Applications of Multilevel Thresholding Algorithms to Transcriptomics Data

Luis Rueda and Iman Rezaeian

School of Computer Science, University of Windsor,
401 Sunset Ave., Windsor, ON, N9B3P4, Canada
{lrueda,rezaeia}@uwindsor.ca

Abstract. Microarrays are one of the methods for analyzing the expression levels of genes in a massive and parallel way. Since any errors in early stages of the analysis affect subsequent stages, leading to possibly erroneous biological conclusions, finding the correct location of the spots in the images is extremely important for subsequent steps that include segmentation, quantification, normalization and clustering. On the other hand, genome-wide profiling of DNA-binding proteins using ChIP-seq and RNA-seq has emerged as an alternative to ChIP-chip methods. Due to the large amounts of data produced by next generation sequencing technology, ChIPseq and RNA-seq offer much higher resolution, less noise and greater coverage than its predecessor, the ChIPchip array.

Multilevel thresholding algorithms have been applied to many problems in image and signal processing. We show that these algorithms can be used for transcriptomics and genomics data analysis such as sub-grid and spot detection in DNA microarrays, and also for detecting significant regions based on next generation sequencing data. We show the advantages and disadvantages of using multilevel thresholding and other algorithms in these two applications, as well as an overview of numerical and visual results used to validate the power of the thresholding methods based on previously published data.

Keywords: microarray image gridding, image analysis, multi level thresholding, transcriptomics.

1 Introduction

Among other components, the genome contains a set of genes required for an organism to function and evolve. However, the genome is only a source of information and in order to function, the genes express themselves into proteins. The transcription of genes to produce RNA is the first stage of gene expression. The transcriptome can be seen as the complete set of RNA transcripts produced by the genome. Unlike the genome, the transcriptome is very dynamic. Despite having the same genome regardless of the type of cell or environmental conditions, the transcriptome varies considerably in differing circumstances because of the different ways the genes may express.

Transcriptomics, the field that studies the role of the transcriptome, provides a rich source of data suitable for pattern discovery and analysis. The quantity and size of these data may vary based on the model and underlying methods used for analysis. In gene

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 26–37, 2011.

expression microarrays, the raw data are represented in terms of images, typically in TIFF format which are approximately 20-30MB per array. These TIFF files are processed and transformed into quantified data used for posterior analysis. In contrast, high throughput sequencing methods (e.g. ChIP-seq and RNA-seq) generate more than 1TB of data, while the sequence files (approximately 20-30GB) are typically used as a starting point for analysis [16]. Clearly, these sequence files are an order of magnitude larger than those from arrays.

1.1 DNA Microarray Image Gridding

Various technologies have been developed to measure the transcriptome, including hybridization or sequence-based approaches. Hybridization-based approaches typically involve processing fluorescently labeled DNA microarrays. Microarrays are one of the most important technologies used in molecular biology to massively explore the abilities of the genes to express themselves into proteins and other molecular machines responsible for different functions in an organism. These expressions are monitored in cells and organisms under specific conditions, and are present in many applications in medical diagnosis, pharmacology, disease treatment, among others. If we consider DNA microarrays, scanning the slides at a very high resolution produces images composed of sub-grids of spots. Image processing and analysis are two important aspects of microarrays, and involve various steps. The first task is gridding, which is quite important as errors are propagated to subsequent steps. Roughly speaking, gridding consists of determining the spot locations in a microarray image (typically, in a sub-grid). The gridding process requires the knowledge of the sub-girds in advance in order to proceed, which is not necessarily available in advance.

Many approaches have been proposed for microarray image gridding and spot detection, being the most widely known the following. The Markov random field (MRF) is one of them, which applies specific constraints and heuristic criteria [15]. Other gridding methods used for gridding include mathematical morphology [8], Bayesian model-based algorithms [1,6], the hill-climbing approach [13], a Gaussian mixture model approach [18], Radon-transform-based method [11], a genetic algorithm for separating sub-grids and spots [5], and the recently introduced maximum margin method [4]. A method that we have proposed and has been successfully used in microarray gridding is the multilevel thresholding algorithm [21], which is discussed in more detail later in the paper.

