# Finding Genomic Features from Enriched Regions in ChIP-Seq Data

Iman Rezaeian
*School of Computer Science*
*University of Windsor*
*Windsor, Canada*
*Email: rezaeia@uwindsor.ca*

Luis Rueda
*School of Computer Science*
*University of Windsor*
*Windsor, Canada*
*Email: lrueda@uwindsor.ca*

***Abstract*—Finding genomic features in ChIP-Seq data has become an attractive research topic lately, because of the power, resolution and low-noise of next generation sequencing, making it a much better alternative to traditional microarrays such as ChIP-chip and other related methods. However, handling ChIP-Seq data is not straightforward, mainly because of the large amounts of data produced by next generation sequencing. ChIP-Seq has widespread over a range of applications in finding biomarkers, especially those associated with important genomic features in epigenomics and transcriptomics, including binding sites, promoters, exons/introns, transcription sites, among others. Efficient algorithms for finding relevant regions in ChIP-Seq data have been proposed, which capture the most significant peaks from the sequence reads. Among these, multilevel thresholding algorithms have been applied successfully for transcriptomics and genomics data analysis, in particular for detecting significant regions based on next generation sequencing data.**

**We show that the Optimal Multilevel Thresholding algorithm (OMT) achieves higher accuracy in detecting enriched regions and genomic features of detected regions on FoxA1 data. OMT finds more gene-related regions (gene, exon, promoter) in comparison with other methods. Using a small number of parameters is another advantage of the proposed method.**

***Keywords*-multi level thresholding; transcriptomics; ChIP-Seq data analysis**

## I. Introduction

Genome-wide mapping of protein-DNA interactions is essential for understanding of transcriptional regulation. Mapping of binding sites for transcription factors and other DNA-binding proteins is essential for decoding gene regulatory networks that underlie different biological processes. Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) is one of those techniques that provides quantitative, genome-wide mapping of target protein binding events [1], [2].

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 reference genome. Detecting protein binding sites from massive sequence-based datasets with millions of short reads represents a truly bioinformatics challenge that requires considerable computational resources, in spite of the availability of programs for ChIP-chip analysis [3].

With the increasing popularity of ChIP-Seq technology, the demand for peak finding methods has increased the need to develop 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 on ChIP-Seq/RNA-Seq mapped reads. Zhang et al. presented a *model-based analysis of ChIP-Seq data* (MACS), which analyzes data generated by short read sequencers [4]. It models the length of the sequenced ChIP fragments and uses it to improve the spatial resolution of predicted binding sites. A two-pass strategy called *Peak-Seq* has been presented in [5]. This strategy compensates for signals 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. *Tree shape Peak Identification for ChIP-Seq* (T-PIC) is a statistical approach for calling peaks that has been recently proposed in [6]. This approach is based on evaluating the significance of a robust statistical test that measures the extent of pile-up reads. Another algorithm for identification of binding sites is *site identification from paired-end sequencing* (SIPeS) [7], which can be used for identification of binding sites from short reads generated from paired-end Illumina ChIP-Seq technology.

One of the possible problems of the existing methods is that the location of detected peaks could be non-optimal. Moreover, for detecting these peaks all of the methods use a set of parameters that may cause variations of the results for different datasets. In [8], we have recently proposed a method for finding significant peaks using optimal multilevel thresholding (OMT). Here, we show that OMT can be efficiently used to find genomic features when used in conjunction with a model for finding the best number of peaks. The results of our experiments show that our method can achieve a higher degree of accuracy than two recently developed methods, MACS and T-PIC, while providing

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

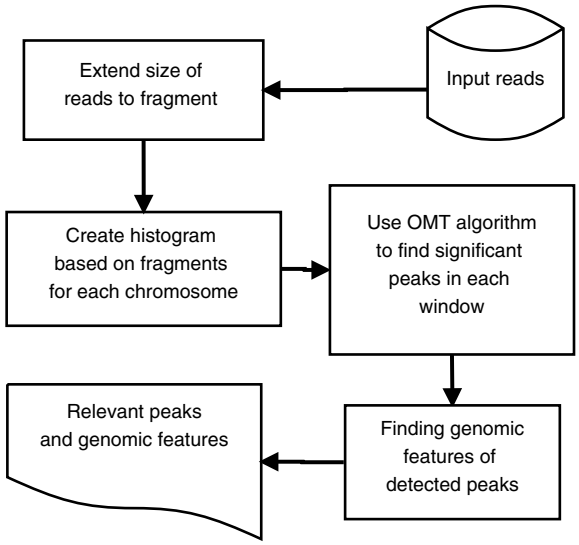


Figure 1. Schematic representation of the process for finding genomic features by using OMT.

flexibility when applying it to different datasets.

