Published in final edited form as:

J Bioinform Comput Biol. 2011 February ; 9(1): 131–148.

A NOVEL APPROACH TO DNA COPY NUMBER DATA SEGMENTATION

Siling Wang^{*,‡}, Yuhang Wang^{*,§}, Yang Xie^{†,¶}, and Guanghua Xiao^{†,||}

^{*}Department of Computer Science and Engineering, Southern Methodist University Dallas, Texas 75205, USA

[†]Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

Abstract

DNA copy number (DCN) is the number of copies of DNA at a region of a genome. The alterations of DCN are highly associated with the development of different tumors. Recently, microarray technologies are being employed to detect DCN changes at many loci at the same time in tumor samples. The resulting DCN data are often very noisy, and the tumor sample is often contaminated by normal cells. The goal of computational analysis of array-based DCN data is to infer the underlying DCNs from raw DCN data. Previous methods for this task do not model the tumor/normal cell mixture ratio explicitly and they cannot output segments with DCN annotations.

We developed a novel model-based method using the minimum description length (MDL) principle for DCN data segmentation. Our new method can output underlying DCN for each chromosomal segment, and at the same time, infer the underlying tumor proportion in the test samples. Empirical results show that our method achieves better accuracies on average as compared to three previous methods, namely Circular Binary Segmentation, Hidden Markov Model and Ultrasome.

Keywords

MDL; model-based; DCN data; segmentation

1. Introduction

DNA copy number (DCN) is the number of copies of DNA at a region of a genome. The alterations of DCN are highly associated with the development of different tumor chromosomal segments which can lead to cancer development.¹ For example, deletions may inactivate tumor suppressor genes, and amplifications may activate oncogenes in the genome. Both deletions and amplifications change the DNA copy numbers.

Recently, microarray technologies are being employed to detect DCN changes at many loci at the same time in tumor samples. They yield data consisting of log₂-transformed fluorescence intensity ratios of the two DNA samples (e.g. tumor and reference normal

© Imperial College Press

Correspondence to: Yuhang Wang.

[‡]silingw@smu.edu

[§]yuhangw@lyle.smu.edu

[¶]Yang.Xie@utsouthwestern.edu

^{||}Guanghua.Xiao@utsouthwestern.edu

DNA samples). The intensity ratios provide information about DNA copy number aberrations. The goal of analysis of array-based DCN data is to find chromosomal segments with DCN aberrations from raw DCN data. In this DCN data analysis problem, there are two important issues:

- The DCN data is often very noisy. This situation is particularly severe for high-density arrays with short probes (cDNA or oligonucleotides).²
- In practice, the tumor samples are often a mixture of tumor cells and normal cells.³ The result on the DCN data is that the copy number aberration signals are diluted. This could cause small copy number changes (in terms of amplitude or segment length or both) to be missed in the subsequent data analysis.

Several methods have been proposed for the DNA copy number data analysis problem in the last few years. There are basically two different types of approaches. The first type includes denoising and smoothing approaches,^{4–6} which try to reduce the noise in the data. The other type is the segmentation approach,^{3,7} which tries to identify the chromosomal segments with copy number aberrations. In this paper, we focus on the segmentation approach. To our best knowledge, no current DCN analysis methods model sample mixture issue explicitly. We developed a novel minimum description length based segmentation method (MDLS) for DCN data analysis. Our MDLS method is designed to detect the underlying DNA copy number for each chromosomal segment, and at the same time, to infer the underlying tumor proportion of tumor tissue and normal tissue.

The remainder of the paper is organized as follows: Sec. 2 gives an overview of the previous methods; Sec. 3 presents the details of our minimum description length based segmentation method (MDLS); Sec. 4 demonstrates the experiment results of MDLS compared to three previous methods on synthetic data and real data. Section 5 concludes the paper with discussions and future work.

2. Previous Methods

There are three popular segmentation methods for the analysis of array-based DCN data: Circular Binary Segmentation (CBS),⁷ a Hidden Markov Model (HMM)³ based approach and Ultrasome.⁸

2.1. Circular Binary Segmentation

Circular binary segmentation (CBS)⁷ was proposed by Olshen and Venkatraman. It recursively applies a statistical test to test the hypothesis that there exists a segment with changed DNA copy number within the segment being considered. This method has the following limitations:

- Although CBS can successfully detect many obvious changes, the authors also admit in their paper that this procedure has low power to detect a change when the difference in means is small (which may be the result of impure samples) or when the width of the changed segment is small.
- It only outputs segments with no DNA copy number annotations.
- It does not address the sample mixture problem explicitly.

