WAVELET FOOTPRINTS AND SPARSE BAYESIAN LEARNING FOR DNA COPY NUMBER CHANGE ANALYSIS

Roger Pique-Regi[1,2*], *En-Shuo Tsau*[1], *Antonio Ortega*[1], *Robert Seeger*[2], *Shahab Asgharzadeh*[2]

[1] Signal and Image Processing Institute, Department of Electrical Engineering,
Viterbi School of Engineering, University of Southern California;
[2] Division of Hematology - Oncology, Childrens Hospital Los Angeles, Department of Pediatrics,
Keck School of Medicine, University of Southern California

ABSTRACT

Alterations in the number of DNA copies are very common in tumor cells and may have a very important role in cancer development and progression. New array platforms provide means to analyze the copy number by comparing the hybridization intensities of thousands of DNA sections along the genome. However, detecting and locating the copy number changes from this data is a very challenging task due to the large amount of biological processes that affect hybridization and cannot be controlled. This paper proposes a new technique that exploits the key characteristic that the DNA copy number is piecewise-constant along the genome. First, wavelet footprints are used to obtain a basis for representing the DNA copy number that is maximally sparse in the number of copy number change points. Second, Sparse Bayesian Learning is applied to infer the copy number changes from noisy array probe intensities. Results demonstrate that Sparse Bayesian Learning has better performance than matching pursuits methods for this high coherence dictionary. Finally, our results are also shown to be very competitive in performance as compared to state-of-the-art methods for copy number detection.

Index Terms— DNA Copy Number, piece-wise constant, detection, denoising, sparse Bayesian learning.

1. INTRODUCTION

Normal human cells have two copies of nearly identical autosomal chromosomes and a pair of sex chromosomes. Thus, for the autosomal genome the DNA copy number is generally two, one copy inherited from each parent. Cancer cells often exhibit genetic aberrations in which chromosome sections may be lost (copy numbers 0 or 1), or replicated many times (copy number greater than 2). Detecting and locating these alterations and determining their functional effects is an essential foundation for improving diagnostic and therapeutic strategies [2].

These genetic material gains and losses can be detected by different methods. One of the first techniques was comparative genomic hybridization (CGH) [3], which basically consists of using a clone with a fluorescent tag that will hybridize specifically with a target section of the genome. A higher copy number will have a larger fluorescent intensity. Moreover, arrays containing thousands of these clones or probes can be used to perform a genome-wide analysis of copy number [4]. Additionally, the probe intensities from genotyping arrays can also be used for the same purpose [5].

Copy number changes correspond to physical losses/gains in genetic material, which typically cannot affect arbitrarily small segments in the genome, so that the actual copy number will be piecewise constant. The hybridization intensities measured with microarrays will be affected by different sources of noise that cannot be controlled in the experiment; e.g. clone fragment lengths, G-C content, cross-hybridization, and others. Therefore, copy number changes are perceived as a change on the statistics of the hybridization intensity which can be modeled as follows:

$$y_m = f(c_m) + \varepsilon_m \tag{1}$$

where y_m is the observed hybridization intensity at genome location m, ϵ_m is a random variable describing a zero mean hybridization noise, and $x_m = f(c_m)$ is the mean hybridization level that corresponds to copy number c_m.

The biological fact that the copy number c_m and hence x_m is piece-wise constant along the genome will be exploited in this paper to build a basis expansion using wavelet footprints [6]. The representation obtained is shown to be maximally sparse, meaning that no more footprints (basis vectors) than copy number changes present are necessary to represent the signal x_m. This representation is much more compact than a standard wavelet-based method, where one copy number change is represented by as many coefficients as levels of decomposition are used. In contrast, with the footprint approach all the discontinuity information is gathered in one footprint. To the best of our knowledge this is the first time that wavelet footprints have been proposed to represent genomic copy numbers.

The goal is to infer where the copy number change points are located, from noisy observed hybridization intensities y. Footprints become useful for our purposes because they provide a representation that will be *sparse* when, as is to be expected, the number of copy number changes is small. Thus, our goal will be to minimize the error in approximating the observed noisy signal using footprints, subject to the number of footprints (copy number changes) being no greater than a given k. Even though the error measure can be a simple quadratic function (e.g., mean square error MSE), the optimization problem is still combinatorial because of the sparsity constraints, so that $\binom{M}{k}$ solutions would have to be evaluated (where M is the length of genome under analysis), for each possible k.