## II. The Peak Detection Method

In ChIP-Seq, 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 reading the alignments produces a histogram with genome coordinates as the $x$-axis and frequency of the aligned reads in each genome coordinate as the $y$-axis. The aim is to find significant peaks corresponding to enriched regions. Each peak can be seen as a homogeneous group (cluster) which is well separated from the others by means of "valleys". In that sense, the problem can be formulated as *one-dimensional clustering*. Figure 1 depicts the process of finding peaks and genomic features corresponding to the regions of interest for the specified protein. Each module is explained in detail in the next few sections.

The first step of the model consists of converting the Input BED file into a histogram. After extending each read to a fragment length based on the direction of each read (forward or backward), each of them is aligned to the reference genome based on its coordinates. Afterwards, for each chromosome, separate histograms for experiment and control data are created for further processing.

Starting from the beginning of the chromosome, a sliding window of minimum size $t$ is applied to the histogram and each window is analyzed separately. The sizes of the windows are not necessarily equal 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 the previous window, the size of the window is increased until a zero value in the histogram is reached. We consider a minimum of $t = 3,000$ in order to ensure that a window covers at least one peak of typical size.

After creating the histogram based on fragments (reads), the histogram is then processed to obtain the optimal thresholding that will determine the locations of peaks. More detail about this procedure can be found in [8].

A dynamic programming algorithm for *optimal* multi-level thresholding was proposed in our previous work [9], which is an extension for irregularly sampled histograms. The optimal thresholding is the one that maximizes the between-class variance. The algorithm runs in $O(kn^2)$ for a histogram of $n$ bins, and has been further improved to achieve linear complexity, i.e. $O(kn)$, by following the approach of [10].

## III. Finding Relevant Peaks and Genomic Features

Finding the correct number of peaks (the number of regions in each window) is crucial in order to fully automate the whole process. For this, we need to determine the correct number peaks prior to applying the multi-level thresholding method. This is found by using an index of validity derived from clustering techniques. We have recently proposed the $\alpha(K)$ index [11], which is the result of a combination of a simple index, $A(K)$, and the well-known $I$ index [12]. By computing and comparing values of $\alpha(K)$ over all possible numbers of clusters, the one with the maximum value of $\alpha(K)$ is the best number of clusters.

After finding the locations of the detected peaks, in a two step process, significant peaks are selected. In the first step, the effective area of each peak is found by shrinking the peak. In the second step, the two sample Cramer-von Mises non parametric hypothesis test [13], with $\alpha = 0.01$, is used to accept/reject peaks based on the comparison between experiment and control histograms corresponding to each peak. Finally, the peaks are ranked and returned as the final relevant peaks.

In the next step, using the information gathered from the UCSC Genome Browser on *NCBI36/hg18* assembly, the genomic features of each detected peak have been investigated. We assign a genomic feature to a peak if that peak overlaps with the region containing that genomic feature. Since a detected peak can be located in a genomic region with different genomic features, it could also have different genomic features. For example, if a specific peak overlaps with an exon and intron simultaneously, we count that peak as an intron *and* an exon.

## IV. Experimental Results

We have used the FoxA1 dataset [4], which contains experiment and control samples of 24 chromosomes. As in

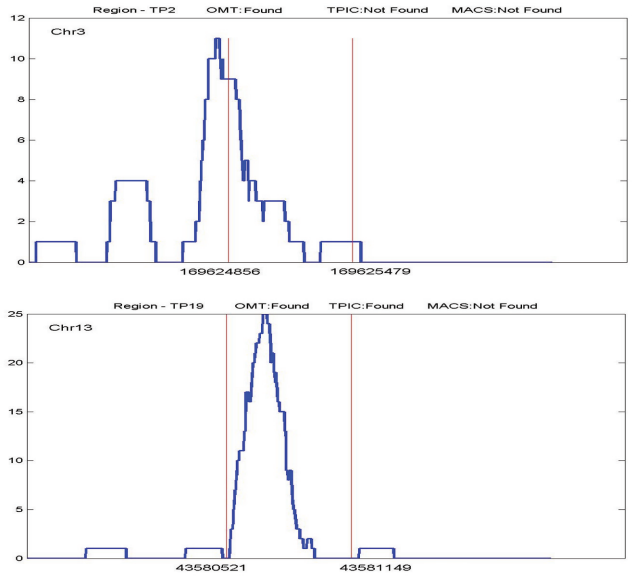


Figure 2. Two true positive regions in chromosomes 3 and 13 of FoxA1 dataset. The $x$-axis corresponds to the genome position in bp and the $y$-axis corresponds to the number of reads. Both peaks are detected by OMT but only the bottom one is detected by T-PIC and none of them is detected by MACS.

