FRAMEWORK FOR THE ANALYSIS OF GENETIC VARIATIONS ACROSS MULTIPLE DNA COPY NUMBER SAMPLES

Abdullah K. Alqallaf[1], *Ahmed H. Tewfik*[1], *Scott B. Selleck*[2], *and Rebecca Johnson*[2]

[1] Department of Electrical and Computer Engineering, University of Minnesota
[2] Department of Genetics, Cell Biology and Development, University of Minnesota
200 Union Street. SE, Minneapolis, MN 55455, USA
alqal001@umn.edu, tewfik@umn.edu, selle011, and john6582@umn.edu

ABSTRACT

Genetic diseases are characterized by the presence of genetic variations. These variations can be described in the form of copy number. Microrray-based Comparative Genomic Hybridization is a high-resolution technique used to measure copy number variations. However, the observed copy numbers are corrupted by noise, making variations breakpoints hard to detect. In this paper, we provide a framework for the analysis of copy number. The first part of the framework uses an extended version of nonlinear diffusion filter as pre-processing technique to denoise the observed data base. The extension accounts for the nonuniform physical distance between probes. The second part uses estimates the relative frequency of local and global genomic variations across multiple samples to identify statistically and biologically significant variations. For evaluation, we provide copy number variations results using simulated and real data samples. We also validate the predicted copy number variation segments of copy number gain and copy number loss using the experimental molecular tests quantitative polymerase chain reaction and show that our proposed approach is superior to popular commercial software.

Index Terms— Copy number variations, Comparative Genomic Hybridization, Smoothing, Edge-preserving, multiple samples.

1. INTRODUCTION

Copy number variations (CNVs) such as deletions and duplications are associated with the development and progression of many genomic diseases such as autism spectrum disorders (ASD). Understanding which genes or genomic locations are involved in the disease development, progression and maintenance in addition to the characterization of genomic disorders in developmental abnormalities, will lead to a better understanding of these complex human diseases as well as identify targets for therapeutic involvement. Oligonucleotide array CGH platform (NimbleGen technology) is an experimental approach for genome-wide scanning of differences in DNA copy numbers (DCN). It provides the capacity to detect copy number differences within LCRs. Unfortunately, these experiments contain many sources of errors due to human factors, array printer performance, labeling, and hybridization efficiency [4]. Due to these errors the observed copy numbers corrupted by noise, making the breakpoints of variation regions hard to detect. Therefore, one should consider denoising the data as a pre-processing step to uncover the true DCN changes before drawing inferences on the patterns of variations in the data samples. Various approaches had been proposed to uncover the true genetic variations. We review some of these techniques in the next section.

In this study, we provide a framework for the analysis of copy number datasets. It consists of two parts: (1) a preprocessing algorithm, and (2) a statistical search model based on the relative frequencies of the variations. The preprocessing algorithm is based on a semi-implicit nonlinear diffusion filter [13,14]. We extend the algorithm to consider the effect of nonuniform physical distance between the probes in the copy number datasets. The statistical analysis method locates and classifies the common variations within the affected samples that share the same variation type with respect to the normal variations in the control samples. This is done to further characterize the rearrangements in previously reported samples and suggest an additional set of genes that may be involved in the disease.

The rest of this paper is structured as follows: Prior work is presented in section 2. In section 3, we introduce extended version of nonlinear diffusion filter. Section 4 is devoted for statistical models searching for common variations across multiple samples. In section 5, we examine a few CNVs predicted by the proposed algorithms and compared their ability to reliably report CNVs validated using the experimental molecular method quantitative polymerase chain reaction (QPCR). Finally, Conclusions based on the observed results are provided in section 6.

2. PRIOR WORK

Generally, Copy Number variations (CNVs) detection techniques fall into two categories: statistical model based approaches and smoothing techniques. In the statistical model based algorithms, the noise free signal and noise models are required. Unfortunately, these models are usually unknown or impossible to describe adequately with simple random processes. In addition, the techniques are computationally costly. Examples of recent and efficient statistical approaches are the lookAhead algorithm [1], the Hidden Markov Models (HMMs) [2], and the CGH segmentation [5]. On the other hand, the smoothing techniques provide an alternative method for processing the DCN data that are characterized by small and long intervals with of sharp transitions and singularities at edges. The techniques are particularly suitable for denoising DCN

data as they do not require a parametric model in finding structures in the data. The main advantage of these techniques is their computational efficiency. Examples of efficient smoothing techniques are the one-dimensional Discrete Wavelet-based methods [3] and the Wavelet footprints [8]. Here, we present the extended version of nonlinear diffusion filter (*NLDF*) in the next section as local selective smoothing technique to denoise the DCN data. Also, we provide comparison study with our previously proposed algorithm based on Sigma filter [6,10,11].

