MAXIMUM LIKELIHOOD PRINCIPLE FOR DNA COPY NUMBER ANALYSIS

Abdullah K. Alqallaf, and Ahmed H. Tewfik

Department of Electrical and Computer Engineering, University of Minnesota
200 Union Street. SE, Minneapolis, MN 55455, USA
alqal001@umn.edu, and tewfik@umn.edu

ABSTRACT

Microarray technologies had been used to measure DNA copy number data. The copy number represents the relative fluorescent intensity level between control and test DNA samples. Variation of this number may lead to many genetic diseases such as cancer. Unfortunately, the observed copy numbers are corrupted by noise due to experimental errors and probes accuracy, making the variations hard to detect. Different techniques had been proposed to denoise the data and to extract the important feature such as the breakpoints from the variant regions. In this paper, we present a robust procedure for the analysis of DNA copy number data based on maximum likelihood principle using global information of the entire data record. We show that Dynamic programming can be used to compute the DNA copy number estimates and reduce the computational complexity. Furthermore, we employ the Minimum Description Length rule to estimate the number of unknown parameters. Using simulated and real data, we show that the proposed method outperforms other popular commercial software and published algorithms.

Index Terms— DNA Copy Number, Comparative Genomic Hybridization, Maximum Likelihood rule, Minimum description Length, Dynamic programming

1. INTRODUCTION

Genetic diseases are characterized by the presence of genetic instabilities. Microrray-based Comparative Genomic Hybridization (aCGH) is a molecular technology used to measure the genetic instabilities in the form of DNA copy number (DCN). It provides a high-resolution method to map and measure relative changes in DCN simultaneously at thousands of genomic loci in logarithmic scale. The ratio reflects the relative fluorescent intensity level between a reference (R) DNA as control sample and test (T) DNA sample possibly diseased. We expect to see $\log_2(T/R)=0$ for normal state, $\log_2(T/R)>0$ for amplification state, and $\log_2(T/R)<0$ for deletion state. These intensity ratios are informative about DNA copy number variations. Due to the logarithmic scale and the probes performance, the data can be approximated as a piecewise function of short and long intervals with different intensity levels that are not uniformly distributed along the genome [10]. However, the DCN data corrupted by noise due to experimental errors. Different methods have been developed to denoise and to detect copy number variations (CNVs) based on statistical models and local-smoothing techniques. These methods have advantages and disadvantages. We shall discuss these issues in the prior work section.

In this paper, we present an optimal method (for large data record) based on Maximum likelihood (ML) principle [4] for the analysis of DCN data. Here we use Dynamic Programming (DP) [5] to compute the unknown parameters and to reduce the computational complexity of our method. Next, we employ the Minimum Description Length (MDL) [6] procedure to estimate automatically the number of variant regions along the entire data which is unknown priori. Unlike the local-smoothing technique, the proposed method considers global information from the entire data record.

The paper is organized as follows. Prior work is presented in section 2. Section 3 presents DCN data modeling as one-dimensional piecewise discrete signal and the formulation of the ML method using DP for the detection of the variant regions and their fluorescent intensity levels along with the estimation of the number of these regions using MDL procedure. Comparison study between our proposed technique and other efficient methods using simulated and real data validated using the experimental molecular method quantitative polymerase chain reaction (QPCR) is presented in section 4. Finally, conclusions based on the observed results are provided in section 5.

2. PRIOR WORK

Various techniques had been proposed for the analysis of the DNA copy number variations. Mainly, they fall into two categories: statistical based models and smoothing techniques. In this section, we briefly review recent and efficient approaches. The Circular Binary Segmentation (CBS) presented by [1] is an example of statistical models. It allows splitting the data record into smaller segments until no more changes are detected in any of the segments obtained from the change-points already found. A different modeling approach involves the use of Hidden Markov Models (HMMs) presented by [2], in which the underlying copy numbers are the hidden states with certain transition probabilities. They study DCN variations that naturally occur in normal populations and using an HMM based approach they compare signals for two individuals and seek intervals of four or more probes in which DCN data are likely to be different. On the other hand, the local-selective smoothing techniques provide alternative methods for processing the DCN data that are characterized by small and long intervals with of sharp transitions and singularities at boundaries edges (breakpoints). The wavelet footprints presented by [7] is an example of smoothing technique. It is used to obtain a basis for representing the DCN data that is maximally sparse and then sparse Bayesian learning is applied to infer the copy number changes from the noisy data. Our previously proposed algorithm based on discretization of the partial

