells, our comb model with the above-mentioned modifications has difficulty in obtaining an appropriate solution. In such a case, one possible way is hand tuning. In hand tuning, we need to set at most three pairs of reference values to determine the two unknown parameters μ and B unequivocally.

For example, if we set log fluorescence ratios $x^{(-1)}, x^{(0)}, x^{(1)}$ to $Z = -1, 0, 1$, respectively, μ and B are unequivocally determined and hence the rest $x^{(-2)}, x^{(2)}, x^{(3)}, \cdots$ corresponding to $Z = -2, 2, 3, \cdots$ are determined automatically.

5 A case study

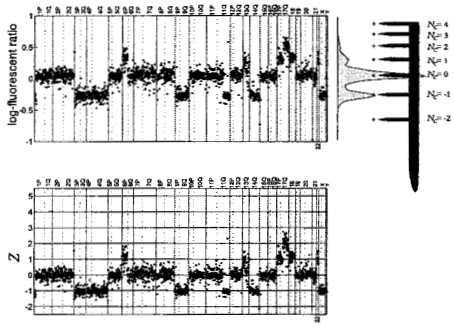

Figure 5: **Demonstration of comb fitting.** Demonstration of comb fitting for a sample which has large and complex chromosomal aberrations.

Figure 5 shows a demonstrative analysis of array CGH observation data obtained from a frozen sample of human neuroblastoma.

In neuroblastoma, it is known that chromosomal aberrations in the first and seventeenth chromosomes have a high correlation with the patient's prognosis. We conducted fluorescence *in situ* hybridization (FISH) observations on these two chromosomes and found the following points.

- The sample cells have three or four copies of both of the first and seventeenth chromosomes.
- The copy number ratio between the Q-arm and the centromere of the seventeenth chromosome, 17q/17cen, is 8/3 or 9/4.

According to these observations, we conclude that this sample has heterogeneity in the ploidy number, which lies between triploid and tetraploid, and the local aberration of 17q is +5 gain.

In Figure 5(a), a gray point denotes original log fluorescence ratio observed at each clone, and a black point denotes the corrected value after block filtering. Figure 5(b) shows the histograms of the original and the filtered log ratios, which are depicted by gray and black colors, respectively. Although the histogram of the original ratios shows a single large peak, that of the filtered ratios shows clear multi-modality which seems to fit the comb model. Accordingly, we applied comb fitting to these block-filtered data. We set the constant σ^2, the variance of each single comb tooth, at the mean variance over blocks extracted by the block filtering; namely, it is set at the mean squared residual of the block filtering divided by the mean number of clones within a block. Although it is possible in principle to estimate σ^2 as another parameter of the likelihood, we regarded it as a constant, because increasing the number of parameters makes it difficult to estimate them against the multi-modality of the likelihood function. As the prior, we used a uniform distribution to reflect the lack of background knowledge.

The results of the comb fitting are shown by '+' marks in Fig. 5(b) which denote comb teeth fitted into the histogram, and Fig. 5(c) shows the mean local aberration values, Z, obtained by comb fitting. The Z value for each clone and its block-wise mean are plotted as gray and black spots, respectively. We can see that the Z values, especially for the block-wise means, densely cluster around integer numbers.

The Z values at the Q-arm of the seventeenth chromosome, 17q, cluster around +6 gain, which differs by one copy from the FISH observation (+5 gain). Our result is consistent with the FISH data, however, if we assume alternatively the baseline (0) of the Z values corresponds to single copy loss. Actually, the second peak of the log likelihood of the comb model corresponded to the alternative solution. Since the difference between their peak heights is small, we may probably obtain the second peak as the best solution if we use appropriate *a priori* knowledge.

We found in Figures 5(a) and 5(c) that log fluorescence ratios and mean local aberrations of clones have noisy distribution centered at block-wise filtered values. Concerning variances of residuals, those in log fluorescence ratio are almost homogeneous at any location (Fig. 5(a)), while those in mean local aberration are large because the mean local aberration itself is large (Fig. 5(c)). This is because the comb tooth intervals are narrow when the mean local aberration is large, and hence expansion from the log fluorescence measurement to the mean local aberration becomes large. Therefore, a large portion of residual variation in copy number aberrations is regarded as due to observation noise in the log fluorescence ratio, rather than as an outcome of local genetic copy number aberration such as homozygous gains or losses.

6 Discussion

Appropriate application of comb fitting often requires subjective setting of parameters based on *a priori* knowledge. Although this may seem, at first glance, to violate the objectivity of data analysis, it is objective enough because the *a priori* knowledge must be expressed explicitly as the prior distribution to meet the probabilistic estimation process with the comb model.