When the error measure is quadratic, there exist several techniques developed to search overcomplete dictionaries that may be applied to solve this problem [7], namely, matching pursuits (MP) [8], orthogonal matching pursuits (OMP) [9], basis pursuit (BP) [7], method

*email: piquereg@usc.edu

The authors thank Dr. Hanni Willenbrock for providing the results of the other copy detection methods [1] .This study was supported in part by grant CA60104 from the National Cancer Institute (R. C. Seeger) and by Child Health Research Career Development Award (S. Asgharzadeh) through grant K12-HD-52954 from National Institute of Health.

of frames (MOF) [7], and sparse Bayesian learning (SBL) [10]. In practice some of them are severely limited due to the high coherence of the wavelet footprint basis, i.e., footprints that correspond to two contiguous discontinuities are highly collinear. Thus, the second new contribution of this paper is to show that sparse Bayesian learning does not exhibit this problem and thus is better suited for this type of dictionary.

Our evaluation is based on a benchmark dataset and performance metrics, proposed by [1] for this particular application, and simulates array-CGH observations with known copy number changes. Compared to other methods to solve this problem the performance of our proposed algorithm is comparable to the best one, DNAcopy [11], but the complexity of our method is significantly lower.

The paper is structured as follows. Section 2 introduces wavelet footprints and develops a basis expansion for piecewise constant signals. Section 3 formulates the copy number change detection problem as a denoising problem. Section 4 proposes SBL for estimating the footprint coefficients. Section 5 presents experimental results and compares to other existing techniques. Finally, the paper concludes summarizing the major contributions and discussing possible extensions.

2. WAVELET FOOTPRINT REPRESENTATION FOR PIECE-WISE CONSTANT SIGNALS

Wavelet footprints [6] have been proposed as a tool to design overcomplete dictionaries for representing piece-wise polynomial signals. The basic idea is that each kind of discontinuity in these signals produces a distinctive set of coefficients in the wavelet domain and that a signal can be characterized by these discontinuities. The footprint is a scale space vector formed by gathering all the wavelets that characterize a discontinuity of a particular kind and location. Then, this set of footprints form an overcomplete dictionary that leads to a representation for these signals that is significantly sparser than what is achievable with a standard wavelet representation.

For piece-wise constant signals and Haar wavelets [6], the wavelet footprint dictionary is formed by a set of vectors, where each vector f_k is a simple step function with one discontinuity between $k-1$ and k, $\sum_{m=1}^{M} f_k(m) = 0$, and $\| f_k \|^2 = 1$ for $k = 1, ..., (M-1)$.

We have used these properties to extend the dictionary in [6] to signals of arbitrary length M (not necessarily a power of 2):

$$f_k(m) = \begin{cases} -\sqrt{\frac{M-k}{kN}} & m < k \\ \sqrt{\frac{k}{M(M-k)}} & m \geq k \end{cases} \tag{2}$$

where $k = 1, ..., (M-1)$, $m = 0, ..., (M-1)$, and $f_0(m) = \frac{1}{\sqrt{M}}$ is defined to be the DC component.

It is easy to see that this dictionary is indeed a basis, since $F = [f_0, f_1, ..., f_{M-1}]$ is a square invertible matrix. With this notation the footprint representation of a piece-wise linear signal x can be compactly defined as:

$$x = Fw \quad w = F^{-1}x \tag{3}$$

One of the most appealing properties of this representation is that it is proved to be maximally sparse for piece-wise constant signals [6]. That is, for any given signal x with exactly K given discontinuities, only $K+1$ coefficients (w components different than 0) are required to perfectly reconstruct the signal. A simpler proof than the one in [6] is obtained by building first the dual basis, and showing that $w(m) = 0$ if and only if $x(m) - x(m-1) = 0$, and therefore there is a one-to-one mapping between the k-th w component and the jump of the discontinuity between $k-1$ and k.