1.2 ChIP-Seq and RNA-Seq Peak Finding

Hybridization-based approaches are high throughput and relatively inexpensive, except for high-resolution tiling arrays that interrogate large genomes. However, these methods have several limitations, which include reliance upon existing knowledge about the genome, high background levels owing to cross-hybridization, and a limited dynamic range of detection owing to both background and saturation of signals [16,26]. Moreover, comparing expression levels across different experiments is often difficult and can require complicated normalization methods.

Recently, the development of novel high-throughput DNA sequencing methods has provided a new method for both mapping and quantifying transcriptomes. These methods, termed ChIP-seq (ChIP sequencing) and RNA-seq (RNA sequencing), have clear advantages over existing approaches and are emerging in such a way that eukaryotic transcriptomes are to be analyzed in a high-throughput and more efficient manner [26].

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is a technique that provides quantitative, genome-wide mapping of target protein binding events [2,17]. In ChIP-seq, a protein is first cross-linked to DNA and the fragments subsequently sheared. Following a size selection step that enriches for fragments of specified lengths, the fragments ends are sequenced, and the resulting reads are aligned to the genome. Detecting protein binding sites from massive sequence-based datasets with millions of short reads represents a truly bioinformatics challenge that has required considerable computational innovation in spite of the availability of programs for ChIP-chip analysis [3,9,18,19].

With the increasing popularity of ChIP-seq technology, a demand for peak finding methods has emerged and it causes developing new algorithms. Although due to mapping challenges and biases in various aspects of existing protocols, identifying peaks is not a straightforward task.

Different approaches have been proposed for detecting peaks based ChIP-seq/RNA-seq mapped reads so far. Zhang et al. presents a Model-based Analysis of ChIP-seq data (MACS), which analyzes data generated by short read sequencers [28]. It models the shift size of ChIP-seq tags, and uses it to improve the spatial resolution of predicted binding sites. A two-pass strategy called PeakSeq has been presented in [20]. This strategy compensates for signal caused by open chromatin, as revealed by the inclusion of the controls. The first pass identifies putative binding sites and compensates for genomic variation in mapping the sequences. The second pass filters out sites not significantly enriched compared to the normalized control, computing precise enrichments and significance. A statistical approach for calling peaks has been recently proposed in [7], which is based on evaluating the significance of a robust statistical test that measures the extent of pile-up reads. Specifically, the shapes of putative peaks are defined and evaluated to differentiate between random and non-random fragment placements on the genome. Another algorithm for identification of binding sites is site identification from paired-end sequencing (SIPeS) [25], which can be used for identification of binding sites from short reads generated from paired-end solexa ChIP-seq technology.

In this paper, we review the application of optimal multilevel thresholding (OMT) to gridding and peak finding problems in transcriptomics. Moreover, a conceptual and practical comparison between OMT and other state-of-the-art approaches is also presented.

2 Optimal Multilevel Thresholding

Multilevel thresholding is one of the most widely-used techniques in different aspects of signal and image processing, including segmentation, classification and object discrimination. Given a histogram with frequencies or probabilities for each bin, the

aim of multilevel thresholding is to divide the histogram into a number of groups (or classes) of contiguous bins in such a way that a criterion is optimized. In microarray image gridding, we compute vertical (or horizontal) running sums of pixel intensities, obtaining histograms in which each bin represents one column (or row respectively), and the running sum of intensities corresponds to the frequency of that bin. The frequencies are then normalized in order to be considered as probabilities. Each histogram is then processed (see below) to obtain the optimal thresholding that will determine the locations of the separating lines.

Consider a histogram H, an ordered set $\{1, 2, \ldots, n-1, n\}$, where the ith value corresponds to the ith bin and has a probability, p_i. Given an image, $A = \{a_{ij}\}$, H can be obtained by means of the horizontal (vertical) running sum as follows: $p_i = \sum_{j=1}^{m} a_{ij}$ ($p_j = \sum_{i=1}^{n} a_{ij}$). We also consider a threshold set T, defined as an ordered set $T = \{t_0, t_1, \ldots, t_k, t_{k+1}\}$, where $0 = t_0 < t_1 < \ldots < t_k < t_{k+1} = n$ and $t_i \in \{0\} \cup H$. The problem of multilevel thresholding consists of finding a threshold set, T^*, in such a way that a function $f : H^k \times [0,1]^n \rightarrow \mathbb{R}^+$ is maximized/minimized. Using this threshold set, H is divided into $k+1$ classes: $\zeta_1 = \{1, 2, \ldots, t_1\}$, $\zeta_2 = \{t_1 + 1, t_1+2, \ldots, t_2\}$, $\ldots$, $\zeta_k = \{t_{k-1}+1, t_{k-1}+2, \ldots, t_k\}$, $\zeta_{k+1} = \{t_k+1, t_k+2, \ldots, n\}$. The most important criteria for multilevel thresholding are the following [12]:

Between class variance:

$$\Psi_{\mathrm{BC}}(T) = \sum_{j=1}^{k+1} \omega_j \mu_j^2 \tag{1}$$

where $\omega_j = \sum_{i=t_{j-1}+1}^{t_j} p_i$, $\mu_j = \frac{1}{\omega_j} \sum_{i=t_{j-1}+1}^{t_j} i p_i$;

Entropy-based:

$$\Psi_{\mathrm{H}}(T) = \sum_{j=1}^{k+1} H_j \tag{2}$$

where $H_j = -\sum_{i=t_{j-1}+1}^{t_j} \frac{p_i}{\omega_j} \log \frac{p_i}{\omega_j}$;

Minimum error:

$$\Psi_{\mathrm{ME}}(T) = 1 + 2 \sum_{j=1}^{k+1} \omega_j (\log \sigma_j - \log \omega_j) \tag{3}$$

where $\sigma_j^2 = \sum_{i=t_{j-1}+1}^{t_j} \frac{p_i (i-\mu_j)^2}{\omega_j}$.

A dynamic programming algorithm for *optimal* multilevel thresholding was proposed in our previous work [12], which is an extension for irregularly sampled histograms. For this, the criterion has to be decomposed as a sum of terms as follows:

$$\Psi(T_{0,m}) = \Psi(\{t_0, t_1, \ldots, t_m\}) \triangleq \sum_{j=1}^{m} \psi_{t_{j-1}+1, t_j} \,, \tag{4}$$

where $1 \leq m \leq k+1$ and the function $\psi_{l,r}$, where $l \leq r$, is a real, positive function of $p_l, p_{l+1}, \ldots, p_r$, $\psi_{l,r} : H^2 \times [0,1]^{l-r+1} \rightarrow \mathbb{R}^+ \cup \{0\}$. If $m = 0$, then $\Psi(\{t_0\}) = \psi_{t_0,t_0} = \psi_{0,0} = 0$. The thresholding algorithm can be found in [12]. In

the algorithm, a table C is filled in, where $C(t_j, j)$ contains the optimal solution for $T_{0,j} = t_0, t_1, \ldots, t_j$, $\Psi^*(T_{0,j})$, which is found from $\min\{t_j\} \leq t_j \leq \max\{t_j\}$. Another table, $D(t_j, j)$, contains the value of t_{j-1} for which $\Psi^*(T_{0,j})$ is optimal. The algorithm runs in $O(kn^2)$, and has been further improved to achieve linear complexity, i.e. $O(kn)$, by following the approach of [14].

2.1 Using Multi-level Thresholding for Gridding DNA Microarray Images

A DNA microarray image contains spots arranged into sub-grids. The image contains various sub-grids as well, which are found in the first stage. Once the sub-grids are found, the spots centers are to be identified. A microarray image can be considered as a matrix $A = \{a_{i,j}\}, i = 1, ..., n$ and $j = 1, ..., m$, where $a_{ij} \in \mathbb{Z}^+$, and A is a sub-grid of a DNA microarray image. The aim of sub-gridding is to obtain vectors, namely $\mathbf{h} = [h_1, ...h_{p-1}]^t$ and $\mathbf{v} = [v_1, ...v_{q-1}]^t$, that separate the sub-grids. Finding the spot locations is done analogously – more details of this, as well as those of the whole process can be found in [21]. The aim of gridding is to find the corresponding spot locations given by the horizontal and vertical adjacent vectors. Post-processing or refinement allows us to find a spot region for each spot, which is enclosed by four lines.