2.2. Hidden Markov Model approach

Several HMM-based methods have been proposed recently.^{3,9,10} One of the most popular methods is the method proposed by Fridlyand *et al.*³ which was further extended with MergeLevels.¹¹ It segments the array CGH data using a K-state HMM with continuous

output. This model tries to identify the optimal state sequence associated with observed data on a chromosome by choosing the state which is individually most likely at each probe with the Forward–Backward algorithm. The optimal HMM parameters are estimated using the Baum–Welch method, with the initialization done heuristically. The best number of states is estimated according to the Akaike information criterion¹² or the Bayesian information criterion.¹³ This approach has the following limitations:

- The states of the HMM do not have any intrinsic meaning (they are not tied to actual DNA copy numbers). Thus post-processing is necessary to actually segment the data.
- It does not address the sample mixture problem explicitly either.

2.3. Ultrasome

Ultrasome is recently introduced by Nilsson *et al.*⁸ It can detect genomic change points in DCN data with really high efficiency and achieved near-linear computational complexity. Ultrasome fits a piecewise constant function by minimizing an equation with two terms. The first term represents consistency with the original data; the second term represents penalty of the number of breakpoints. This is really similar to our method. Although the authors claimed that they can get near-linear complexity by using a special data structure, we suspect that they may use a heuristic method without exploring all possible segmentation situations. Because our method explores all possible segmentation situations, it has much better performance than Ultrasome. We will show the comparisons between all three previous methods and our method in Sec. 4.

3. The MDLS Method

Our proposed MDLS method is based on the Minimum Description Length (MDL) principle.¹⁴ In this section, we will first introduce the MDL principle and discuss the differences between MDLS and previous MDL methods. Then we will discuss the MDLS method in detail.

3.1. The Minimum Description Length principle

The Minimum Description Length (MDL) principle provides a heuristic solution to the model selection problem. The MDL principle, proposed by Jorma Rissanen, was used to evaluate the goodness of fit by minimizing the coding length of data.¹⁴ It is based on the following notion: trying to find as many regularities as possible in the data to compress the description length of the data, i.e. to describe the data using the least number of bits. MDL has the following properties:

- Occam's Razor¹⁵: MDL makes a trade-off between goodness-of-fit and complexity of the models.
- No overfitting: MDL automatically and inherently avoids overfitting.

All model selection methods choose a trade-off between goodness-of-fit and complexity of the models.^{16–18} The MDL principle provides a particular approach to achieve such a trade-off.^{15,19} In practice, such trade-offs lead to much better predictions of test data than the simplest or the most complex fitting.

Here is the formula¹⁴ to describe MDL:

$$\text{Min}_{(\theta)} DL(x) = -\ln P_{\theta}(x) + DL_{\theta}, \quad (1)$$

where $DL(x)$ is the description length function, θ is a set of parameters of the model, $P_{\theta}(x)$ is the probability density function of x by parameter vector θ . DL_{θ} is the description length of model.

$$DL_{\theta} = (n_k/2) \ln(N), \quad (2)$$

where n_k is the number of parameters in model k and N is the total number of data points. According to Eq. (1), the description length of sequence x is a sum of two terms. The first term is the data description length of sequence x given the model with parameter vector θ . The larger the possibility for observing sequence x with model with parameter vector θ , the smaller the description length for sequence x . This is reasonable because if a model fits the data better, the description length of the data should be smaller using the model. The second term is the description length for the model. It depends on the number of parameters. If the number of parameters is larger, the model is more complex. So the description length for the model will be larger, thus increasing the total description length. Therefore, when the MDL principle chooses a model with the minimum description length, it will make the trade-off between goodness-of-fit and complexity of the models.