3. DENOISING WITH WAVELET FOOTPRINTS

The compact representation developed in the previous section is very useful to estimate x from a degraded observation y generated as in (1). If only very few copy number changes $K << M$ are present, then $x = Fw$ has a very sparse representation in the wavelet footprint basis, while the noise ϵ spreads to all w components. Under this scenario, the problem is formulated as finding $\hat{x} = F\hat{w}$ that is closest to the observed y subject to that less than K components of $\hat{w}$ are different than 0 . More formally:

$$\hat{w} = \arg\min_{w} \|y - Fw\|_2^2 - \lambda \|w\|_0 \tag{4}$$

where

$$\|x\|_p^p = \sum_{m=1}^{M} |x_n|^p \qquad \|w\|_{p\to 0} = \sum_{m=1}^{M} I(w_n \neq 0) \tag{5}$$

with $p = 2$ being the Euclidean norm, and $p = 0$ is a measure of sparsity (number of components different than 0). The square-norm is the most widely used metric to measure goodness of fit [7, 8, 9, 10] , and by increasing (decreasing) the parameter λ a solution with higher (lower) sparsity is obtained.

Since the imposed constraints $\|w\|_0$ are not linear, computing the minimizing solution of (4) would require to solve $\binom{M}{k}$ least squares problems. This approach is intractable for chromosome lengths M and number of discontinuities K that are typical for our application. Thus, either, i) a greedy optimization strategy is used (e.g. MP[8], OMP[9]), or ii) the sparsity conditions have to be relaxed (e.g. MOF[7], BP[7], SBL [10]).

Matching pursuits methods (MP, OMP) are greedy and only guaranteed to converge to the optimal solution if the dictionary is indeed an orthogonal basis or only one vector in the dictionary is used for the representation. In basis pursuit the $\|w\|_0$ norm is replaced by a $\|w\|_1$ norm, and now the problem can be solved by convex programming and is guaranteed to converge, but the solution is only guaranteed to be the same as in $\|w\|_0$ case if the dictionary coherence is below some bound [12]. These limitations of matching pursuit and basis pursuit methods are particularly problematic in the case of wavelet footprint dictionaries, since footprints that correspond to contiguous discontinuities are highly collinear. The coherence C indeed approaches 1 (i.e. the worst case) as M increases:

$$C = \max_{k \neq j} \langle f_k, f_j \rangle \quad \langle f_k, f_j \rangle = \sqrt{\tfrac{k(M-j)}{(M-k)j}} \text{ if } k < j \tag{6}$$

One of our contributions is to observe that sparse Bayesian learning does not exhibit this problem and thus is better suited for this type of dictionaries.

4. WAVELET FOOTPRINTS AND SBL

The optimization problem defined in (4) can be formulated from a Bayesian estimation point of view, as was done in [10] for the case of overcomplete dictionaries. If a normal likelihood model is assumed for the observations $p(y|w) \sim \mathcal{N}(Fw, \sigma^2 I)$, and an appropriate prior is used for the weights $p(w) \sim \exp(-\|w\|_p^p)$, then $\hat{w}$ in (4) is indeed the maximum a posteriori (MAP) estimate:

$$\begin{aligned} \hat{w}_{MAP} &= \arg\max_{w} p(w|y) \\ &= \arg\max_{w} p(y|w)\, p(w) \\ &= \arg\min_{w} -\log p(y|w) - \log p(w) \end{aligned} \tag{7}$$

In SBL [10, 13], the prior distribution for the weights is specified as a hierarchical prior:

$$p(w|\alpha) = \prod_{k=0}^{M-1} \mathcal{N}\left(w_k|0, \alpha_k^{-1}\right) \quad (8)$$

where the α is a vector of hyperparameters that are distributed according to a gamma distribution:

$$p(\alpha) = \prod_{k=0}^{M-1} \Gamma(\alpha_k|a,b) \quad (9)$$

This prior is very useful for the following reasons [10, 13]. First, given the hyperparameters α, the posterior weight distribution (10) is normal, and the weights are estimated as the posterior mean $\hat{w} = \mu$.

$$p\left(w|y,\alpha,\sigma^2\right) = \mathcal{N}(w|\mu,\Sigma) \quad (10)$$

$$\Sigma = \left(\sigma^{-2}F'F + \mathrm{diag}(\alpha)\right)^{-1} \qquad \mu = \sigma^{-2}\Sigma F'y \quad (11)$$

