PeakSeg: Peak detection via constrained optimal Segmentation

Toby Dylan Hocking, Guillem Rigaill, and Guillaume Bourque

Abstract

Peak detection is a central problem in ChIP-seq data analysis, and current algorithms for this task are unsupervised and mostly effective for a single data type (e.g. histone H3K4me3 profiles with sharp peaks). We propose PeakSeg, a supervised peak detection algorithm based on constrained optimal segmentation, which is easy to tune since it has only one free parameter: the optimal number of peaks. We propose to tune it using annotated regions in a supervised penalty function learning problem, which we show results in state-of-the-art peak detection for both sharp H3K4me3 and broad H3K36me3 data types.

1 Introduction to supervised ChIP-seq peak detection

Chromatin immunoprecipitation sequencing (ChIP-seq) is a genome-wide assay to profile histone modifications and transcription factor binding sites, with many experimental and computational steps [Bailey et al., 2013]. Briefly, each experiment yields a set of sequence reads which are aligned to a reference genome, and then the data are interpreted by counting the number of aligned reads at each genomic position. In this paper we propose a new method for peak calling these data, which is a binary classification problem for each genomic position. The positive class is enriched (peaks) and the negative class is background noise.

More concretely, a ChIP-seq profile on a single chromosome with d base pairs is a vector $\mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_d \end{bmatrix} \in \mathbb{Z}_+^d$ of counts of aligned sequence reads. A peak detection algorithm can be described as a function $c : \mathbb{Z}_+^d \rightarrow \{0,1\}^d$ which returns 0 for background noise and 1 for a peak. In contrast to the supervised method proposed in this paper, most previous algorithms are unsupervised since they define a peak detector c using only the profile data $\mathbf{y}$.

In supervised peak detection [Hocking et al., 2014], there are n annotated samples, and each sample $i \in \{1, \dots, n\}$ has a profile $\mathbf{y}_i \in \mathbb{Z}_+^d$ and a set of annotated regions R_i which defines a non-convex annotation error function

$$E[c(\mathbf{y}_i), R_i] = \text{FP}[c(\mathbf{y}_i), R_i] + \text{FN}[c(\mathbf{y}_i), R_i]. \tag{1}$$

The annotation error counts the number of false positive (FP) and false negative (FN) regions, so it takes values in the non-negative integers. The goal is to find a peak caller with minimal error on some test profiles:

$$\underset{c}{\text{minimize}} \sum_{i \in \text{test}} E[c(\mathbf{y}_i), R_i]. \tag{2}$$

2 Related work

In the benchmark data set of Hocking et al. [2014], there are two different histone mark types: H3K4me3 (sharp peaks) and H3K36me3 (broadly enriched regions). The best peak detection algorithm for these H3K4me3 data was macs [Zhang et al., 2008], and the best for H3K36me3 was HMCan [Ashoor et al., 2013]. Both of these algorithms are unsupervised, but were calibrated using the annotated region labels to choose the best scalar significance threshold hyperparameter via grid search.

The ChIP-seq segmentation model we propose is a constrained version of the model proposed by Cleynen and Lebarbier [2013]. Specifically, they proposed to search all possible change-points to find the optimal segmentation, but we propose to constrain the possible change-points to the subset of models that can be interpreted as peaks.

3 Unsupervised PeakSeg: segmenting one ChIP-seq profile

After fixing a maximum number of segments $1 \leq s_{\text{max}} \leq d$, the unconstrained maximum likelihood segmentation problem is defined for any $s \in \{1, 2, \dots, s_{\text{max}}\}$ as

$$\begin{aligned} \hat{\mathbf{m}}^s(\mathbf{y}) = \ & \operatorname*{arg\,min}_{\mathbf{m} \in \mathbb{R}^d} \quad \rho(\mathbf{m}, \mathbf{y}) \\ & \text{such that} \quad \text{Segments}(\mathbf{m}) = s, \end{aligned} \tag{3}$$