When producing the microarrays, based on the layout of the printer pins, the number of sub-grids or spots is known. But due to misalignments, deformations, artifacts or noise during producing the microarray images, these numbers may not be available. Thus, it is important that the gridding algorithm allows some flexibility in finding these parameters, as well as avoiding the use of other user-defined parameters. This is what the thresholding methods endeavor to do, by automatically finding the best number of thresholds (sub-grids or spots) – more details in the next section.

2.2 Using Multi-level Thresholding for Analyzing ChIP-Seq/RNA-Seq Data

In ChIP-seq and RNA-seq analysis, a protein is first cross-linked to DNA and the fragments subsequently pruned. Then, the fragments ends are sequenced, and the resulting reads are aligned to the genome. The result of read alignments produces a histogram in such a way that the x axis represents the genome coordinate and the y axis the frequency of the aligned reads in each genome coordinate. The aim is to find the significant peaks corresponding to enriched regions. For this reason, a non-overlapping moving window is used. By starting from the beginning, a dynamic window of minimum size t is being applied to the histogram and each window that could be analyzed separately. The size of the window could be different for each window to prevent truncating a peak before its end. Thus, for each window a minimum number of t bins is used and, by starting from the end of previous window, the size of window is increased until a zero value in the histogram is reached.

The aim is to obtain vectors $C_{w_i} = [c^1_{w_i}, ...c^n_{w_i}]^t$, where w_i is the i^{th} window and C_{w_i} is the vector that contains n threshold coordinates which correspond to the i^{th} window. Figure 1 depicts the process of finding the peaks corresponding to the regions of interest for the specified protein. The input to the algorithm includes the reads and the output of the whole process is the location of the detected significant peaks by using optimal multilevel thresholding combined with our recently proposed α index.

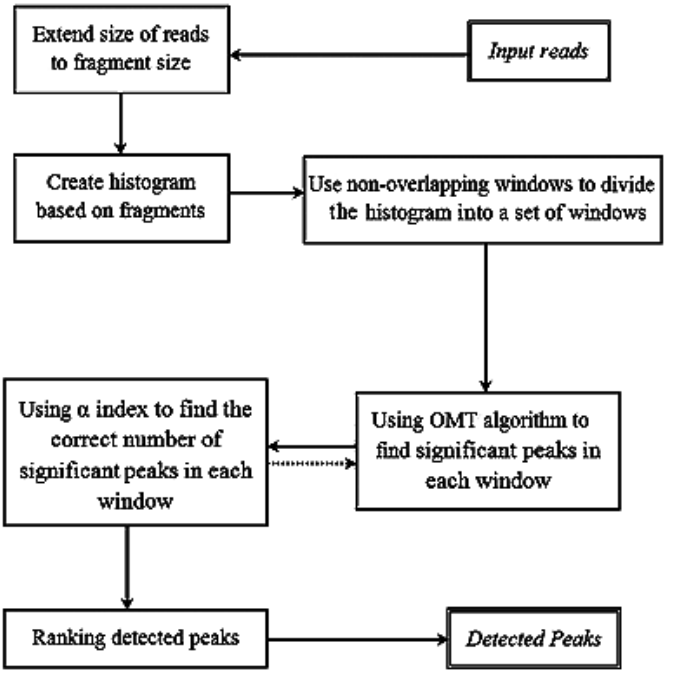

Fig. 1. Schematic representation of the process for finding significant peaks

3 Automatic Detection of the Number of Clusters

Finding the correct number of clusters (number of sub-grids or spots or the number of regions in each window in ChIP-seq/RNA-seq analysis) is one of the most challenging issues. This stage is crucial in order to fully automate the whole process. For this, we need to determine the correct number clusters or thresholds prior to applying multi-level thresholding methods. This is found by applying an index of validity (derived from clustering techniques) and testing over all possible number of clusters (or thresholds) from 2 to $\sqrt{n}$, where n is the number of bins in the histogram. We have recently proposed the $\alpha(x)$ index, which is the result of a combination of a simple index and the well-known I index [23] as follows:

$$\alpha(K) = \sqrt{K}\frac{I(K)}{A(K)} = \frac{\left(\frac{E_1}{E_K} \times D_K\right)^2}{\sqrt{K}\Sigma_{i=1}^{K} p(t_i)}. \tag{5}$$