Second, by treating the weights w as hidden variables, the maximum likelihood estimation for the hyperparameters α can be obtained by the EM algorithm:

$$E\,\text{Step}: \quad E_{w|y,\alpha^{(l)},\sigma^2}\left(w_i^2\right) = \Sigma_{ii} + \mu_i^2 \quad (12)$$

$$M\,\text{Step}: \quad \alpha_i^{(l+1)} = \frac{1+2a}{\Sigma_{ii}+\mu_i^2+2b} \quad (13)$$

Finally, the unconditional prior $p(w) = \int p(w|\alpha)\,p(\alpha)\,d\alpha$ is a multivariate t-student distribution and approximates $\|w\|_0$ better than the l_1 norm used in BP, as shown in Figure 1.

The sparsity of the solution obtained by SBL is studied in detail, and compared to BP, in [10] for the case of arbitrary dictionaries. Their results and Figure 1 demonstrate SBL as a better measure of sparseness than BP as compared to $\|w\|_0$. This is even more important in the case of highly coherent dictionaries such as the wavelet footprint that we use here (2), where the bounds in BP performance [12] do not apply. Consider, as an example, the case where the noise-free signal has only one discontinuity at i, such that $w_i = 2$ (approximation A), but where the noise levels are such that $w_{i-1} = 1$ and $w_i = 1$ (approximation B) provide better MSE fit to the noisy data. Then in a BP framework, both approximations lead to the same sparseness within BP, since $|2| = |1| + |1|$. Thus, BP would select approximation B, as it has better MSE approximation cost and the same sparseness. In contrast, in SBL the cost of approximation B will be twice as much as that of approximation A, so that the sparser approximation (A) is more likely to be selected. Note that these kinds of situations would not arise if the dictionary had lower coherence[12]. Because coherence is high, it is possible for approximations A and B to be comparable in terms MSE; indeed, in our example, the two solutions will be different in *only two samples*, namely i and $i-1$.

The EM algorithm is guaranteed to improve the solution after each step and will always converge [10], but it may converge to a local minimum instead of the global minimum. However, these local minima are indeed always sparse, Theorem 2 in [10]. Nevertheless, the corresponding weights w can be individually thresholded to escape from a local minimum. An automatic procedure to select the threshold which controls for the expected false discovery rate of the discovered breakpoints has been employed, details can be found in [14].

The noise σ^2 can also be jointly estimated by the EM algorithm as in [10]. However, since each chromosome in the genome is analyzed independently, and σ^2 is assumed to be the same for all chromosomes, it is more robust to estimate σ^2 for all the genome before applying the EM algorithm in each chromosome. In this paper, σ^2 is estimated as in [6].

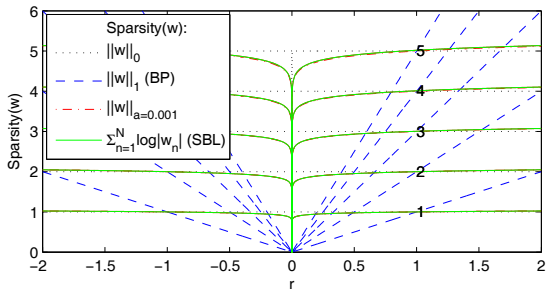

Fig. 1. SBL and BP sparsity metrics compared to the desired (quasi)norm-0. Each curve is the corresponding sparsity measure for a vector with $K = 1, \ldots, 5$ equal coefficients different than 0; $w_{1,\ldots,K} = r$ and $w_{K+1,\ldots,M} = 0$.

Direct computation of (11) for an arbitrary dictionary F would require $O(M^3)$ operations [10, 13]. However, we have found that for our particular F in (2), $(F'F)^{-1}$ is a symmetric tridiagonal matrix; and this can be exploited to obtain the Σ_{ii} and μ_i in each EM step (12) in $O(M)$ computations. This makes the computation of each EM step very efficient, but the overall complexity of our SBL algorithm depends on how many steps are required to converge. In our experiments the convergence has always been reached in a constant number of steps. But in the future, it would be desirable to obtain an upper bound by analyzing how fast the objective function decreases in each step.

