BAYESIAN DETECTION OF RECURRENT COPY NUMBER ALTERATIONS ACROSS MULTIPLE ARRAY SAMPLES

Roger Pique-Regi[1,2,*], *Jordi Monso-Varona*[1], *Antonio Ortega*[1], *Shahab Asgharzadeh*[2]

[1]Signal and Image Processing Institute, Viterbi School of Engineering,
University of Southern California; [2]Department of Pediatrics,
Childrens Hospital of Los Angeles, University of Southern California
*rpique@ieee.org

ABSTRACT

Copy number alterations (CNA) affecting small portions of chromosomes are difficult to identify. Advances in microarray technology now allow very high resolution scans of large cohorts of samples but at the price of severe noise degradation. Our proposed genome alteration detection algorithm (GADA) has been shown to be a highly accurate and efficient approach to analyze a single array sample. In this paper, the sparse Bayesian learning (SBL) used in GADA is extended to find CNA on multiple samples that share breakpoint positions but may have different magnitude of alteration. Our model is especially well suited to analyze sample replicates, i.e., multiple arrays from the same specimen. Our results show that replicates greatly improve the accuracy and robustness in detection. In some cases, a single replicate sample offers an accuracy equivalent to a 2-fold increase in the signal to noise ratio, while reducing by up to a 50% the detection of false CNA caused by outliers. The computational cost of the algorithm is essentially linear $\mathcal{O}(NM)$ in the number of the microarray probes M and samples N. In conclusion, the multiple sample GADA (N-GADA) presented here appears to be a promising tool for finely locating small CNAs that are shared across multiple samples.

1. INTRODUCTION

Copy number alterations (CNA) represent deviations from the normal number of DNA copies generally found in the genome of some organism (e.g., two for diploid cells). In humans, these alterations are known to be present in both normal and diseased cells. Examples include: chromosome 21 trisomy in Down's syndrome, amplification of MYCN proto-oncogene in Neuroblastoma, and loss of RB tumor repressor in Retinoblastoma. Recent advances in the microarray technology enabling high resolution genomic scans of large cohort of individuals have revealed presence of short CNAs that are repeated across normal genomes (i.e., polymorphic CNAs) [1] constituting a completely new source of unstudied natural genetic variation. Small alterations are the most difficult to detect and the ones most likely to lead to false detections because of severe noise degradation. A joint analysis of many samples would undoubtedly increase the performance in detecting small CNAs, but nearly all currently available algorithms only analyze one sample at a time.

In previous work [2, 3] we developed a copy number detection approach called GADA (genome alteration detection algorithm) that achieved excellent performance in single-sample CNA detection. Compared to other state-of-the-art methods, using standard evaluation datasets and benchmarks [4], GADA obtained the highest accuracy and was at least 100 times faster. GADA is based on a compact linear algebra representation of the array probe intensities as a piece-wise constant (PWC) vector and makes use of a two step detection approach. In the first step, sparse Bayesian learning (SBL [5, 6]) identifies all potentially interesting breakpoints that delimitate the CNA. The second step uses a backward elimination (BE) procedure to statistically rank the identified breakpoints, allowing a flexible control of the false discovery rate (FDR).

In this paper we extend GADA to detect CNA across multiple samples (N-GADA). The method is especially suited to detect CNAs from sample replicates, since the underlying breakpoint locations should be the same, but the mean magnitude of the array probe measurements may be different. These differences may be due to sample contamination, amount of material, or other uncontrolled effects that cannot be corrected. Compared to the large number of algorithms proposed for single-sample CNA analysis, there are very few approaches dealing with the multiple sample problem [7, 8, 9, 10]. Two of them [7, 8] are post-processing techniques to refine the results obtained by a given single-sample algorithm and do not propose a joint model. The other two approaches [9, 10] propose models that only encourage overlap among CNAs across samples. In contrast, our approach is unique in the sense that it encourages recurrent breakpoint positions. More precisely, the SBL hierarchical prior is modified to encourage the selection of breakpoints delimiting CNA at similar positions across the samples under analysis. We hypothesize that this may be a more powerful model when there is underlying evidence that the alterations start and end at recurrent positions, as it is the case of sample replicates and possibly of CNA polymorphisms. In order to evaluate N-GADA we used simulation and real datasets of pairs of replicate samples with the same underlying copy

1-4244-2372-9/08/$20.00 ©2008 IEEE

Authorized licensed use limited to: IEEE Xplore. Downloaded on November 26, 2008 at 14:39 from IEEE Xplore. Restrictions apply.