For maximizing $I(K)$ and minimizing $A(K)$, the value of $\alpha(K)$ must be maximized. Thus, the best number of thresholds K^* based on the α index is given by:

$$K^* = \underset{1\leq K\leq \delta}{\operatorname{argmax}}\, \alpha(K) = \underset{1\leq K\leq \delta}{\operatorname{argmax}} \frac{\left(\frac{E_1}{E_K} \times D_K\right)^2}{\sqrt{K}\Sigma_{i=1}^{K} p(t_i)}. \tag{6}$$

4 Comparison of Transcriptomics Data Analysis Algorithms

4.1 DNA Microarray Image Gridding Algorithms Comparison

A conceptual comparison of microarray image gridding methods based on their features is shown in Table 1. The methods included in the comparison are the following: (i) Radon transform sub-gridding (RTSG) [11], (ii) Bayesian simulated annealing gridding (BSAG) [1], (iii) genetic-algorithm-based gridding (GABG) [5], (iv) hill-climbing gridding (HCG) [13], (v) maximum margin microarray gridding (M^3G) [4], and the optimal multilevel thresholding algorithm for gridding (OMT) [21]. As shown in the table, OMT does not need any number-based parameter, and hence making it much more powerful than the other methods. Although the index or thresholding criterion can be considered as a "parameter", this can be fixed by using the between class criterion. In a previous work, we have "fixed" the index of validity to the α *index* and the *between class* as the thresholding criterion. As can also be observed in the table, most algorithms and methods require the use of user-defined and subjectively fixed parameters. One example is the GABG, which needs to adjust the mutation and crossover rates, probability of maximum and minimum thresholds, among others. It is critical then to adjust these parameters for specific data, and variations may occur across images of different characteristics.

Table 1. Conceptual comparison of recently proposed DNA microarray gridding methods

Method	Parameters	Sub-grid Detection	Spot Detection	Automatic Detection No. of Spots	Rotation
Rueda07	n: Number of sub-grids	$\surd$	$\times$	$\times$	$\surd$
Antoniol04	α ,β: Parameters for balancing prior and posterior probability rates	$\times$	$\surd$	$\surd$	$\surd$
Zacharia08	μ , c :Mutation and Crossover rates, p_{max}: probability of maximum threshold, p_{low}: probability of minimum threshold, f_{max} : percentage of line with low probability to be a part of grid, T_p: Refinement threshold	$\surd$	$\surd$	$\surd$	$\surd$
Rueda06	λ , σ: Distribution parameters	$\times$	$\surd$	$\surd$	$\times$
Bariamis10	c: Cost parameter	$\times$	$\surd$	$\surd$	$\surd$
OMT	None[1]	$\surd$	$\surd$	$\surd$	$\surd$

[1] The only parameters that would be needed in the proposed method are the "thresholding criterion" and the "index of validity". These two "parameters" are methodological, not number-based, and hence making OMT less dependent on parameters.

4.2 Comparison of Algorithms for ChIP-Seq and RNA-Seq Analysis

A conceptual comparison between thresholding algorithms and other ChIP and RNA-Seq methods based on their features is shown in Table 2. The methods included in the comparison are the following: (i) GLobal Identifier of Target Regions (CLITR) [22], (ii)

Table 2. Conceptual comparison of recently proposed methods for ChIP-seq and RNA-seq data

Method	Peak selection criteria	Peak ranking	Parameters
GLITR	n: Classification by height and relative enrichment	Peak height and fold enrichment	Target FDR, number nearest neighbors for clustering
MACS v1.3.5	Local region Poisson p value	p value	p-value threshold, tag length, m-fold for shift estimate
PeakSeq	Local region binomial p value	q value	Target FDR
Quest v2.3	height threshold, background ratio	q value	KDE bandwidth, peaks height, sub-peak valley depth, ratio to background
SICER v1.02	p value from random background model, enrichment relative to control	q value	Window length, gap size, FDR (with control) or E-Value (no control)
SiSSRs v1.4	$N^+ - N^-$ sign change, $N^+ + N^-$ threshold in region	p value	FDR, $N^+ + N^-$ threshold
T-PIC	Local height threshold	p value	Average fragment length, significance p value, minimum length of interval
OMT	number of ChIP reads minus control reads in window	volume	Average fragment length