where $\rho(\mathbf{m}, \mathbf{y}) = \sum_{j=1}^{d} m_j - y_j \log m_j$ is the loss function corresponding to maximum likelihood inference of a Poisson distribution with mean parameter m_j. The model complexity $\text{Segments}(\mathbf{m}) = 1 + \sum_{j=2}^{d} I(m_j \neq m_{j-1})$ is the number of segments, where I is the indicator function. Although it is a non-convex optimization problem, the sequence of segmentations $\hat{\mathbf{m}}^1(\mathbf{y}), \dots, \hat{\mathbf{m}}^{s_{\text{max}}}(\mathbf{y})$ can be computed in $O(s_{\text{max}} d^2)$ time using dynamic programming (DP) algorithms [Bellman, 1961], or in $O(s_{\text{max}} d \log d)$ time using pruned DP [Rigaill, 2010, Cleynen et al., 2014].

We refer to (3) as the unconstrained problem since $\hat{\mathbf{m}}^s(\mathbf{y})$ is the most likely segmentation of all possible models with s segments. Several unconstrained models are shown on the left of Figure 1, and for example the 2nd segment of the model with $s = 3$ segments appears to capture the peak in the data. To construct a peak detector c, first define the sign of the change before base $j \in \{2, \dots, d\}$ as

$$S_j(\mathbf{m}) = \text{sign}(m_j - m_{j-1}), \tag{4}$$

with $S_1(\mathbf{m}) = 0$ by convention. Furthermore we define the number of peaks at base $j \in \{1, \dots, d\}$ as

$$P_j(\mathbf{m}) = \sum_{k=1}^{j} S_k(\mathbf{m}). \tag{5}$$

In general for the unconstrained model $P_j(\mathbf{m}) \in \mathbb{Z}$, and for example in Figure 1 there is a position j for which $P_j\left[\hat{\mathbf{m}}^5(\mathbf{y})\right] = 2$ (since the mean changes up, up, down, down). We would like to constrain the number of peaks $P_j(\mathbf{m}) \in \{0, 1\}$ so that we can use $c(\mathbf{y}) = \mathbf{P}\left[\tilde{\mathbf{m}}^s(\mathbf{y})\right]$ as a peak detector, where $\mathbf{P}[\mathbf{m}] = \left[\begin{array}{ccc} P_1(\mathbf{m}) & \cdots & P_d(\mathbf{m}) \end{array} \right] \in \{0, 1\}^d$. That results in the constrained problem

$$\begin{aligned} \tilde{\mathbf{m}}^s(\mathbf{y}) = \ & \operatorname*{arg\,min}_{\mathbf{m} \in \mathbb{R}^d} \quad \rho(\mathbf{m}, \mathbf{y}) \\ & \text{such that} \quad \text{Segments}(\mathbf{m}) = s, \\ & \qquad\qquad\quad P_j(\mathbf{m}) \in \{0, 1\} \text{ for all } j \in \{1, \dots, d\}. \end{aligned} \tag{6}$$

Another way to interpret the constrained problem (6) is that the sequence of changes in the segment means $\mathbf{m}$ must begin with a positive change and then alternate: up, down, up, down, ... (and not up, up, down). Thus the even