3. METHODS AND MATERIALS

Nonlinear diffusion filter is an efficient and unconditionally stable algorithm. It had been presented in [13,14] based on semi-implicit scheme for discretizing diffusion equation. The filter operation is practically performed by numerically solving the continuous nonlinear partial differential equation (PDE).

3.1. One-dimensional Nonlinear Diffusion Filter

A good model for describing DNA copy number data is:

$$y[i] = f[i] + \varepsilon_i. \qquad i = 1, 2, \ldots, N \tag{1}$$

where $y[i]$ and $f[i]$ are the observed and true intensities of the DCN data probe at i^{th} location along the x-axis respectively. The ε_i represent independent identically distributed (*i.i.d.*) random variable from the Gaussian distribution of zero mean and σ^2 variance. The main idea of the original NLDF algorithm proposed by [13] is to determine a set of smoothed versions u(x, t) of the observed noisy signal y as an approximation of the true signal. Let

$$\partial_t u = \mathrm{div}[g(\|\nabla u\|)\ \nabla u] \tag{2}$$

where div is the divergence operator, ∇ is the gradient operator, $\|\nabla u\|$ is the gradient magnitude, and $g(\|\nabla u\|)$ is an edge-stopping function (diffusivity). Here we use Weickert edge-stopping function presented in [13]

$$g(|s|) = \begin{cases} 1 & |s| < 0, \\ 1 - \exp\left(\dfrac{-3.315}{(|s|/\lambda)^4}\right) & |s| \geq 0. \end{cases} \tag{3}$$

where λ is a positive constant that play role of a contrast parameter. The main purpose of the edge-stopping function g(s) is to control the diffusion process. To ensure the sharpness of the breakpoints in the smoothed version of the original signal, it should be a non-negative decreasing function. It is chosen to satisfy $g(s) \to 0$ as $s \to \infty$ so that the diffusion is stopped across edges.

As in [13,14], we linearly approximated the gradient as intensity differences, $\nabla u = u_j - u_i$, between the processing intensity i and its neighboring intensity $j \in W(i)$, where $W(i)$ denotes the spatial set of the neighbors of intensity i. We discretized the diffusion equation using semi-implicit scheme as follows:

$$\frac{u_i^{k+1} - u_i^k}{\tau} = \sum_{j \in W(i)} \frac{g_{i,j}^k}{\Delta x^2}(u_i^k - u_j^k). \tag{4}$$

Here, k denotes the discrete time steps (iterations) and $g_{i,j}$ is the diffusivity belong to the connection between intensities i and j. The constant τ is a positive scalar that determines the diffusion rate and Δx is the grid size. The matrix form of the semi-implicit scheme can be written as:

$$\begin{aligned} \frac{u^{k+1} - u^k}{\tau} &= A(u^k)u^{k+1}, \\ u^{k+1} &= (I - \tau A(u^k))^{-1}u^k, \\ u^{k+1} &= B(u^k)u^k. \end{aligned} \tag{5}$$

Here, $A(u^k) = (\alpha_{i,j}(u^k))$

$$\alpha_{i,j}(u^k) = \begin{cases} \dfrac{g_{i,j}^k}{\Delta x^2} & for\ j \in W(i), \\ -\displaystyle\sum_{n \in W(i)} \frac{g_{i,n}^k}{\Delta x^2} & for\ j = i, \\ 0 & else. \end{cases} \tag{6}$$

Note that A is a tridiagonal invertible matrix with $\alpha_{i,j} \neq 0$ for i=j and I is the identity matrix. For further details and proofs see [13,14].