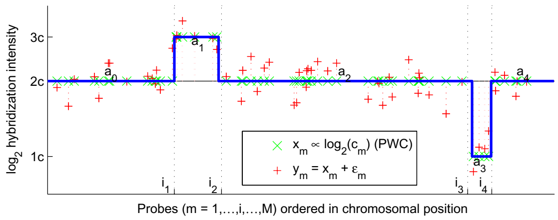

Figure 1. Graphical representation of the observation model (1) using a chromosome section with 2 alterations as an example. The underlying mean hybridization intensity x_m is piece-wise constant (PWC) and discrete valued depending on the number of DNA copies. The observed hybridization intensities y_m do not follow this expected behavior due to degradation by hybridization noise ϵ_m.

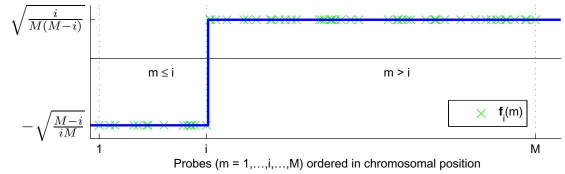

Figure 2. Step vector f_i with a breakpoint between probe i and $i+1$. The step vectors have been normalized to have unit norm, $\sum_{m=1}^{M} (f_i(m))^2 = 1$, and average zero for $i > 0$, $\sum_{m=1}^{M} (f_i(m)) = 0$.

number profile. Our results show that replicates greatly improve the accuracy and robustness of detection while maintaining a very good computational efficiency.

The paper is structured as follows. The extended N-GADA approach and its implementation are presented in Section 2. Section 3 is devoted to presenting the results, and conclusions are discussed in Section 4.

2. N-GADA FOR MULTIPLE SAMPLES

In this section we extend the GADA approach [3] so that it can handle multiple samples. First, we review the PWC representation for genome CNA we introduced in [2], which is a maximally sparse representation in terms of the number of breakpoints. Second, we extend the SBL hierarchical prior to model sparse breakpoints occurring at similar locations across multiple samples; and we briefly describe how to efficiently fit the resulting model using the EM algorithm [11]. Finally, we detail the new multiple sample implementation of the BE procedure to control for the false discovery rate (FDR).

2.1. PWC representation

Most CNA detection algorithms model microarray measurements as follows:

$$y_m = x_m + \epsilon_m, \tag{1}$$

where y_m are the log-intensities of each probe m measured in the microarray, x_m represents the copy number effect, and ϵ_m a zero-mean hybridization noise. In Figure 1, we can observe that x_m is piece-wise constant (PWC) and discrete valued (DIS). These two characteristics are the consequence of every piece of the genome being represented in a cell by an integer number of DNA strands (usually two copies for the human autosome). Thus, the probe hybridization intensities y_m fluctuate around a mean value x_m that depends on the underlying number of DNA copies.

Using vector notation, the model of (1) can be written as:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon} = \boldsymbol{F}\boldsymbol{w} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{x}$ has been replaced by its representation in terms of the PWC basis, $\boldsymbol{Fw}$, where the columns of $\boldsymbol{F}$ are normalized step vectors $\boldsymbol{f}_i$ as in Figure 2. With this representation, any PWC vector $\boldsymbol{x}$ with K breakpoints ($\mathcal{I} = \{i_1, \ldots, i_K\}$) can be compactly represented by a linear combination of K step vectors $\boldsymbol{f}_i$ plus a constant vector $\boldsymbol{f}_0 = 1/\sqrt{M}(1, \ldots, 1)$. The number of copy number changes is very small compared to the number of probes, $K << M$, so we can can exploit these sparseness properties to infer the most likely copy number alterations.

2.2. Sparse Bayesian Learning for multiple samples

CNA detection can be formulated using SBL as the problem of finding the maximum a posteriori (MAP) estimate [3]:

$$\begin{aligned} \hat{\boldsymbol{w}}_{MAP} &= \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y}) = \arg\max_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w}) \\ &= \arg\min_{\boldsymbol{w}} -\log p(\boldsymbol{y}|\boldsymbol{w}) - \log p(\boldsymbol{w}) \end{aligned} \tag{3}$$

where the observation model $p(\boldsymbol{y}|\boldsymbol{w})$ specifies a goodness of fit measure and the prior distribution for the weights $p(\boldsymbol{w})$ specifies the sparseness constraints. Here, we extend our previously proposed model [3] to multiple samples. Assuming noise to be normal and independent across probes m and samples n, for a given underlying CNA profile for each sample, $\boldsymbol{x}^n = \boldsymbol{F}\boldsymbol{w}^n$, the observation model would be:

$$p\left(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^N | \boldsymbol{w}^1, \ldots, \boldsymbol{w}^N\right) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{F}\boldsymbol{w}^n, \sigma_n^2 \boldsymbol{I}\right) \tag{4}$$

and the prior distribution for the weights is specified as a hierarchical prior:

$$p\left(\boldsymbol{w}^1, \ldots, \boldsymbol{w}^N | \boldsymbol{\alpha}\right) = \prod_{n=1}^{N} \prod_{m=1}^{M-1} \mathcal{N}\left(w_m^n | 0, \alpha_m^{-1}\right) \tag{5}$$

where $\boldsymbol{\alpha}$ is a vector of hyperparameters that are distributed according to a gamma distribution:

$$p(\boldsymbol{\alpha}) = \prod_{m=1}^{M-1} \Gamma(\alpha_m | a, b). \tag{6}$$

Notice that here the $\boldsymbol{\alpha}$ hyperparameters are shared across multiple samples. This is in contrast to the application of SBL in 1-GADA, which implies that a different set of a hyperparameters is used for each sample. The role of the hyperparameter α_m is to control the likelihood of the presence of a breakpoint at a particular position of the genome

Authorized licensed use limited to: IEEE Xplore. Downloaded on November 26, 2008 at 14:39 from IEEE Xplore. Restrictions apply.

but without imposing any restriction on the actual magnitude of the breakpoint w_m^n and its corresponding CNA.

The mathematical procedures to fit this multiple sample model and to infer the CNA breakpoints are basically the same as in 1-GADA [3]. We also use the EM algorithm, exploiting the conjugacy properties between the gamma and normal distributions, as well as the properties of our PWC representation (i.e., the matrix structure for $\boldsymbol{F}$). The E-step is the same as before but repeated for each of the samples; i.e., finding the posterior distribution given the hyperparameters and the observation:

$$p\left(\boldsymbol{w}^n|\boldsymbol{y}^n, \boldsymbol{\alpha}, \sigma_n^2\right) = \mathcal{N}\left(\boldsymbol{w}^n|\boldsymbol{\mu}^n, \boldsymbol{\Sigma}_n\right) \quad (7)$$

$$\boldsymbol{\Sigma}_n = \left(\sigma_n^{-2}\boldsymbol{F}^t\boldsymbol{F} + \mathrm{diag}(\boldsymbol{\alpha})\right)^{-1} \quad (8)$$

$$\boldsymbol{\mu}_n = \sigma_n^{-2}\boldsymbol{\Sigma}_n\boldsymbol{F}^t\boldsymbol{y}_n \quad (9)$$

The M-step, on the other hand, takes all the samples into account in computing the $\boldsymbol{\alpha}$ hyperparameters:

$$\hat{\alpha}_m = \frac{2a + N}{\sum_n \left(\Sigma_{mm}^n + (\mu_m^n)^2\right) + 2b} \quad (10)$$

The EM algorithm requires very few iterations to converge in our experiments; and all required operations in each iteration can be performed in a linear number of steps $\mathcal{O}(NM)$. This is clear for the M-step, and we already demonstrated in [3] that the operations required to compute $\boldsymbol{\mu}$ (9) and the diagonal of $\boldsymbol{\Sigma}$ (8) is $\mathcal{O}(M)$ for each sample, since we can exploit the fact that $(\boldsymbol{F}^t\boldsymbol{F})^{-1}$ is a tridiagonal matrix.

2.3. Backward Elimination for multiple samples

In our previous work [3] the statistical significances of breakpoints returned by SBL were ranked by a simple BE procedure using a standard linear regression model. Here, this is done within the SBL algorithm but taking into account the statistical evidence observed across multiple samples. For a single sample, both approaches are essentially equivalent; but the new approach can exploit better the information gathered by SBL about the multiple samples (i.e., the α parameters). In the new procedure, after the SBL has converged for the first time to a set of breakpoints with high sensitivity, each breakpoint is statistically scored as

$$t_m = \sqrt{\sum_n \frac{{\mu_m^n}^2}{\Sigma_{mm}^n}} \quad (11)$$

and the lowest scoring breakpoint is recursively eliminated from the model. Each elimination is carried out by setting $w_m = 0$ and repeating the EM algorithm described in Section 2.2. The sensitivity vs. FDR trade-off is controlled by stopping the procedure when all the remaining breakpoints have a score higher than a critical value T.