[6], the experiment and control histograms 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 whose initial size is 3,000bp. To avoid truncating peaks in boundaries, each window is extended until the value of the histogram at the end of the window becomes zero.

Computing the enrichment score for each method proceeds as follows. Random intervals from the genome are created by selecting the same number of intervals with the same lengths from each chromosome as in the called peaks but with random starting locations. Then, the number of occurrences of the binding motif in the called peaks and the random intervals are counted. The enrichment score is the ratio of the number of occurrences in the called peaks divided by the number of occurrences in the random intervals.

### A. Comparison with Other Methods for ChIP-Seq Analysis

Table I shows a comparison between OMT and two recently proposed methods, MACS [4] and T-PIC [6]. As shown in the table, the number of significant peaks detected by OMT is higher than those of the other two methods. This implies that OMT is able to find some significant peaks that are not detected by the other two methods. Also, the enrichment ratio for OMT is far higher than MACS and higher than T-PIC. However, the average size of the peaks is smaller than the other two methods which implies that OMT is able to detect significant peaks more precisely.

Table I
COMPARISON BETWEEN OMT AND TWO RECENTLY PROPOSED METHODS, MACS AND T-PIC, BASED ON NUMBER OF DETECTED PEAKS, MEAN LENGTH OF DETECTED PEAKS AND ENRICHMENT SCORE.

| Dataset | Method of Comparison | OMT | T-PIC | MACS |
|---|---|---|---|---|
| FoxA1 | Detected peaks | 20,032 | 17,619 | 13,639 |
| | Mean length of peaks | 306 | 510 | 394 |
| | Enrichment ratio | 2.62 | 2.54 | 1.68 |

A conceptual comparison of OMT with MACS and T-PIC based on their features is shown in Table II. As shown in the table, the other algorithms require some parameters to be set by the user based on the particular data to be processed, including $p$-values, $m$-fold, window length, among others. OMT is the algorithm that requires the smallest number of parameters. Only the average fragment length is needed.

### B. Analysis of Genomic Features

We have also biologically validated the peaks detected by OMT on the results of independent qPCR experiments for the FoxA1 protein. For this, we considered 25 true positives and 7 true negatives (regions) reported in [14]. The results of other two well-known methods, T-PIC and MACS, are included in the comparison. Table IV shows the result of this biological validation on each method. As the other two methods, OMT has been able to reject all true negatives. Although OMT finds a larger number of regions, OMT shows a high sensitivity, finding more true positives than T-PIC and MACS. As an example, two true positive regions in chromosomes 3 and 13 of FoxA1 are shown in Figure 2. Both peaks are detected by OMT but only the bottom one is detected by T-PIC and none of them is detected by MACS.

In another experiment, we compared the type and corresponding number of regions found by these three methods in the FoxA1 dataset. Table III shows the percentage of regions which are located in gene, promoter, intron and exon areas as well as inter-genetic regions. OMT was able to detect more regions corresponding to genes, promoters and exons, while the percentage of detected regions within inter-genetic area by our proposed method is less than number of regions corresponding to the other two methods. In contrast, the number of detected regions corresponding to the introns found by OMT is not higher than the other two methods.