The one-dimensional semi-implicit scheme of *Nonlinear Diffusion filter* procedure is as follows:

1- Start with the noisy signal y as initial condition: $u^0 = y$.
2- Calculate A and B from (5).
3- Apply the Thomas (Gaussian elimination) algorithm to solve a set of linear equations in the form of tridiagonal system matrix in the form of $Bu=d$. It can be summarized as follows:

   a) Decompose B into the product of a lower bidiagonal matrix L and an upper bidiagonal matrix, respectively.
   b) Solve $Lv=d$ for y by forward substitution.
   c) Solve $Ru=v$ by backward substitution.

The scheme is computationally efficient. It requires ($5N$-4) multiplications/divisions, and ($3N$-3) addition/subtractions. Hence the algorithm is linear in N. It is stable for every strictly diagonally dominant system matrix.

3.2. Extended Nonlinear Diffusion Filter

Most prior works considered the DNA copy number profiles as discrete signals under the assumption that the probes are uniformly distributed along the chromosomes. This assumption may lead to wrong decisions with false positive or/and false negative points.

In this section, we consider the nonuniform physical distance between the probes and demonstrate the performance of the extended algorithm using simulated and real data examples.
Hence, we remodeled the DCN data model discussed in the previous section as nonuniformly distributed discrete signals as follows:

$$y[x_i] = f[x_i] + \varepsilon_i. \qquad i = 1, 2, \ldots, N \tag{7}$$

where x_i in this case is the nonuniform distributed probe at i^{th} location along the x-axis. The x_i's are not uniformly distributed and the distance between two adjacent probes x_i and x_{i+1} may vary randomly. The $y[x_i]$ and $f[x_i]$ are the observed and true intensities of the DCN data probe location x_i respectively. The ε_i represent *i.i.d.* random variable from the Gaussian distribution of zero mean and σ^2 variance. The suggested procedure to solve the spacing distance effect can be summarized as follows:

1. Insert artificial markers between the original probes based on the average distance of the adjacent probes.

$$d_{avg} = \frac{\sum_{i=1}^{N-1}(x_{i+1} - x_i)}{N-1}, \tag{8}$$

2. Apply the nearest neighbor interpolation to create the artificial probes.
3. Apply the original nonlinear diffusion filter steps.

3.3. Performance

As shown in Figure 1, the receiver operating characteristic (ROC) curves for MDA-MB-453 data sample of *Coriel cell lines* demonstrate that our previously proposed Sigma filter based algorithm [11] provides superior performance to denoise the DCN data compared to other efficient proposed techniques such as LookAhead algorithm [1], wavelet-based [3], CGH segmentation [5] and HMM [2]. It also has lower computational complexity of O(N) compared to the statistical approaches, LookAhead algorithm O($N^{1.5}$), and CGH segmentation O(N^2) . We therefore use it in our comparison study with the semi-implicit nonlinear diffusion filter (*NLDF*). Table 1 presents a comparison study between the Sigma filter and *NLDF* based on the average of the root mean square error (RMSEs) values of 300 simulated data sets generated randomly according to real data distributions at different noise levels. See [7] for details. The results of the average RMSE values in Table 1 show that on average the *NLDF* outperforms the *Sigma filter*. In addition, the *NLDF* that considers the nonuniform spacing effect between probes achieved better performance than the *NLDF* that does not consider that effect. Figure 2 shows that the Nonlinear Diffusion Filter (*NLDF*) has better performance compared to Sigma filter for different noise levels in simulated data. The *NLDF* ROC curve rides on the top of *Sigma filter* curve. The superior performance of the *NLDF* is due to the selective smoothing ability of the edge-stopping function.

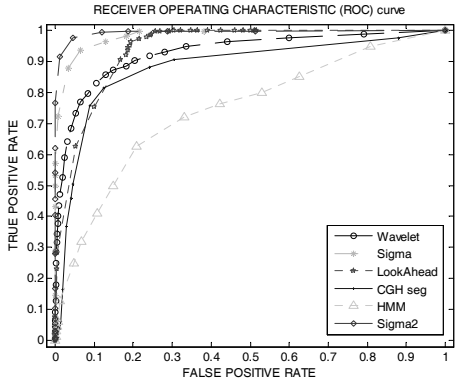

Figure 1. Receiver operating characteristic (ROC) curves for MDA-MB-453 data sample of Coriel cell line of *Sigma filter*, *Wavelet*, *LookAhead*, *CGH segmentation, HMM*, and *Extended Sigma filter* (*Sigma2*) algorithms.