Model-based Analysis of ChIP-seq (MACS)[28], (iii) PeakSeq [20], (iv) quantitative enrichment of sequence tags (Quest) [24], (v) SICER [27], (vi) Site Identification from Short Sequence Reads (SiSSRs) [10], (vii) Tree shape Peak Identification for ChIP-seq (T-PIC) [7], and (viii) the optimal multilevel thresholding algorithm, OMT. As shown in the table, all algorithms require some parameters to be set by the user based on the particular data to be processed, including p-values, FDR, number os nearest neighbors, peak height, valley depth, window length, gap size, among others. OMT is the algorithm that requires almost no parameter at all. Only the average fragment length is needed, but this parameter can be easily estimated from the underlying data. In practice, if enough computational resources are available, the fragment length would not be needed, since the OMT algorithm could be run directly on the whole histogram.

5 Experimental Analysis

This section is necessarily brief and reviews some experimental results as presented in [21]. For the experiments, two different kinds of DNA microarray images have been used, which were obtained from the Stanford Microarray Database (SMD) the Gene Expression Omnibus (GEO). The images have different resolutions, number of sub-grids and spots. We have used the between-class variance as the thresholding criteria, since it is the one that delivers the best results. All the sub-grids in each image are detected with a 100% accuracy, and also spot locations in each sub-grid can be detected efficiently with an average accuracy of 96.2% for SMD dataset and 96% for GEO dataset. Figure 2 shows the detected sub-grids from the AT-20387-ch2 image (left) and the detected spots in one of sub-grids (right). As shown in the figure, the proposed method precisely detects the sub-grids location at first, and in the next stage, each sub-grid is divided precisely into the corresponding spots with the same method.

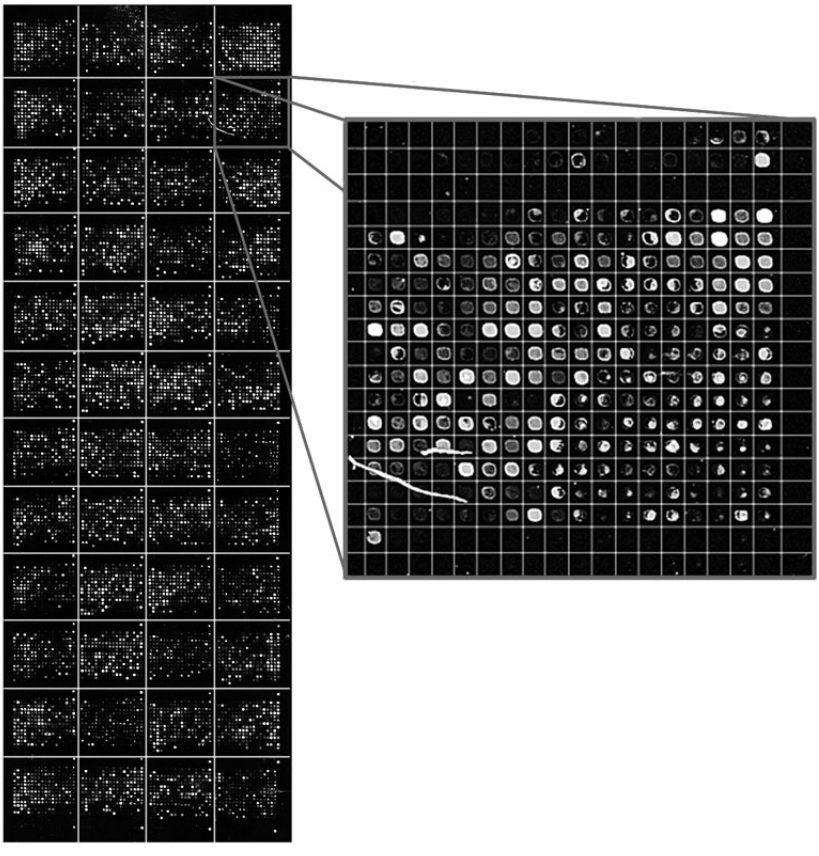

Fig. 2. Detected sub-grids in AT-20387-ch2 microarray image (left) and detected spots in one of sub-grids (right)

In addition to this, some experimental, preliminary results for testing performance of the OMT algorithm on ChIP/RNA-seq data are shown here. We have used the FoxA1 dataset [28], which contains experiment and control samples of 24 chromosomes. The experiment and control histogram were generated separately by extending each mapped position (read) into an appropriately oriented fragment, and then joining the fragments based on their genome coordinates. The final histogram was generated by subtracting the control from the experiment histogram. To find significant peaks, we used a non-overlapping window with the initial size of 3000bp. To avoid truncating peaks in boundaries, each window is extended until the value of the histogram at the end of the window becomes zero. Figure 3 shows three detected regions for chromosomes 9 and 17 and their corresponding base pair coordinates. It clear from the pictures that the peaks contain a very high number of reads, and then these regions are quite likely to represent binding sites, open reading frames or other bio-markers. A biological assessment of these bio-markers can corroborate this.