From the Bayesian point of view, data analysis of all sorts includes inevitable bias from a researcher's subjective *a priori* beliefs about the analysis targets. Therefore, what is important for sound data analysis is to explicitly express *a priori* knowledge in the form of prior probability. The subjective tuning and hand tuning of the comb fitting, discussed in section 4.3, were based on such an idea.

When the prior probability is available, we can update it by using *a posteriori* knowledge obtained from the observed data, which enhances the objectivity of the analysis. The empirical tuning and recursive tuning of the prior, discussed in section 4.2, were based on this idea.

7 Conclusion

We have developed a method called comb fitting, which determines the copy number of DNA corre-

sponding to each BAC clone from each fluorescence ratio measured by array CGH.

Automatic comb fitting is available even when there is considerable contamination by normal cells, or when tumor cells have heterogeneous ploidy numbers. We also proposed modifications using *a priori* knowledge and/or hand tuning, which help comb fitting when automatic fitting is hard to apply, as in cases where large contamination of normal cells or large heterogeneity in local aberrations exists.

References

[1] Carvalho, B., Ouwerkerk, E., Meijer, G. A. and Ylstra, B.: High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides, *J Clin Pathol*, Vol. 57, No. 6, pp. 644–646 (2004). Evaluation Studies.

[2] Daruwala, R. S., Rudra, A., Ostrer, H., Lucito, R., Wigler, M. and Mishra, B.: A versatile statistical analysis algorithm to detect genome copy number variation, *Proc. Natl. Acad. Sci. USA*, Vol. 101, No. 46, pp. 16292–16297 (2004).

[3] Eilers, P. H. C. and de Menezes, R. X.: Quantile smoothing of array CGH data, *Bioinformatics*, Vol. 21, No. 7, pp. 1146–1153 (2005). Evaluation Studies.

[4] Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. and Jain, A. N.: Hidden Markov models approach to the analysis of array CGH data, *Journal of Multivariate Analysis*, Vol. 90, No. 1, pp. 132–153 (2004).

[5] Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L. and Porter, P.: Denoising array-based comparative genomic hybridization data using wavelets, *Biostatistics*, Vol. 6, No. 2, pp. 211–226 (2005).

[6] Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F. and Barillot, E.: Analysis of array CGH data: from signal ratio to gain and loss of DNA regions, *Bioinformatics*, Vol. 20, No. 18, pp. 3413–3422 (2004). Evaluation Studies.

[7] Jong, K., Marchiori, E., Meijer, G., Vaart, A. V. D. and Ylstra, B.: Breakpoint identification and smoothing of array comparative genomic hybridization data, *Bioinformatics*, Vol. 20, No. 18, pp. 3636–3637 (2004). Evaluation Studies.

[8] Lai, W. R., Johnson, M. D., Kucherlapati, R. and Park, P. J.: Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data, *Bioinformatics*, Vol. 21, No. 19, pp. 3763–3770 (2005).

[9] Myers, C. L., Dunham, M. J., Kung, S. Y. and Troyanskaya, O. G.: Accurate detection of aneuploidies in array CGH and gene expression microarray data, *Bioinformatics*, Vol. 20, No. 18, pp. 3533–3543 (2004). Evaluation Studies.

[10] Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.-J.: A statistical approach for array CGH data analysis, *BMC Bioinformatics*, Vol. 6, No. 1, p. 27 (2005).

[11] Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B.-M. and Gray, J. W.: High resolution analysis of DNA copy number variation using comparative geneomic hybridization to microarrays, *Nature Genetics*, Vol. 20, pp. 207–211 (1998).

[12] Willenbrock, H. and Fridlyand, J.: A comparison study: applying segmentation to array CGH data for downstream analyses, *Bioinformatics*, Vol. 21, No. 22, pp. 4084–4091 (2005).

Acknowledgment

The authors are grateful to Dr. Burt G Feuerstein, Dr. Donna Albertson and Dr. Dan Pinkel (Cancer Center and Department of Laboratory Medicine, University of California) for array CGH data acquisition, to Dr. Jane Fridlyand (Cancer Center and Department of Laboratory Medicine, University of California) for fluorescence data prepocessing, to Dr. Yasuhiko Kaneko (Division of Chemotherapy, Saitama Cancer Center Research Institute) for FISH analysis of a demonstration sample and to Dr. Akira Nakagawara (Chiba Cancer Center Research Institute) for discussions on characteristics of neuroblastoma samples.