σ	Sigma filter	Ext-Sigma filter	NLDF	Ext-NLDF
0.1	0.0408	0.0312	0.0248	0.0201
0.3	0.0641	0.0507	0.0445	0.0384
0.5	0.1090	0.0853	0.0893	0.0721

Table 1. The average of root mean square errors (RMSE's) values of 300 simulated data samples with 3 different noise levels using *Sigma filter* and *NLDF*.

4. STATISTICAL ANALYSIS MODELS

After filtering the DCN samples, we need to apply a statistical analysis to characterize the randomness of the copy variations and classify the genes involved in the targeted diseases. Here, we discuss an approach to resolve the challenge encountered in [9] while mapping the true breakpoints across multiple samples and especially in the complex *LCRs* regions. We use two statistical scores, the Global score and the Interval score to measure the significance of common variations across the sample at each probe location and interval probes locations, respectively.

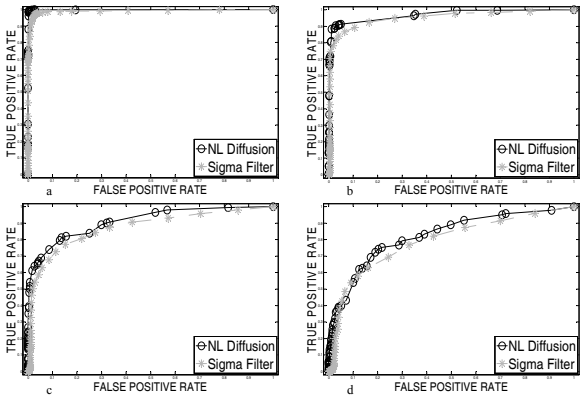

Figure 2. Receiver operating characteristic (ROC) curves for Sigma filter and nonlinear diffusion filter using simulated data with different noise level. a) 10%, b) 20%, c) 30%, and d) 50% of the original signal.

4.1. Global score

The global score (G-score) is the number of variations of the same type (gains or losses) that have occurred at each genomic location according to the thresholds provided by [9]. Suppose that a set of M filtered DCN samples each with N probes is represented in a matrix form $\hat{\boldsymbol{Y}}$. Given a column vector of the filtered DCN data of the same type $\hat{\boldsymbol{y}}_{\boldsymbol{n}}=\{\hat{y}_{s,n}\}$ at position n, we have

$$G[n]=\frac{\sum_{s\in M} v_{s,n}}{M}. \quad n=1,2,\ldots,N \qquad (9)$$

Here, s represents the samples of the variation of the same type and $v_{s,n}$ is a binary number equal to 1 if the variation is present and 0 otherwise. The score $G[n]$ will indicate whether the genomic location n contains a significant variation of the same type within the samples M. A higher the G-score indicates a higher confidence associated to a decision made at a given location. For simplicity, we consider only 2 types of variations as it illustrated in Figure 3.

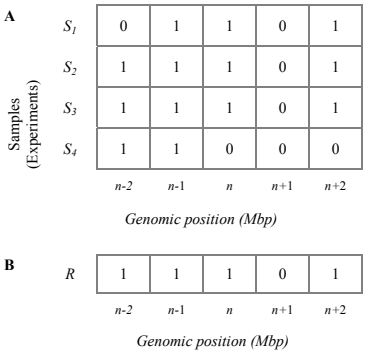

Figure 3. A) Illustration of the statistical analysis across multiple samples for a given variation interval. B) Display of the statistical method resultant sample (R) for the corresponding variant interval of the given samples.

4.2. Interval score (Class discovery)

The statistical score searching for common genomic intervals across the filtered samples is similar to searching globally for

common variations as described above. However, instead of measuring the number of variant probes *s* at each genomic location, here we measure the number of common variant intervals *C* across the samples that share the same type of variations (class). Particularly, we are looking for *I* and *C* that maximize the Global-score, where *I* is a continuous interval of genomic locations that share the same type of variations. Specifically,

$$G[C, I] = \frac{\sum_{m \in C} \sum_{n \in I} v_{m,n}}{M}. \quad (10)$$

Our initial set of 29 filtered data samples from 14 controls and 15 children with autism indicates that some CNVs have a significantly higher frequency than others. Even with such a small sample size we are already observing statistically significant CNVs associated with autism. For example, CNV2 on chromosome 15 (a deletion) was represented in 8/15 children with autism as compared to only 1/14 controls.