5. EXPERIMENTAL RESULTS

5.1. Results with simulated data

In order to evaluate the proposed algorithm we have employed a benchmark dataset developed by [1] that simulates real life array-CGH observations with the advantage that the copy number change points are known. Since for this specific application the most important objective is to accurately locate the copy number change, the evaluation metrics that have been chosen are:

- Sensitivity: $\frac{\#\,\text{discontinuities detected correctly}}{\text{Total}\,\#\,\text{of discontinuities present}}$
- False Discovery Rate: $\frac{\#\,\text{discontinuities detected incorrectly}}{\text{Total}\,\#\,\text{discontinuities detected}}$

First, we compare different signal reconstruction techniques for our wavelet footprint dictionary. In each method, a threshold can be adjusted to allow for more or less discontinuities in the reconstruction, and is represented by a point in a curve in Figure 2. This figure shows that SBL outperforms matching pursuits methods (MP, MMP, OMP) on all operating points. Note that, as discussed earlier, the high coherence in the dictionary explains the relatively poor behavior of BP; optimizing the l_1 norm does not lead necessarily to a sparse solution in this case.

Second, we compare our proposed technique to other existing methods for copy number detection (see Figure 3). The only other technique that offers comparable but slightly lower performance than our method is the DNAcopy.

The DNAcopy approach [11] is based on a recursive circular binary segmentation, where each segment is recursively broken into two or three smaller segments. In each step, the problem is solved by arranging the segment as a circle and finding the best positions

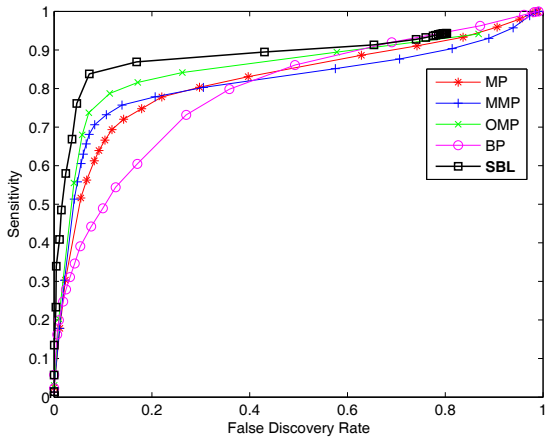

Fig. 2. Receiver operational curves for sensitivity vs. false discovery rate in detecting real copy number changes within a $w = 2$ sample precision window in Willenbrock dataset [1].

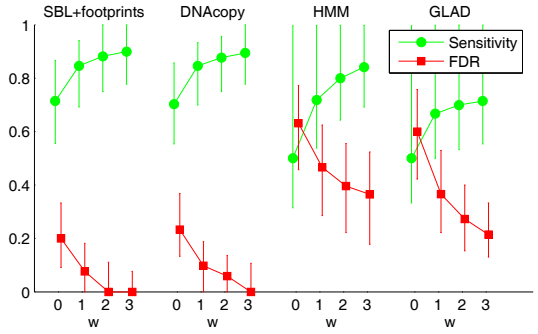

Fig. 3. Median Sensitivity and FDR for detecting real copy number changes within a $w = 0...3$ sample precision window in Willenbrock dataset [1]. The first 3 methods were already analyzed in [1] and the same results are reported. The result obtained by the new proposed method has one of the best performances

for two breakpoints. The key differences compared to our algorithm are that i) each step requires $O(M^2)$ computation, and ii) it follows a greedy strategy that is not guaranteed to find the optimal solution. Our SBL algorithm explores all possible solutions weighted by the bayesian model, resulting on SBL having a better average performance as indicated by the results. More recently, [15] proposed an approximate method that recursively finds a copy number interval with $O(M\epsilon^{-2})$ cost each time. This can be thought as a faster version of DNAcopy with the cost of some additional degradation.

5.2. Results with a human neuroblastoma cell line

With real microarray data the exact location of the copy number changes is not known, and the evaluation metrics employed in the previous section cannot be used. Nevertheless, we experimented the new algorithm with tumor samples (analyzed with the Affymetrix 500K mapping array set) with very encouraging results, as exemplified in Figure 4. This is the result of analyzing a human neuroblastoma cell line (SK-N-BE2) known to have amplification of the MYCN gene, located at 2p24.1. The first peak in the top plot corresponds to this already known gain of MYCN, and the second one is under investigation to determine candidate genes. The bottom plot shows a gain of the q arm on chromosome 17, which commonly occurs in high-risk neuroblastomas and cell lines.