MDL has been used in many areas, such as polygon fitting, 2D curve fitting, piecewise linear fitting, piecewise constant fitting, character recognition, moving object segmentation and motion estimation.^{15,19} However, the previous MDL methods for curve fitting cannot be used directly to solve the DCN data analysis problem. There are several reasons: first, most of previous methods deal with piecewise linear fitting, while the DCN data analysis problem requires piecewise constant fitting; second, there are several previous methods^{20,21} dealing with piecewise constant fitting. There is no constraint on the y -axis values of the horizontal segments in the piecewise constant fitting. However, in our case, those values should reflect biological meaning — DCN. The DNA copy numbers in homogeneous tumor tissue sample should be integers, not arbitrary values. For the above reasons, we devised a novel MDLS method for DCN data analysis.

3.2. Our MDLS approach

Basically, our MDLS approach produces a piecewise constant fitting with biological constraints, that is, the fitted y -axis values for the horizontal segments correspond to integer DCN values. We explain the algorithm detail of MDLS in the section below.

3.2.1. Calculating MDL for a segment—The first step of our MDLS method is to calculate the MDL for an arbitrary segment in the given DCN data. We assume that the DCN data within a chromosomal segment with the same DCN follow a Gaussian distribution, described by Eq. (3), where μ is the mean, σ is the standard deviation. We use Eqs. (4) and (5) to determine the μ and σ for the segment $[i, j]$, where i and j are the indices of end points of segment $[i, j]$.

$$\varphi_{\mu, \sigma^2}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (3)$$

$$\mu_{i,j} = \frac{1}{j-i+1} \sum_{n=i}^j x_n, \quad (4)$$

$$\sigma_{i,j}^2 = \frac{1}{j-i+1} \sum_{n=i}^j (x_n - \mu_{i,j})^2. \quad (5)$$

We assume that the test sample is a mixture of homogeneous tumor cells and normal cells. The proportion of tumor cells in the sample is denoted by P_t , and the DCN at a particular chromosomal location is denoted by c . Then theoretically the DCN $\log_2$ -Ratio data at this particular chromosomal location can be described by the following equation:

$$S = \log_2 \frac{cP_t + 2(1 - P_t)}{2}. \quad (6)$$

If the proportion of tumor cells P_t is already known, the DCN for segment $[i, j]$ can be calculated by the following equation:

$$c_{i,j} = \frac{2 \cdot 2^{\mu_{i,j}} - 2(1 - P_t)}{P_t}. \quad (7)$$

Because $\mu_{i,j}$ is the mean observed signal for segment $[i, j]$, for the whole segment $[i, j]$ we have

$$\mu_{i,j} = \log_2 \frac{c_{i,j}P_t + 2(1 - P_t)}{2}. \quad (8)$$

In ideal experimental conditions with minimum noise, $c_{i,j}$ calculated in this way should be an integer, because DCNs should be integers. However, in real data, due to noise, the observed signal $c_{i,j}$ is typically a real value, instead of an integer. Therefore, we round the value $c_{i,j}$ to the closest integer:

$$\hat{c}_{i,j} = \text{round}(c_{i,j}). \quad (9)$$

Then Eq. (10) is used to find the adjusted mean signal $\mu'_{i,j}$.

$$\mu'_{i,j} = \log_2 \frac{\hat{c}_{i,j}P_t + 2(1 - P_t)}{2}. \quad (10)$$

By using μ , σ and the definition of the probability density function of Gaussian distribution, we can calculate the first term of the description length function [Eq. (1)].

$$\ln p(x[i, j] | \mu'_{i,j}, \sigma_{i,j}^2) = -\frac{j-i+1}{2} \ln(2\pi\sigma_{i,j}^2) - \frac{\sum_{n=i}^j (x_n - \mu'_{i,j})^2}{2\sigma_{i,j}^2}. \quad (11)$$

The second term of the description length function can be easily calculated with Eq. (2).

Note that the above equation is based on the assumption that the proportion of tumor cells P_t is already known. However, in reality this correct value P_t is not known, but to be inferred from the data. To this end, we used a hill climbing algorithm to find the optimal value, as discussed in Sec. 3.2.2.