3. RESULTS

In this section we evaluate the proposed N-GADA algorithm for the case where $N = 2$ replicates are available, but results extend to other N. We employed the artificial dataset conceived by Willenbrock [4], which consists of 500 samples of 20 chromosomes with 100 probes where the underlying CNA are known and the noise is i.i.d. Gaussian. We generated the sample replicates using the same ground truth but with an independent new noise realization $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$, with uniformly distributed noise power $\sigma \sim U(0.1, 0.2)$ and tissue mixture $p \sim U(0.3, 0.7)$ parameters. These kind of simulations [4] may not reflect all possible scenarios, but constitute the most widely used method for quantitative evaluation.

These 2×500 samples are used to compare the performance of N-GADA to two other alternatives (Figure 3). The algorithms that combine both samples, i.e., 2-GADA and naive averaging, greatly improve the accuracy in breakpoint detection in comparison to the case in which no replicates are available (1-GADA). Roughly, a sample replicate would be equivalent to a two fold increase of the signal to noise ratio on a single sample. The results obtained by naive averaging are slightly worse than those of the 2-GADA approach; because the former assumes that breakpoints and segment reconstruction levels are the same while in the latter only the breakpoints are the same. On this simulation dataset, the reconstruction levels for each sample in the pair change depending on the tissue mixture parameter p.

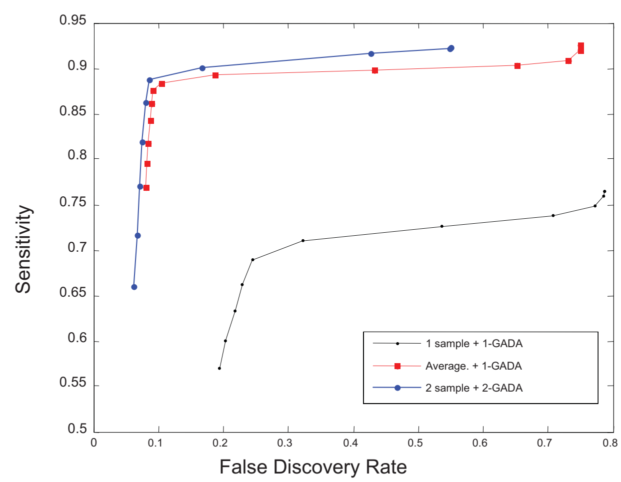

Figure 3. PROC operational curves for the mean sensitivity vs. FDR in detecting real copy number changes at their exact location. Black curve consist of applying 1-GADA to each of the two samples independently. Red curve combines the two samples by a weighted average into a single sample which is analyzed by 1-GADA. Blue curve is the proposed M-GADA approach. The benchmark metrics sensitivity and FDR are the same as originally defined in [4] in terms of CNA breakpoint detection.

In order to further assess the performance in terms of robustness, we randomly introduced single probe outliers (extreme values) in only one of the samples in each pair in a simulation dataset. Ideally, we would like to avoid false detections that are only supported by one of the samples. The single-sample algorithm and the one based on sample

Authorized licensed use limited to: IEEE Xplore. Downloaded on November 26, 2008 at 14:39 from IEEE Xplore. Restrictions apply.

averaging cannot distinguish these outliers and nearly all of them will cause false detection. On the other hand, 2-GADA reduces false detection caused by these outliers by about 50%.

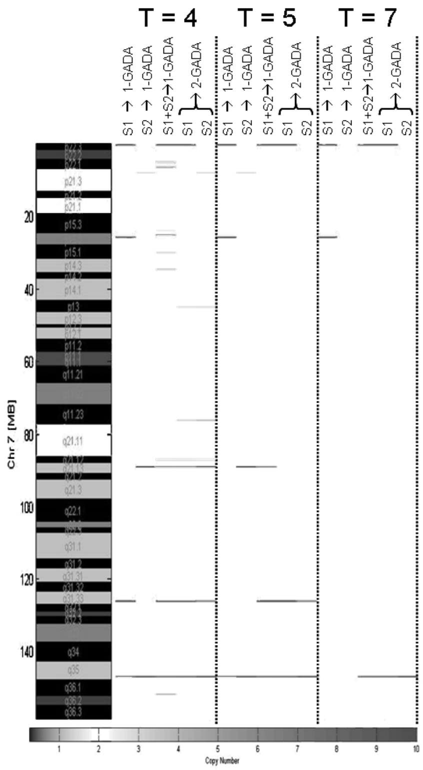

Figure 4. Visual representation of the detected CNA using different algorithms and settings (columns) on two replicates (S1 and S2) of a normal human sample (NA01416) analyzed using Affymetrix 500K (Nsp) platform. Columns are divided into three sections, each representing a different threshold T used for CNA detection. In each group, the first two columns correspond to the independent analysis of S1 and S2 using 1-GADA, the third column is the result of applying 1-GADA to the S1 and S2 weighted average, and the last two columns in each group (4th and 5th) are the outputs corresponding to S1 and S2 resulting of the 2-GADA joint analysis. For each claimed CNA, red tones represent amplification and blue tones loss of genetic material.