## V. DISCUSSION AND CONCLUSION

We have presented a multi-level thresholding algorithm that can be applied to an efficient analysis of ChIP-Seq data to find significant peaks and genomic features. OMT can be applied to high-throughput next generation sequencing data with different characteristics, and allows us detecting

Table II
CONCEPTUAL COMPARISON OF RECENTLY PROPOSED METHODS FOR $ChIP-Seq$ DATA.

| Method | Peak selection criteria | Peak ranking | Parameters |
|---|---|---|---|
| MACS | local region Poisson $p$-value | $p$-value | $p$-value threshold, tag length, $m$-fold for shift estimate |
| 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 | $p$-value | average fragment length |

Table III
COMPARISON BETWEEN OUR PROPOSED METHOD, MACS AND T-PIC, BASED ON THE PERCENTAGE OF DETECTED REGIONS WHICH BELONG TO DIFFERENT GENOMIC FEATURES.

| Method | Number of Regions | Genes | | Exons | | Introns | | Promoters | | Inter-genetic Regions | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Regions | % | Regions | % | Regions | % | Regions | % | Regions | % |
| MACS | 13,639 | 12,125 | 88.90 | 976 | 7.16 | 11,689 | 85.70 | 688 | 5.05 | 7,533 | 55.23 |
| T-PIC | 17,619 | 15,529 | 88.14 | 1,336 | 7.58 | 15,325 | 86.98 | 793 | 4.50 | 8,794 | 49.91 |
| OMT | 20,032 | 19,557 | 97.63 | 1,941 | 9.69 | 17,258 | 86.15 | 1,296 | 6.47 | 9,155 | 45.70 |

Table IV
COMPARISON OF OMT, MACS AND T-PIC, BASED ON THE NUMBER OF TRUE POSITIVE (TP) AND TRUE NEGATIVE (TN) DETECTED PEAKS.

| | OMT | T-PIC | MACS |
|---|---|---|---|
| TP | 15 | 13 | 12 |
| TN | 0 | 0 | 0 |

significant regions on ChIP-Seq data. OMT has been shown to be sound and robust in experiments. Finding more genomic features in comparison with two other state of the art methods, MACS and T-PIC, and using fewer parameter are other interesting features of OMT.

## REFERENCES


[1] A. Barski and K. Zhao, "Genomic location analysis by chip-seq," *Journal of Cellular Biochemistry*, no. 107, pp. 11–18, 2009.

[2] P. Park, "Chip-seq: advantages and challenges of a maturing technology," *Nat Rev Genetics*, vol. 10, no. 10, pp. 669–680, 2009.

[3] D. Reiss, M. Facciotti, and N. Baliga, "Model-based deconvolution of genome-wide dna binding," *Bioinformatics*, vol. 24, no. 3, pp. 396–403, 2008.

[4] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, , and X. Liu, "Model-based analysis of chip-seq (macs)," *Genome Biology*, vol. 9, no. 9, p. R137, 2008.

[5] J. Rozowsky, G. Euskirchen, R. Auerbach, Z. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. Gerstein, "Peakseq enables systematic scoring of chip-seq experiments relative to controls," *Nature Biotechnology*, vol. 27, no. 1, pp. 66–75, 2009.

[6] V. Hower, S. Evans, and L. Pachter, "Shape-based peak identification for chip-seq," *BMC Bioinformatics*, vol. 11, no. 81, 2010.

[7] C. Wang, J. Xu, D. Zhang, Z. Wilson, and D. Zhang, "An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data," *BMC Bioinformatics*, vol. 41, no. 1, pp. 117–129, 2008.

[8] I. Rezaeian and L. Rueda, "A new algorithm for finding enriched regions in chip-seq data," *ACM Conference on Bioinformatics, Computational Biology and Biomedicine - to appear*, 2012.

[9] L. Rueda, "An Efficient Algorithm for Optimal Multilevel Thresholding of Irregularly Sampled Histograms," *Proceedings of the 7th International Workshop on Statistical Pattern Recognition*, pp. 612–621, 2008.

[10] M. Luessi, M. Eichmann, G. Schuster, and A. Katsaggelos, "Framework for efficient optimal multilevel image thresholding," *Journal of Electronic Imaging*, vol. 18, 2009.

[11] L. Rueda and I. Rezaeian, "A fully automatic gridding method for cdna microarray images," *BMC Bioinformatics*, vol. 12, p. 113, 2011.

[12] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1655, 2002.

[13] T. W. Anderson, "On the Distribution of the Two-Sample Cramer-von Mises Criterion," *Ann. Math. Statist.*, vol. 33, pp. 1148–1159, 1962.

[14] M. Lupien, J. Eeckhoute, C. A. Meyer, Q. Wang, Y. Zhang, W. Li, J. S. Carroll, X. S. Liu, and M. Brown, "FoxA1 Translates Epigenetic Signatures into Enhancer-driven Lineage-specific Transcription," *Cell*, vol. 132, no. 6, pp. 958–970, 2008.