3.2.2. Using dynamic programming to find the optimal segment structure on a chromosome—For a chromosome with n probes, there will be 2^{n-1} possible different ways to segment it. A naive brute force algorithm could be implemented to calculate the description length (DL) for each structure. By comparing the DLs, the minimum one can be chosen as the optimal one. The segment structure with MDL is the best structure. The running time of brute force algorithm would be exponential. Here we propose a dynamic programming algorithm to solve this problem in polynomial time. The key idea is this recursive equation:

$$MDL_{i,l} = \begin{cases} DL_{i,1} & \text{if } l=1 \\ \min_{1 \leq j \leq l-1} (DL_{i,j}, MDL_{i,j} + MDL_{i+j,l-j}) & \text{if } l > 1 \end{cases},$$

where $MDL_{i,l}$ represents the MDL for a segment with length l , that starts at i th probe location on a chromosome. $DL_{i,l}$ denotes the description length for the segment starting from probe position i with length l , assuming that the whole segment has the same DCN. This value can be calculated by Eq. (1). The model with one DCN for the whole segment may not have the MDL as compared with other models with different sub-segments, each having a different DCN. Therefore we calculate description lengths for the segment with all possible sub-segments situations. Among all the description lengths with and without sub-segments, the MDL can be obtained. By employing the above formula, the MDLs of small segments can be used to calculate the MDLs of large segments. In this way, the running time is reduced to $O(n^3)$, where n is the number of probes on a chromosome.

Below is the pseudocode for computing MDL for a test sample, assuming that the tumor cell proportion is P_t :

COMPUTE-MDL(P_t) :

```

1:  tot MDL  $\leftarrow 0$ 
2:  for each chromosome  $chr_k$  do
3:    for  $i \leftarrow 1$  to  $len\_chr_k$  do
4:       $MDL[i, 1] \leftarrow Compute\_dl(i, 1)$ 
5:       $div[i, 1] \leftarrow 0$ 
6:       $sig\_fit[i, 1] \leftarrow \mu'_{i,i}$ 
7:       $DCN\_fit[i, 1] \leftarrow \hat{c}_{i,i}$ 
8:    end for
9:    for  $l \leftarrow 2$  to  $len\_chr_k$  do
10:     for  $i \leftarrow 1$  to  $(len\_chr_k - l + 1)$  do
11:        $temp\_min \leftarrow Compute\_dl(i, l)$ 
12:        $temp\_div \leftarrow 0$ 

```

```

13:   for  $j \leftarrow 1$  to  $(l - 1)$  do
14:     if  $temp\_min > (MDL[i, j] + MDL[i + j, l - j])$  then
15:        $temp\_min \leftarrow (MDL[i, j] + MDL[i + j, l - j])$ 
16:        $temp\_div \leftarrow j$ 
17:     end if
18:   end for
19:    $MDL[i, l] = temp\_min$ 
20:    $div[i, l] \leftarrow temp\_div$ 
21:    $sig\_fit[i, l] \leftarrow \mu'_{i,i+l-1}$ 
22:    $DCN\_fit[i, l] \leftarrow \hat{c}_{i,i+l-1}$ 
23: end for
24: end for
25:  $tot\ MDL \leftarrow tot\ MDL + MDL[l, len\_chr_k]$ 
26: end for
27: return  $totMDL$ 

```

3.2.3. Estimating the true tumor cell proportion P_t —The above *COMPUTE_MDL* algorithm assumes that the tumor cell proportion P_t is already known. However, P_t is a value to be inferred from the data. Intuitively, the model assuming the correct P_t value should fit the data best as compared to other models with different P_t values. Therefore the MDL computed with correct P_t value should be the smallest among the MDLs computed with other assumed P_t values. To solve the search problem, we employed the hill climbing algorithm with multiple starting points. Our approach works as follows. First, our MDLS method calculates the MDLs for 9 tumor cell proportions of 0.1, 0.2, ..., 0.9. By using the 9 tumor cell proportions as starting points, the hill climbing algorithm is employed to find local minima around the 9 starting points. Comparing the local minima, the tumor cell proportion associated with the smallest MDL is chosen as our inferred tumor cell proportion. The inferred tumor cell proportion found in this way may not be the globally optimal solution, as with any other local search algorithm. Our synthetic results show that the inferred tumor cell proportion is very close to the hidden true tumor cell proportion. The mean deviation between inferred tumor cell proportions and true tumor cell proportions is less than 0.01.

We implemented the MDLS method in C language and Matlab. Our program can be freely downloaded from <http://engr.smu.edu/~swang/research.html>.