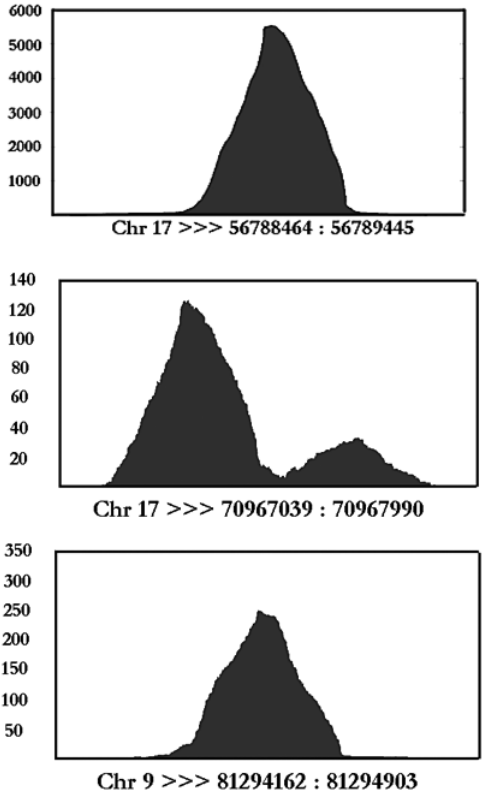

Fig. 3. Three detected regions from FoxA1 data for chromosomes 9 and 17. The x axis corresponds to the genome position in bp and the y axis corresponds to the number of reads.

6 Discussion and Conclusion

Transcriptomics provide a rich source of data suitable for pattern analysis. We have shown how multilevel thresholding algorithms can be applied to an efficient analysis of transcriptomics and genomics data by finding sub-grids and spots in microarray images, as well as significant peaks in high-throughput next generation sequencing data. OMT can be applied to a wide range of data from different sources and with different characteristics, and allows data analysis such as sub-grid and spot detection in DNA microarray image gridding and also for detecting significant regions on ChIP and RNA-seq data. OMT has been shown to be statistically sound and robust to noise in experiments and it is able to use on different approaches with a little change – this is one the most important features of this algorithm.

Thresholding algorithms, though shown to be quite useful for transcriptomics and genomics data analysis, are still emerging tools in these areas, and open the possibility for further advancement. One of the problems that deserves attention is the use of other thresholding criteria, including minimum error, entropy-based and others. For these two

criteria the algorithm still runs in quadratic or n-logarithmic complexity, and which make the whole process sluggish. Processing a whole genome or even a chromosome for finding peaks in ChIP or RNA-seq is still a challenge, since it involves histograms with several million bins. This makes it virtually impossible to process a histogram at once, and so it has to be divided into several fragments. Processing the whole histograms at once is one of the open and challenging problems that deserve more investigation. Next generation sequence data analysis is an emerging and promising area for pattern discovery and analysis, which deserve the attention of the research community in the field.

Acknowledgments. This work has been partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada.