differential equation (PDE), nonlinear diffusion filter (*NLDF*) [9], is another example of local-smoothing techniques. Although the techniques are computationally efficient, they use local information about the data to detect the variations, which may lead to an increase in the false positive rate. Unlike the statistical models and the smoothing techniques, in the next section, we present our method based on the Maximum likelihood rule to detect the DCN variations. It has low computational complexity due to the use of DP and it uses global information from the data to reduce the false positive rate.

3. MATHEOD AND MATERIALS

In this section, we present a technique based on the principle of maximum likelihood estimation to extract the important feature of the DNA copy number data. It is divided into 3 parts. 1) data modeling 2) variant regions estimation 3) Estimation of the number of variant regions.

3.1. Data modeling

DNA copy number observations are traditionally modeled as one-dimensional discrete time series with multilevel and jumps at unknown transition times, corrupted by additive white Gaussian noise (AWGN) of variance σ^2 [10]. Specifically,

$$y[n]=f[n]+w[n]. \qquad n=0,1,2,\ldots, N\text{-}1 \tag{1}$$

where $y[n]$ is the observed DCN data and $w[n]$ is AWGN and $f[n]$ is the true DCN signal to be estimated. Then, we define

$$f[n]=\begin{cases} A_1 & n=0,1,\cdots,n_1-1,\\ A_2 & n=n_1,n_1+1,\cdots,n_2-1,\\ \vdots & \\ A_M & n=n_{M-1},n_{M-1}+1,\cdots,N-1. \end{cases}$$

or in the compact form

$$f[n]=\sum_{i=1}^{M} A_i\left[u[n_{i-1}]-u[n_i]\right], \tag{2}$$

where $n_0=0<n_1<n_2< \ldots <n_{M-1}<n_M=N$ and $u[n]$ is the unit step function. Here N is the length of DCN data.

3.2. Variant regions estimation

Based on the data assumption of the previous section and assuming that we are given the number of variant regions M, the next step is to apply the principle of maximum likelihood (ML) to estimate the breakpoints and intensity levels corresponding to these regions. The i^{th} variant region can be characterized by the PDF $p_i([y[n_{i-1}]:y[n_i\text{-}1]];A_i)$, where A_i and n_i are the unknown parameters representing the intensity level and the breakpoint, respectively. Moreover, each variant region is assumed to be statistically independent of all other regions. Hence, the PDF of the entire data record can be written as

$$\begin{aligned} p(\mathrm{y};\mathrm{A},\mathrm{n}) &= \prod_{i=1}^{M} p_i(y[n_{i-1}:n_i-1];A_i) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{M}\left[\sum_{n=n_{i-1}}^{n_i-1}\left(y[n]-A_i\right)^2\right]\right], \end{aligned} \tag{3}$$

Taking the logarithm yields

$$\sum_{i=1}^{M}\ln p_i\left(y[n_{i-1}:n_i-1];A_i\right),$$

$$\begin{aligned} \Delta_i[n_{i-1},n_i-1] &= -\ln p_i(\mathrm{y}[n_{i-1},n_i-1];\hat{\theta}_i) \\ &= \frac{1}{2\sigma^2}\sum_{n=n_{i-1}}^{n_i-1}\left(y[n]-\hat{A}_i\right)^2+\frac{n_i-n_{i-1}}{2}\ln(2\pi\sigma^2), \end{aligned}$$

So we need to minimize $\sum_{i=1}^{M}\Delta_i[n_{i-1}:n_i-1]$ over $\boldsymbol{n}=\{n_1,\ldots,n_{M\text{-}1}\}$.