4. Experiment Results

4.1. Synthetic data generation

Willenbrock and Fridlyand¹¹ described a simulation model for generating synthetic array CGH data. Their model was designed realistically using a real primary breast tumor dataset of 145 samples. The synthetic DCN data on a chromosome were generated as follows:

1. As suggested in Ref. ¹¹, chromosomal segments with DNA copy numbers 0, 1, 2, 3, 4, and 5 were generated with the probability 0.01, 0.08, 0.81, 0.07, 0.02, and 0.01, respectively.
2. Following the same model in Ref. ¹¹, each sample was assumed to be a mixture of tumor cells and normal cells, and the proportion of tumor cells P_t was chosen at 0.33, 0.45, 0.56, 0.67, 0.78, and 0.9. The expected $\log_2$ ratio of intensity computed as Eq. (6), is then the latent true signal.
3. Gaussian noise of mean 0 and variance σ^2 was added to the latent true signal.
4. Nonequispaced probes were placed on the chromosome. The distances between adjacent probes were randomly drawn from the empirical distribution of distances obtained from the UCSF HumArray2 BAC array.
5. In order to investigate the capability of different DCN data analysis algorithms to identify chromosomal segments with abnormal copy numbers of different lengths, we generated synthetic data containing abnormal segments and normal segments with different lengths separately. Specifically, the lengths for the segments with aberrant copy numbers were chosen at 10th, 30th, 50th, 70th, and 90th percentiles of the corresponding empirical length distribution given in Ref. ¹¹. We also produced a set of data without abnormal segment length restriction.

Using this model, for each of the five segment lengths in step 5 and each noise level σ of 0.1, 0.125, 0.15, 0.175, and 0.2, we generated 20 samples, each containing 20 chromosomes. Each chromosome has a length of 200 Mb and around 150 probes. Therefore there are 30 different cases for each segment length. Table 1 shows the 30 different cases in the order of decreasing signal-to-noise ratio.

4.2. Results on synthetic data

In this section, we present the results on synthetic data to evaluate the performance of our MDLS method. A probe locus depending on the DCN at that locus should be classified as normal, gain or loss. Our MDLS method outputs integer DCNs. Therefore it is straightforward to assign one of the three class labels (normal, gain and loss) to each probe locus. In the synthetic data, the true class label for each probe is already known. If the class label determined by a DCN analysis method is different from the corresponding true class label for a probe locus, this is considered as a misclassification. Accuracy is defined as the number of correctly classified probes divided by the total number of the probes in the data. Because CBS, HMM and Ultrasome do not explicitly model the underlying copy number, and only output segments without class labels, additional steps are needed to calculate accuracies for CBS and HMM. In order to calculate their accuracies, a pair of thresholds $(-t, t)$ are set on their output $\log_2$ ratio values. If the value at a probe locus in the output falls between $-t$ and t , this probe is classified as normal. If the value is above t , it is classified as gain. Otherwise if the value is below $-t$, it is classified as loss. t takes values 0.01, 0.02, 0.03, ..., 2.0. Each different value may result in a different accuracy for CBS, HMM and Ultrasome. In our comparisons, we used the best accuracies among those obtained from different threshold t as the accuracies for CBS, HMM and Ultrasome. Similar procedures are applied to calculate sensitivities and specificities of CBS, HMM and Ultrasome.

Figure 1 shows the comparison of the performances for MDLS, CBS, HMM and Ultrasome when the segment length is of the 10th percentile. Row one shows accuracy versus the proportion of tumor cells P_t , row two shows sensitivity versus the proportion of tumor cells P_t , row three shows specificity versus the proportion of tumor cells P_t . Although our comparison method favors CBS, HMM and Ultrasome, our MDLS method gets better accuracies as compared to other methods, as shown in Fig. 1 row one. Figure 1 row two and

row three also show that the sensitivities and specificities of MDLS are also better than CBS and HMM. Although Ultrasome has the best performance in specificity among all four methods, the sensitivity of Ultrasome is worst in all four methods. This indicates that Ultrasome is conservative to detect some of the aberrant regions in the genome. As shown in Fig. 1 row four, our MDLS method can predict the copy numbers of segments with an accuracy of more than 90% in most cases.