Our results on real data are also in accordance to the findings obtained using simulation data. Figure 4 shows a visual representation of some of the CNA detected on 3 different FDR operating points (T settings) for a pair of replicate samples (S1, S2) analyzed with Affymetrix 500K platform. The CNA found are very short segments because the samples are from a healthy human subject (NA01416). We can observe the higher sensitivity of the 2-GADA approach on the deletion on q35; the CNA is retained for a higher significance setting $T = 7$ while it is removed on the single-sample approaches. This higher sensitivity can also be achieved by the sample averaging procedure, but this naive combination may cause more spurious false CNA (see 3rd column, $T = 4$). On the other hand, the 2-GADA approach is more robust since it retains the information of the origin of each observation. This can also be seen on an S2 outlier in q21.13 $T = 5$; 2-GADA eliminates this false alteration since it is not supported on (S1) one of the two samples, while in naive averaging this outlier causes a false detection. In terms of computational speed, the 2-GADA approach performance is very competitive, with computational complexity linear in the number of probes M and samples N.

4. CONCLUSION

This paper presents a novel approach N-GADA to solve the problem of finding CNA with breakpoints at recurrent positions across multiple samples. N-GADA extends the single-sample algorithm GADA presented in [3] using a Bayes hierarchical prior for the breakpoints that is shared across all the samples. Simulation and real data results show that the proposed approach achieves a higher accuracy and robustness to outliers when sample replicates are available. The resulting approach retains a linear complexity in the number of samples and probes. Thus, the approach can be considered a promising tool to discover small alterations that are recurrent across many samples.

5. ACKNOWLEDGMENTS

This research has been supported in part by grants K12-CA60104 from the NIHs Child Health Research Career Development Award Program and the Pre-Institute Award from the Pediatric Brain Tumor Foundation.

6. REFERENCES

[1] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, et al., "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–54, 2006.

[2] R. Pique-Regi, E. S. Tsau, A. Ortega, R. C. Seeger, and S. Asgharzadeh, "Wavelet footprints and sparse bayesian learning for DNA copy number change analysis," in *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, April 2007.

[3] R. Pique-Regi, J. Monso-Varona, A. Ortega, R. Seeger, T. Triche, and S. Asgharzadeh, "Sparse representation and Bayesian detection of genome copy number alterations from microarray data," *Bioinformatics*, vol. Accepted, 2007.

[4] H. Willenbrock and J. Fridlyand, "A comparison study: applying segmentation to array CGH data for downstream analyses.," *Bioinformatics*, vol. 21, no. 22, pp. 4084–91, 2005.

[5] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[6] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE-Trans-SP*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[7] C. Rouveirol, N. Stransky, P. Hupe, P. L. Rosa, E. Viara, E. Barillot, and F. Radvanyi, "Computation of recurrent minimal genomic alterations from array-CGH data," *Bioinformatics*, vol. 22, no. 7, pp. 849–56, 2006.

[8] S. J. Diskin, T. Eck, J. Greshock, Y. P. Mosse, T. Naylor, J. Stoeckert, C. J., B. L. Weber, J. M. Maris, and G. R. Grant, "STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments," *Genome Res*, vol. 16, no. 9, pp. 1149–58, 2006.

[9] S. P. Shah, W. L. Lam, R. T. Ng, and K. P. Murphy, "Modeling recurrent DNA copy number alterations in array CGH data," *Bioinformatics*, vol. 23, no. 13, pp. i450–8, 2007.

[10] D. Lipson, Y. Aumann, A. Ben-Dor, N. Linial, and Z. Yakhini, "Efficient calculation of interval scores for DNA copy number data analysis.," *J Comput Biol*, vol. 13, no. 2, pp. 215–28, 2006.

[11] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley, 1997.

Authorized licensed use limited to: IEEE Xplore. Downloaded on November 26, 2008 at 14:39 from IEEE Xplore. Restrictions apply.