4.3. Statistical Significance

To date, some attention was given to search for significance of CNVs (deletions or duplications) that overlap across multiple case samples with respect to control samples as null model [1,11]. In this paper, we apply the *t*-statistic test and assign *p*-value to each genomic location by using a multiple testing corrected permutation approach. Probes with $p<0.05$ were termed significant. Results of the comparison study based on the calculation of *p*-value are not shown due to space limitation.

5. VALIDATION USING QUANTITATIVE POLYMERASE CHAIN REACTION (QPCR) METHODS

In this section, we examine a few CNVs predicted by both the segmentation software 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 laboratory. As shown in Table 2, Quantitative PCR was performed on a set of 6 samples, 3 normal controls (C1 - C3) and 3 (A1 - A3) children with autism using oligonucleotide array CGH along with the reference sample for the chromosome 7q and chromosome 10q segments for nucleotide positions (70061077- 70061395) and (77927368- 77927714), respectively.

Tested Region	Sample ID	Segment Analysis	Sigma Filter	NLDF	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
Chromosome10 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 *segmentation analysis*, *Sigma filter*, and *NLDF* algorithms for CNV detection, validated by QPCR. Sample IDs with "A" are autistic individuals and with "C" are control individuals, respectively.

Table 2 shows that within the two tested regions that were determined by QPCR to be either deleted (loss) or duplicated (gain) in 6 samples, the segmentation analysis correctly predicted only 4 of the CNVs. The copy number gain and loss found in samples C2 and C3 was not predicted by the segmentation but is readily predicted by examination of the *NL* diffusion filtered data.

6. CONCLUSIONS

In this paper, we proposed a new algorithm for detecting copy number variations. The algorithm is based on nonlinear diffusion filtering. We demonstrated its superior performance using real and synthetic data. This superior performance is due to the discriminative characteristics governed of the edge-stopping function that the nonlinear diffusion filter uses. Even better performance can be achieved by considering the effect of the nonuniform physical distance between the probes of the DCN samples. To characterize the randomness that incentives, we preformed a statistical analysis on the filtered samples searching for common variations that occur with high frequency to provide insight into the patterns of these variations. Finally, the experimental molecular method QPCR confirms the superior performance of *NLDF*.

7. REFERENCES

[1] Doron L., Yonatan A., Amir B., Nathan L., and Zohar Y. (2006). Efficient Calculation of Interval Scores for DNA Copy Number Data Analysis. Journal of Computational Biology, Pp. 215-228.

[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] David Donho et al, WAVELAB 802, "http://www.stat.stanford.edu/~wavelab/".

[4] Churchill GA. Fundamentals of experimental design for cDNA microarray. Nat Genet Sup. 2002;32:490. doi:10.1038/ng1031.

[5] Franck P., Stephane R. Marc L., Christian V., and Jean-Jacques D. A statistical approach for array CGH data analysis. BMC Bioinformatics 2005, 6:27 doi:10.1186/147-2105-6-27.

[6] J.S. Lee, "Digital Image Smoothing and the Sigma Filter," CVGIP, Vol 24, pp. 255-269, 1983.

[7] Wang, Y. and Wang S. (2007) `A novel stationary wavelet denoising algorithm for array-based DNA copy number data', Int. J. Bioinformatics Research and Applications, available at "http://engr.smu.edu/~yuhangw/papers/ waveletdenoising_dcn.pdf".

[8] 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,

[9] 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.

[10] A. Alqallaf, and A. Tewfik. "DNA COPY NUMBER DETECTION AND SIGMA FILTER" GENSIPS 2007.

[11] Alqallaf, Abdullah K. and Tewfik, Ahmed H. "Signal Processing Techniques and Statistics for the Analysis of Human Genome Associated with Behavior Abnormalities" Statistical Signal Processing, 2007. SSP apos;07. IEEE/SP 14th Workshop on Volume, Issue, 26-29 Aug. 2007 Page(s):36 – 38 DOI: 10.1109/SSP.2007.4301213.

[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] P. Perona and J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 7, p.629, 1990.

[14] J. Weickert, B.M. ter Haar Romeny, and M. Viergever, "Efficient and Reliable Schemes for Nonlinear Diffusion Filtering," *IEEE Transactions on Image Processing*, Vol. 7, No. 3, p.398, 1998.