Figure 2 shows the comparison of the performances for MDLS, CBS, HMM and Ultrasome when the segment length is without restriction. In this case, MDLS achieves better accuracies as compared to other methods, as shown in Fig. 2 row one. According to Fig. 2 row two and row three, the sensitivities and specificities of MDLS are still better than CBS and HMM. But the improvements are not as significant as those shown in Fig. 1 row two and row three. This indicates that MDLS is suitable for the situations with small segment lengths. The results of Ultrasome in Fig. 2 also show its conservation to detect aberrant segments in the genome. As shown in Fig. 2 row four, our MDLS method can predict the copy numbers of segments with an accuracy of more than 94% in all cases.

Figure 3(a) and 3(b) show box plots of inferred tumor cell proportions of MDLS when segment length is of the 10th percentile and without restriction, respectively. The horizontal dotted lines indicate the underlying true tumor cell proportions. As shown in Fig. 3, the inferred tumor cell proportions are very close to the true tumor cell proportion. The mean deviation between inferred tumor cell proportions and true tumor cell proportions is less than 0.01.

Due to the limitation of space, we are not including results from all cases in this paper. The complete experiment results can be found in the supplemental material at <http://lyle.smu.edu/~swang/research.html>.

4.3. Results on real data

We also tested our MDLS method on two public microarray datasets containing DNA copy number data. The first dataset contains 33 small cell lung cancer (SCLC) tumors, 13 SCLC cell lines, 19 bronchial carcinoids, and 9 gastrointestinal (GI) carcinoids.²² The DNA copy number data can be downloaded from

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21468>. The dataset was obtained through Agilent SurePrint 180 K G3 Human CGH Microarray or Agilent 105K Human Genome CGH Microarray. The second dataset is from 26 different primary Glioblastoma Multiforme (GBM) tumors.²³ The raw data can be downloaded from <http://smd.stanford.edu/>. In our experiments, the normalized data are used. It is available at <http://www.chip.org/~ppark/Supplements/Bioinformatics05b.html>.

In Voortman *et al.*,²² the authors observed 14 recurrent regions of amplification and 6 recurrent regions of loss. By using our MDLS methods, we also discovered all amplifications and losses which were found in the article. Figure 4 shows two examples of MDLS results for the dataset which coincide with the results in Voortman *et al.*²² Amplifications around the fibroblast growth factor receptor 1 (FGFR1) gene and the Janus kinase 2 (JAK2) gene are shown in Fig. 4(a) and 4(b), separately.

We chose the second dataset because it was used as a benchmark in a recent paper²⁴ where 11 algorithms for array CGH data analysis were compared.

Figure 5 shows the fitting result for chromosome 7 in the sample GBM29. This chromosome is characterized by small amplified regions with high magnitude. Amplifications around the

EGFR locus have been implicated in some tumors. In Fig. 5, we can see that there are three amplifications around EGFR, which is consistent with the results in previous paper.²⁴

Figure 6 shows the fitting result for chromosome 13 in the sample GBM31. On this chromosome, there is a large region of loss with low magnitude. As can be observed in Fig. 6, our method captured this loss region.

5. Conclusions and Discussions

DNA copy number data analysis is an important task because it might lead to the discovery of novel tumor suppressor genes and oncogenes. In this paper, we proposed a novel minimum description length principle based segmentation method (MDLS) for the analysis of DCN data. In MDLS, the tumor sample is explicitly modelled as a mixture of tumor cells and normal cells, reflecting what occurs in real situation. Experiment results on realistically generated synthetic data showed that the performance of MDLS is significantly better than that of three widely used methods (CBS, HMM and Ultrasome). In addition to the DCN for each chromosomal segment, MDLS also outputs the inferred tumor cell proportion for the test sample.

Empirical results show that MDLS is more sensitive to signal than CBS, HMM and Ultrasome, especially in the situations when the chromosomal segments with aberrant DCNs are small. Because our MDLS method can identify small deletions and small amplifications, as evidenced by experimental results, we believed that MDLS is better suited for high density CGH array data. As the density of microarrays increases, there is increased possibility to find micro-deletions. Tanay and Siggia²⁵ discussed that short insertions and deletions are a significant factor in the mutational input that feeds into the evolutionary process. Rodriguez *et al.*²⁶ discovered micro-deletions in the tumor suppressor gene PTEN. Currently, the running time of MDLS is $O(n^3)$, where n is the number of probes on a chromosome. As future work, we plan to reduce its running time, possibly through approximation.

Acknowledgments

This study was supported in part by grants from the National Science Foundation (DMS-0907562) and the National Institute of Health (NCI 1R01CA152301-01 and NIDA 1R21DA027592).