Or by retaining the data-dependent terms

$$\begin{aligned} J(\mathrm{A},\mathrm{n}) &= \sum_{i=1}^{M}\Delta_i[n_{i-1}:n_i-1] \\ &= \sum_{i=1}^{M}\left[\sum_{n=n_{i-1}}^{n_i-1}\left(y[n]-\hat{A}_i\right)^2\right], \end{aligned} \tag{4}$$

where the joint MLE of $\boldsymbol{A}=[A_1, A_2, \ldots, A_M]^{\mathrm{T}}$, $\boldsymbol{n}=[n_1, n_2, \ldots, n_M]^{\mathrm{T}}$ are found by maximizing (3) or minimizing (4) over $\boldsymbol{n}$. Here we do not assume knowledge of the intensity levels so they must be jointly estimated with the breakpoints. Clearly, if the breakpoints n_i's of each variant region were known, the MLE of their corresponding intensity levels A_i's would be given by the sample mean of the data over each variant region as

$$\hat{A}_i=\frac{1}{n_i-n_{i-1}}\sum_{n=n_{i-1}}^{n_i-1}y[n], \qquad for\ i=1:M \tag{5}$$

where $n_0=0$ and $n_M=N$-1.

In summary, to maximize (3) we need to choose a set of breakpoints n_i's, estimate their intensity levels A_i's, sum the logarithm of their PDFs, and repeat the process for all possible set of breakpoints to determine which set yield the maximum. The difficulty of doing so is the exponential growing rate of the number of the possible variant regions M or in general $O(N^M)$. To solve the problem, we need to minimize the data-dependent term $\boldsymbol{J}(\boldsymbol{A}, \boldsymbol{n})$ over $\boldsymbol{A}$ and $\boldsymbol{n}$ or equivalently to minimize $\boldsymbol{J}(\hat{\boldsymbol{A}}, \boldsymbol{n})$ over $\boldsymbol{n}$. Fortunately, (4) exhibits the Markov property, which allows us to apply the technique of DP to reduce the computational complexity to a more manageable level. The computational complexity of DP is linearly proportional with the number of variant regions M. The mathematical formulation of the problem in terms of DP can be summarized as follows. To begin the recursion for minimum $\boldsymbol{J}(\hat{\boldsymbol{A}}, \boldsymbol{n})$ over $\boldsymbol{n}$, we include some restrictions and constraints on the breakpoints. Let,

$$I_k(L)=\min_{\substack{n_1,n_2,\ldots,n_{k-1}\\ n_0=0,n_k=L+1}}\sum_{i=1}^{k}\Delta_i\left[n_{i-1},n_i-1\right], \tag{6}$$

where $1\le n_1< n_2< \ldots < n_{k\text{-}1}\le L$.

$$\begin{aligned} I_k(L) &= \min_{n_{k-1}=L+1}\ \min_{\substack{n1,n2,\ldots,n_{k-2}\\ n_0=0}}\sum_{i=1}^{k}\Delta_i\left[n_{i-1},n_i-1\right], \\ &= \min_{n_{k-1}=L+1}\ \min_{\substack{n1,n2,\ldots,n_{k-2}\\ n_0=0}}\sum_{i=1}^{k-1}\Delta_i\left[n_{i-1},n_i-1\right]+\Delta_k\left[n_{k-1},n_k-1\right], \\ &= \min_{n_{k-1}=L+1}\left[I_{k-1}(n_{k-1}-1)+\Delta_k\left[n_{k-1},n_k-1\right]\right], \\ &= \min_{n_{k-1}}\left[I_{k-1}(n_{k-1}-1)+\Delta_k\left[n_{k-1},L\right]\right], \\ I_k(L) &= \min_{k-1\le n_{k-1}\le L}\left[I_{k-1}(n_{k-1}-1)+\Delta_k\left[n_{k-1},L\right]\right], \end{aligned} \tag{7}$$