References

1. Ceccarelli, B., Antoniol, G.: A Deformable Grid-matching Approach for Microarray Images. IEEE Transactions on Image Processing 15(10), 3178–3188 (2006)
2. Barski, A., Zhao, K.: Genomic location analysis by chip-seq. Journal of Cellular Biochemistry (107), 11–18 (2009)
3. Buck, M., Nobel, A., Lieb, J.: Chipotle: a user-friendly tool for the analysis of chip-chip data. Genome Biology 6(11), R97 (2005)
4. Bariamis, D., Maroulis, D., Iakovidis, D.: M^3G: Maximum Margin Microarray Gridding. BMC Bioinformatics 11, 49 (2010)
5. Zacharia, E., Maroulis, D.: Micoarray image gridding via an evolutionary algorithm. In: IEEE International Conference on Image Processing, pp. 1444–1447 (2008)
6. Antoniol, G., Ceccarelli, M.: A Markov Random Field Approach to Microarray Image Gridding. In: Proc. of the 17th International Conference on Pattern Recognition, pp. 550–553 (2004)
7. Hower, V., Evans, S., Pachter, L.: Shape-based peak identification for chip-seq. BMC Bioinformatics 11(81) (2010)
8. Angulo, J., Serra, J.: Automatic Analysis of DNA Microarray Images Using Mathematical Morphology. Bioinformatics 19(5), 553–562 (2003)
9. Johnson, W., Li, W., Meyer, C., Gottardo, R., Carroll, J., Brown, M., Liu, X.S.: Model-based analysis of tiling-arrays for chip-chip. Proceedings of the National Academy of Sciences 103(33), 12457–12462 (2006)
10. Jothi, R., Cuddapah, S., Barski, A., Cui, K., Zhao, K.: Genome-wide identification of in vivo proteindna binding sites from chip-seq data. Nucleic Acids Research 36(16), 5221–5231 (2008)
11. Rueda, L.: Sub-grid Detection in DNA Microarray Images. In: Proceedings of the IEEE Pacific-RIM Symposium on Image and Video Technology, pp. 248–259 (2007)
12. Rueda, L.: An Efficient Algorithm for Optimal Multilevel Thresholding of Irregularly Sampled Histograms. In: Proceedings of the 7th International Workshop on Statistical Pattern Recognition, pp. 612–621 (2008)
13. Rueda, L., Vidyadharan, V.: A Hill-climbing Approach for Automatic Gridding of cDNA Microarray Images. IEEE Transactions on Computational Biology and Bioinformatics 3(1), 72–83 (2006)
14. Luessi, M., Eichmann, M., Schuster, G., Katsaggelos, A.: Framework for efficient optimal multilevel image thresholding. Journal of Electronic Imaging 18 (2009)
15. Katzer, M., Kummer, F., Sagerer, G.: A Markov Random Field Model of Microarray Gridding. In: Proceeding of the 2003 ACM Symposium on Applied Computing, pp. 72–77 (2003)

16. Malone, J., Oliver, B.: Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biology 9(1), 34 (2011)
17. Park, P.J.: Chip-seq: advantages and challenges of a maturing technology. Nat. Rev. Genetics 10(10), 669–680 (2009)
18. Qi, Y., Rolfe, A., MacIsaac, K.D., Gerber, G., Pokholok, D., Zeitlinger, J., Danford, T., Dowell, R., Fraenkel, E., Jaakkola, T.S., Young, R., Gifford, D.: High-resolution computational models of genome binding events. Nat. Biotech. 24(8), 963–970 (2006)
19. Reiss, D., Facciotti, M., Baliga, N.: Model-based deconvolution of genome-wide dna binding. Bioinformatics 24(3), 396–403 (2008)
20. Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, Z., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., Gerstein, M.: Peakseq enables systematic scoring of chip-seq experiments relative to controls. Nat. Biotech. 27(1), 66–75 (2009)
21. Rueda, L., Rezaeian, I.: A fully automatic gridding method for cdna microarray images. BMC Bioinformatics 12, 113 (2011)
22. Tuteja, G., White, P., Schug, J., Kaestner, K.H.: Extracting transcription factor targets from ChIP-Seq data. Nucleic Acids Res. 37(17), e113 (2009)
23. Maulik, U., Bandyopadhyay, S.: Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(12), 1650–1655 (2002)
24. Valouev, A., Johnson, D., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R., Sidow, A.: Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Meth. 5(9), 829–834 (2008)
25. Wang, C., Xu, J., Zhang, D., Wilson, Z., Zhang, D.: An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. BMC Bioinformatics 41(1), 117–129 (2008)
26. Wang, Z., Gerstein, M., Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10(1), 57–63 (2009)
27. Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., Peng, W.: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25(15), 1952–1958 (2009)
28. Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nusbaum, C., Myers, R., Brown, M., Li, W., Liu, X.S.: Model-based analysis of chip-seq (macs). Genome Biology 9(9), R137 (2008)