References

1. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nat Genet.* 2005; 37 Suppl:S11–S17. [PubMed: 15920524]
2. Bilke S, Chen QR, Whiteford CC, Khan J. Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays. *Bioinformatics.* 2005; 21(7): 1138–1145. [PubMed: 15539449]
3. Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. Hidden Markov models approach to the analysis of array CGH data. *J Multivar Anal.* 2004; 90:132–153.
4. Wang Y, Wang S. A novel stationary wavelet denoising algorithm for array-based DNA copy number data. *Int J Bioinform Res Appl.* 2007; 3(2):206–222. [PubMed: 18048189]
5. Beheshti B, Braude I, Marrano P, Thorner P, Zielenska M, Squire JA. Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization. *Neoplasia.* 2003; 5(1):53–62. [PubMed: 12659670]
6. Eilers PH, de Menezes RX. Quantile smoothing of array CGH data. *Bioinformatics.* 2005; 21(7): 1146–1153. [PubMed: 15572474]
7. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics.* 2004; 5:557–572. [PubMed: 15475419]

8. Nilsson B, Johansson M, Al-Shahrour F, CAE, Ebert BL. Ultrasome: Efficient aberration caller for copy number studies of ultra-high resolution. *Bioinformatics*. 2009; 25(8):1078–1079. [PubMed: 19228802]
9. Shah SP, Xuan X, Deleeuw RJ, Khojasteh M, Lam W, Ng R, Murphy KP. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*. 2006; 22(14):431–439.
10. Marioni JC, Thorne NP, Tavare S. BioHMM: A heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*. 2006; 22(9):1144–1146. [PubMed: 16533818]
11. Willenbrock H, Fridlyand J. A comparison study: Applying segmentation to array CGH data for downstream analyses. *Bioinformatics*. 2005; 21:4084–4091. [PubMed: 16159913]
12. Akaike H. Fitting autoregressive models for prediction. *Ann Inst Stat Math*. 1969; 21:243–247.
13. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978; 6:461–464.
14. Rissanen J. Modeling by the shortest data description. *Automatica*. 1978; 14:465–471.
15. Hansen MH, Yu B. Model selection and the principle of minimum description length. *J Am Stat Assoc*. 2001; 96:746–774.
16. Wallace CS, Georgeff MP. A general objective for inductive inference. Technical Report. 1983; 32
17. Wallace CS, Georgeff MP. A general selection criterion for inductive inference. Technical Report. 1984; 44
18. Lavielle M. Optimal segmentation of random processes. *IEEE Trans Signal Process*. 1988; 46:1365–1373.
19. Barron A, Rissanen J, Yu B. The minimum description length principle in coding and modeling. *IEEE Trans Information Theory*. 1998; 44(6):2743–2760.
20. Wang Z, Willett P. Two algorithms to segment white gaussian data with piecewise constant variances. *IEEE Trans Signal Process*. 2003; 51(2):373–385.
21. Nowak RD, Figueiredo MA. Unsupervised segmentation of poisson data. 15th Int Conf Pattern Recognition. 2000; 3:3159–3163.
22. Voortman J, Lee JH, Killian JK, Suuriniemi M, Wang Y, Lucchi M, Smith WI Jr, Meltzer P, Wang Y, Giaccone G. Array comparative genomic hybridization-based characterization of genetic alterations in pulmonary neuroendocrine tumors. *Proc Natl Acad Sci USA*. 2010; 107(29):13040–13045. [PubMed: 20615970]
23. Bredel M, Bredel C, Juric D, Harsh GR, Vogel H, Recht LD, Sikic BI. High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res*. 2005; 65(10):4088–4096. [PubMed: 15899798]
24. Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*. 2005; 21(19):3763–3770. [PubMed: 16081473]
25. Tanay A, Siggia ED. Sequence context affects the rate of short insertions and deletions in flies and primates. *Genome Biol*. 2008; 9(2)
26. Rodriguez J, Guiteau J, Nazareth L, Reid J, Goss J, Gibbs R, Gingras M. Sequencing the full-length of the phosphatase and tensin homolog (PTEN) gene in hepatocellular carcinoma (HCC) using the 454 gs20 and Illumina GA DNA sequencing platforms. *World J Surg*. 2008

Biographies

Siling Wang is currently a Ph.D. candidate in the Department of Computer Science and Engineering, Southern Methodist University in Dallas, TX. His research interests include issues related to bioinformatics, algorithms for array-based DNA copy number data analysis and feature selection algorithms for high-throughput biological data.