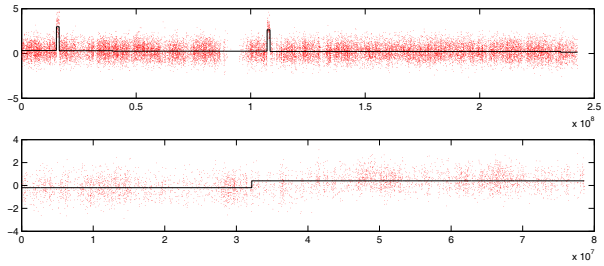

Fig. 4. Detection of two regions of chromosome 2 (top) and one region in chr 17 (bottom) with amplified copy number from of a neuroblastoma tumor cell line (SK-N-BE2) analyzed with an Affymetrix 500K mapping array set and our copy number detection technique.

6. CONCLUSIONS

In this paper a sparse representation for the genome copy number has been developed using a wavelet footprints technique. SBL has been shown to have better performance than other learning methods when the representation dictionary is highly coherent as is the case for wavelet footprints. Compared to other existing techniques for copy number detection, the new developed technique has a very competitive performance both in terms of FDR and sensitivity, as well as in terms of computational complexity. Additionally, the threshold can be automatically adjusted to control for the expected False Discovery Rate [14].

Finally, in future work, we will work on analyzing the overall computational complexity of the SBL algorithm; and if it is possible to incorporate in our work and extend the interval finding strategies in [15] to approximate directly the overall optimal segmentation solution.

7. REFERENCES

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

[2] Donna G Albertson, Colin Collins, Frank McCormick, and Joe W Gray, "Chromosome aberrations in solid tumors.," *Nat Genet*, vol. 34, no. 4, pp. 369–76, 2003.

[3] A Kallioniemi, O P Kallioniemi, D Sudar, D Rutovitz, J W Gray, F Waldman, and D Pinkel, "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.," *Science*, vol. 258, no. 5083, pp. 818–21, 1992.

[4] J R Pollack, C M Perou, A A Alizadeh, M B Eisen, A Pergamenschikov, C F Williams, S S Jeffrey, D Botstein, and P O Brown, "Genome-wide analysis of DNA copy-number changes using cDNA microarrays.," *Nat Genet*, vol. 23, no. 1, pp. 41–6, 1999.

[5] Jing Huang, Wen Wei, Jane Zhang, Guoying Liu, Graham R Bignell, Michael R Stratton, P Andrew Futreal, Richard Wooster, Keith W Jones, and Michael H Shapero, "Whole genome DNA copy number changes identified by high density oligonucleotide arrays.," *Hum Genomics*, vol. 1, no. 4, pp. 287–99, 2004.

[6] P. Dragotti and M. Vetterli, "Wavelet footprints: Theory, algorithms, and applications," *IEEE-Trans-SP*, vol. 51, no. 5, pp. 1306–1323, May 2002.

[7] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998, Also appeared in SIAM Review 43(1), 129-159 (2001).

[8] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE-Trans-SP*, vol. 41, no. 12, pp. 3397–3415, 1993.

[9] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," 1993.

[10] D.P. Wipf and B.D. Rao, "Sparse bayesian learning for basis selection," *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[11] Adam B Olshen, E S Venkatraman, Robert Lucito, and Michael Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data.," *Biostatistics*, vol. 5, no. 4, pp. 557–72, 2004.

[12] D.L. Donoho, M. Elad, and V.N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

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

[14] R. Piqué-Regí, Jordi Monso-Varona, A. Ortega, Timothy Triche, Robert Seeger, and S. Asgharzadeh, "Sparse representation and bayesian detection of genome copy number changes from array data," *Submitted to Bioinformatics*, 2007.

[15] Doron Lipson, Yonatan Aumann, Amir Ben-Dor, Nathan Linial, and Zohar Yakhini, "Efficient calculation of interval scores for DNA copy number data analysis.," *J Comput Biol*.