The recursion process involved in the implementation of the DP algorithm to compute the solution assuming a minimum length of the variant region of one data sample, can be summarized as follows:

1) For k=1 (the maximum likelihood for all one-data sample regions),

a) Compute $I_1[L]$ for L=0:N-1 using (7) as

$$I_1(L) = \Delta_1[n_0 = 0, n_1 - 1 = L] = \Delta_1[0, L] = \sum_{n=0}^{L} \left(y[n] - \hat{A}_1\right)^2,$$

This is the minimum least square errors when the data record $[n_{k-1}, L]$ is used to estimate the mean, where

$$\hat{A}_1 = \frac{1}{L+1} \sum_{n=0}^{L} y[n],$$

b) Store the result in $I_1[L]$.

2) For k=2, (the maximum likelihood for all two-data sample regions)

a) Compute $I_2[L]$ for L=1:N-1 using (7) as

$$I_2(L) = \min_{1 \le n_1 \le L} \left[I_1(n_1 - 1) + \Delta_2[n_1, L]\right],$$

where we already determined $I_1(n_1\text{-}1)$ from the first step and the second term is determined using (4)

b) Store the result in $I_2[L]$.

c) For each L determine the value of n_1 that minimizes $I_2[L]$ and call it $n_1(L)$.

3) Repeat the same calculations of step 2 for k=3:M-1 and L=k-1:N-1.

4) Finally, for k=M, (the maximum likelihood for the last variant region where the solution to our original problem occurs)

a) Compute $I_M[L]$ for L=N-1 using (7) as

$$I_M(L) = \min_{M-1 \le n_{M-1} \le L} \left[I_1(n_{M-1} - 1) + \Delta_M[n_{M-1}, L]\right],$$

here, we denote $\hat{n}_{M-1}$ as the minimum value of $n_{M\text{-}1}$ by as $n_2(N\text{-}1)$.

b) Using backward recursion, the estimated breakpoints are found as

$$\begin{aligned} \hat{n}_{M-1} &= n_{M-1}(N-1), \\ \hat{n}_{M-2} &= n_{M-2}(\hat{n}_{M-1} - 1), \\ &\vdots \quad \vdots \\ \hat{n}_1 &= n_1(\hat{n}_2 - 1). \end{aligned}$$

c) Using the estimated breakpoints, the estimated intensity level $\hat{A}_i$ for each variant region is easily found using (5).

3.3. Estimation of the number of variant regions

In the previous section, we assumed that the number of variant regions M is known and we applied the principle of maximum likelihood (ML) to estimate the variant regions breakpoints and their corresponding intensity levels. However, the number of variant regions of the DCN data is unknown a priori and need to be estimated. For this we apply a technique termed the minimum description principle (MDL) presented by [3] to estimate the number of variant regions before we apply the ML rule to estimate the unknown parameters.

$$MDL(k) = -\ln \prod_{i=1}^{k} p_i\left(y[\hat{n}_{i-1}], \ldots, y[\hat{n}_i - 1]; \hat{\theta}_i\right) + \frac{m_k}{2} \ln N, \quad (8)$$

where m_k is the number of estimated parameters or equivalently the dimensionality of the unknown parameters θ_i and $\left\{\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{k-1}, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\right\}$ is the MLE for the k-variant region of the entire data record for k=1, 2, …, K. Here K is the maximum number of variant regions chosen by the user. By definition $\hat{n}_0 = 0, \hat{n}_M = N.$ if the dimension of $\hat{\theta}_i$ is q_i, then

$$m_k = \sum_{i=1}^{k} q_i + k - 1, \quad (9)$$

where the first term represents the estimated $\hat{\theta}_i$ and the second term represents the k-1 breakpoints. The first term of (8) representing the negative log likelihood function for k variant region computed using DP as described in the previous section. For this, there is no extra computation required when the number of the variant regions is unknown.

4. RESULTS

In this section, we provide comparison study between our proposed algorithm and other recent techniques including our previously proposed algorithms, to detect the variation regions in the DCN data.