Yuhang Wang is an Assistant Professor in the CSE Department at Southern Methodist University. He received the Ph.D. degree in Computer Science from Dartmouth College in March 2005. Before joining SMU, he worked as a Postdoctoral Research Fellow at Harvard Medical School and Brigham and Women's Hospital. His research interests include

bioinformatics, medical image analysis, data mining, machine learning, pattern recognition, multimedia information retrieval, and computational geometry.

Yang Xie, Ph.D., MPH, is an Assistant Professor in Division of Biostatistics, and the Director of Bioinformatics core at the Harold C. Simmons Comprehensive Cancer Center at UTSW. Dr. Xie's research expertise includes molecular profiling data integration and analysis, Bayesian modeling, database development and management, risk modeling and computational biology.

Guanghua Xiao, Ph.D., is an Assistant Professor in Division of Biostatistics, Department of Clinical Sciences at UTSW. Dr. Xiao's primary statistical expertise is high-throughput data analysis and he has extensive experience with gene expression microarrays, next-generation sequencing data, ChIP-chip, array CGH (comparative genomic hybridization) and integrated analysis.

Fig. 1.

Experiment results when abnormal segment length is of the 10th percentile. Row one shows accuracy versus the proportion of tumor cells P_t , row two shows sensitivity versus the proportion of tumor cells P_t , row three shows specificity versus the proportion of tumor cells P_t , row four shows accuracy of copy number by MDLS versus the proportion of tumor cells P_t . Column one through five represents noise level 0.1, 0.125, 0.15, 0.175 and 0.2, respectively. Circle represents MDLS's results. Plus sign represents CBS's results. Star sign represents HMM's results. Cross sign represents Ultrasome's results.

Fig. 2.

Experiment results without abnormal segment length restriction. Row one shows accuracy versus the proportion of tumor cells P_t , row two shows sensitivity versus the proportion of tumor cells P_t , row three shows specificity versus the proportion of tumor cells P_t , row four shows accuracy of copy number by MDLS versus the proportion of tumor cells P_t . Column one through five represents noise level 0.1, 0.125, 0.15, 0.175 and 0.2, respectively. Circle represents MDLS's results. Plus sign represents CBS's results. Star sign represents HMM's results. Cross sign represents Ultrasome's results.

(a)

Fig. 3. Box plots of inferred tumor cell proportions by MDLS (a) when segment length is of the 10th percentile, (b) when segment length is without restriction.

Fig. 4.
Two examples of MDLS results for Voortman's dataset.

Fig. 5.
Fitting result of MDLS on chromosome 7 in the Glioblastoma Multiforme sample GBM29.

Fig. 6.
Fitting result of MDLS on chromosome 13 in the Glioblastoma Multiforme sample GBM31.

Table 1

The decreasing signal-to-noise ratios for different tumor cell proportion (TCP) and noise level situations.

Index	TCP	Noise level	S/N ratio	Index	TCP	Noise level	S/N ratio
1	0.9	0.1	5.36	16	0.67	0.175	2.38
2	0.78	0.1	4.75	17	0.78	0.2	2.38
3	0.9	0.125	4.29	18	0.56	0.15	2.37
4	0.67	0.1	4.17	19	0.45	0.125	2.34
5	0.78	0.125	3.80	20	0.33	0.1	2.20
6	0.9	0.15	3.57	21	0.67	0.2	2.08
7	0.56	0.1	3.56	22	0.56	0.175	2.04
8	0.67	0.125	3.33	23	0.45	0.15	1.95
9	0.78	0.15	3.17	24	0.56	0.2	1.78
10	0.9	0.175	3.06	25	0.33	0.125	1.76
11	0.45	0.1	2.93	26	0.45	0.175	1.67
12	0.56	0.125	2.85	27	0.33	0.15	1.47
13	0.67	0.15	2.78	28	0.45	0.2	1.46
14	0.78	0.175	2.71	29	0.33	0.175	1.26
15	0.9	0.2	2.68	30	0.33	0.2	1.10