4.1. Simulated data

In this section, we provide two simulated example to compare the detection capabilities of the proposed methods. The first example is a comparison between the proposed methods based on the average of the root mean square errors (RMSEs) values of 100 simulated data sets generated randomly according to real data distributions using three different noise levels.

σ^2	MLE	NLDF	CBS	Wavelet	HMM
0.25	0.0421	0.0482	0.0513	0.0637	0.0852
0.50	0.0835	0.0881	0.0917	0.1051	0.1245
0.75	0.0945	0.1023	0.1120	0.2116	0.2355

Table 1. Comparison between the proposed methods based on RMSE's values.

From the results in Table 1, we can observe that the RMSEs values of MLE method outperform the proposed methods for detecting CNVs by 4 – 15%.

Figure 1 shows another example based on the receiver operating characteristic (ROC) curves for simulated data with σ^2=0.25.

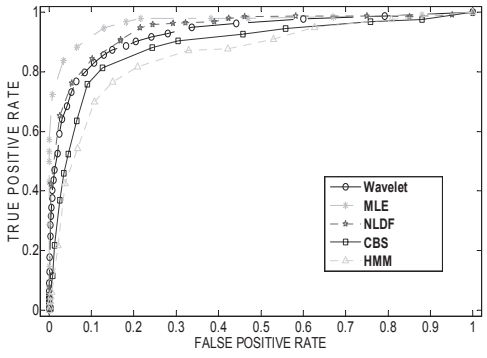

Figure 1. Receiver operating characteristic (ROC) curves for the proposed methods using simulated data.

Although the ROC curves of Figure 1 show that the MLE method is just slightly better than proposed methods, the detection performance decreases dramatically for these methods compared to MLE as the noise level increases.

4.2. Real data (Self-self experiment)

In this section, we present an experimental study to compare the performance of our proposed method, MLE, and CBS algorithm based on false positive rate. The same DNA sample is used as the test and reference. In other word, we compare the DNA sample with itself in the aCGH process to generate the DCN data as described in section 1. In the ideal case, the intensity level is $\log_2(T/R) = \log_2(1)=0$. However, due to the experimental noise, we expect to detect only one segment with relatively small intensity level value. Otherwise, the detected segments would be considered as false positives.

Chrm ID	Array 1		Array 2		Array 3	
	CBS	MLE	CBS	MLE	CBS	MLE
7	1	1	1	1	1	1
10	1	1	3	1	7	3
15	1	1	3	2	1	1
17	1	1	3	1	1	1
22	5	2	3	2	7	3

Table 2. Representation of the number of the detected CNVs using CBS and MLE methods in the three sample arrays for each chromosome.

In comparison with CBS, our purposed method, MLE, detects lower number of false positives as shown in Table 2. Details about the choice of chromosomes can be found in [11].

4.3. Validation

In this section, we examine a few CNVs predicted by both the segmentation software (CBS) provided by NimbleGen and the proposed algorithm and compare their ability to reliably report CNVs validated using the experimental molecular method quantitative polymerase chain reaction (QPCR) [12] in the lab. As shown in Table 2, Quantitative PCR was performed on a set of 6 samples, 3 normal controls (C1–C3) and 3 children with autism (A1–A3) using oligonucleotide array CGH along with the reference sample provided by [11] for the chromosome 7q and chromosome 10q segments for nucleotide positions (70061077–70061395) and (77927368–77927714), respectively.

Tested Region	Sample ID	CBS	NLDF	MLE	QPCR
Chromosome 7 70061077–70061395	A1	no change	no change	no change	no change
	A2	gain	gain	gain	gain
	A3	gain	gain	gain	gain
	C1	no change	no change	no change	no change
	C2	no change	gain	gain	gain
	C3	no change	no change	no change	no change
Chromosome 10 77927368–77927714	A1	loss	loss	loss	loss
	A2	loss	loss	loss	loss
	A3	loss	loss	loss	loss
	C1	loss	loss	loss	loss
	C2	no change	gain	gain	gain
	C3	no change	loss	loss	loss

Table 2. Comparison study of CBS, NLDF, and MLE algorithms for CNV detection, validated by QPCR.

In this example, the two tested regions that were determined by QPCR to be either deleted (loss) or duplicated (gain) in 6 samples, the CBS correctly predicted only 4 of the CNVs. The copy number gain and loss found in samples C2 and C3 was not predicted by the CBS but is readily predicted by examination of the NLDF and MLE methods. Moreover, the absolute mean-values of CNVs predicted by MLE are relatively higher than those predicted by NLDF.

5. CONCLUSIONS

In this paper, we investigated the performance characteristics of our proposed method. It is based on the Maximum likelihood principle to estimate the variant regions breakpoints and their corresponding intensity levels. The comparison study shows that our method achieves better detection capabilities by considering the global information from the entire data record. Dynamic Programming used to compute and reduce the computational complexity of the proposed method. The robustness of our method is due to the use of the Minimum description length procedure for estimating the number of variant regions automatically with no extra cost. Finally, the experimental molecular method QPCR confirms our results.

6. REFERENCES

[1] Venkatraman ES and Olshen AB. (2007). "A faster circular binary segmentation algorithm for the analysis of array CGH data". *Bioinformatics*. **23**:657–663.

[2] Fridlyand J., Snijders A., Pinkel, D., Albertson D. G. and Jain, A. N. Application of Hidden Markov Models to the analysis of the array CGH data. (Special Genomic Issue of Journal of Multivariate Analysis, June 2004, V. 90, pp. 132-153).

[3] S. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*, Prentice-Hall, 1993.

[4] S. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*, Prentice-Hall, 1998.

[5] R. E. Larson and J. L. Castie, *Principles of Dynamic Programming, vol. I, II*, Marcel Dekker Inc., NY, 1982.

[6] J. Rissanen, .Modeling by Shortest Data Description,. *Automatica*, vol. 14, 1978, pp.465-471.

[7] Pique-Regi R, Tsau ES,Ortega A, Seeger RC, Asgharzadeh S: "Wavelet footprints and sparse Bayesian learning for DNA copy number change analysis ", in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Hawaii 2007.

[8] A. Alqallaf, and A. Tewfik. (2007) "DNA Copy Number Detection and Sigma Filter". GENSIPS, pp.1–4 doi:10.1109/GENSIPS.2007.4365825.

[9] Alqallaf, A. K. and Tewfik, A. H. Selleck, S.B. Johnson, R. "Framework for the analysis of genetic variations across multiple DNA copy number samples" in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Las Vegas 2008 Page(s): 553-556. doi: 10.1109/ICASSP.2008.4517669.

[10] Wang, Y. and Wang S. "A novel stationary wavelet denoising algorithm for array-based DNA copy number data", *Int. J. Bioinformatics Research and Applications* 2007, Vol. 3, No. 2, pp. 206–222.

[11] J. Balciuniene, N. Feng, K. Iyadurai, B. Hirsch, L. Charnas, B. R. Bill, M. C. Easterday, J. Staaf, L. Oseth, D. Czapansky-Beilman, D. Avramopoulos, G. H. Thomas, A.. Borg, D. Valle, L. A. Schimmenti, and S. B. Selleck "Recurrent 10q22-q23 Deletions: A Genomic Disorder on 10q Associated with Cognitive and Behavioral Abnormalities" *Am. J. Hum. Genet.* 2007;80:938–947. doi: 10.1086/513607.

[12] Weksberg R, et al. "A method for accurate detection of genomic microdeletions using real-time quantitative PCR." BMC Genomics. 2005 Dec 13; 6:180.

[13] Aguirre, A.J., Brenman, C., Bailey, G., Sinha, R., Feng, B., Leo, C. "High-resolution characterization of the pancreatic adenocarcinoma genome". 2004 PNAS 101, 9067-9072.