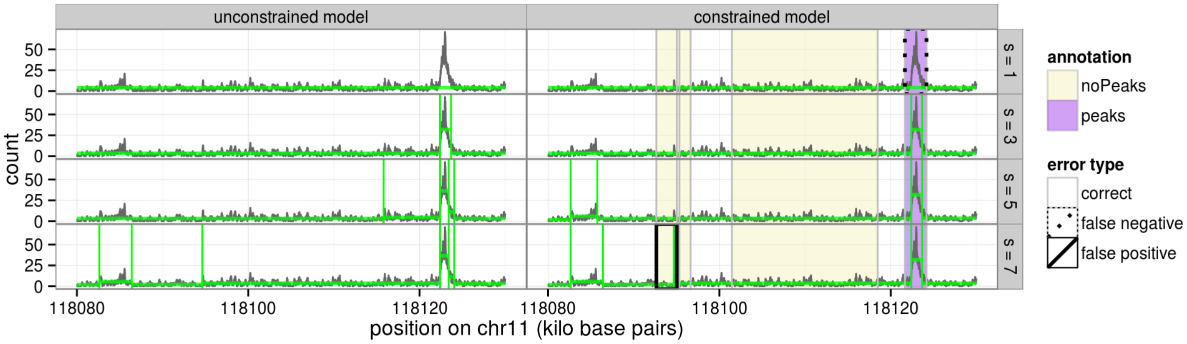

Figure 1: Example profile $\mathbf{y}$ (black), with green horizontal lines for the segmentation mean $\mathbf{m}$, and green vertical lines to emphasize change-points. For this particular profile $\mathbf{y}$, the models are equivalent $\hat{\mathbf{m}}^s(\mathbf{y}) = \tilde{\mathbf{m}}^s(\mathbf{y})$ for $s \in \{1, 3\}$ segments but not for $s \in \{5, 7\}$. For the constrained models, the 2nd, 4th, ... are interpreted as peaks (5), whose accuracy can be quantified using the annotations (1).

numbered segments (2nd, 4th, etc) may be interpreted as peaks, and the odd numbered segments (1st, 3rd, etc) may be interpreted as background. Figure 1 shows a profile where the constraint is necessary to detect peaks for models with $s \in \{5, 7\}$ segments. We propose to use dynamic programming to compute the sequence of maximum likelihood models $\tilde{\mathbf{m}}^1(\mathbf{y}), \dots, \tilde{\mathbf{m}}^{s_{\text{max}}}(\mathbf{y})$ satisfying this up-down constraint. The algorithm is in $O(s_{\text{max}} d^2)$ time, where d is the number of data points, using the compression scheme proposed by Cleynen et al. [2014].

4 Supervised PeakSeg: learning a penalty function

After computing the constrained maximum likelihood segmentations for each sample $i \in \{1, \dots, n\}$, the only question that remains is: how many segments? To predict a sample-specific number of segments, we propose to use the annotated regions to learn a penalty function [Hocking et al., 2013]. Briefly, we define the optimal number of segments

$$s^*(\lambda, \mathbf{y}) = \operatorname*{arg\,min}_{s \in \{1,3,\dots,s_{\text{max}}\}} \rho\left[\tilde{\mathbf{m}}^s(\mathbf{y}), \mathbf{y}\right] + \lambda s, \tag{7}$$

$\lambda \in \mathbb{R}_+$ is a positive penalty value. We define sample-specific penalty values $\log \lambda_i = f(\mathbf{x}_i) = \beta + \mathbf{w}^\intercal \mathbf{x}_i$, which is an affine function with parameters $\beta \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^m$ that will be learned. We used an $m = 2$-dimensional feature vector $\mathbf{x}_i = \left[\ \log \max \mathbf{y}_i \quad \log d_i\ \right]$ where d_i is the number of base pairs for sample/chromosome i. The learning algorithm amounts to minimizing a smooth convex loss $\ell_i : \mathbb{R} \rightarrow \mathbb{R}_+$ which depends on the annotated region data R_i:

$$\hat{f} = \operatorname*{arg\,min}_{f} \sum_{i=1}^{n} \ell_i\left[f(\mathbf{x}_i)\right]. \tag{8}$$

Since ℓ_i is smooth and convex, this problem can be easily solved using gradient-based algorithms. We used the accelerated gradient method of the FISTA algorithm [Beck and Teboulle, 2009].

To make a prediction on a test sample with profile $\mathbf{y}$ and features $\mathbf{x}$, we compute the predicted penalty $\hat{\lambda} = \exp \hat{f}(\mathbf{x})$, the predicted number of segments $\hat{s} = s^*(\hat{\lambda}, \mathbf{y})$, and finally the predicted peaks $\mathbf{P}\left[\tilde{\mathbf{m}}^{\hat{s}}(\mathbf{y})\right]$.

5 Results: state-of-the-art peak detection for two data types

We downloaded 7 benchmark annotated region data sets.[1] Since the DP solver that we implemented has $O(s_{\text{max}} d^2)$ time complexity, we considered only the subset of samples which had at most $d \leq 100,000$ data points to segment in each labeled genomic window. This left a database of 4,628 annotated regions (out of 12,826 annotated regions in the entire benchmark). Finally, since these genomic windows are relatively small, we set the maximum number of segments $s_{\text{max}} = 19$, and for each profile $\mathbf{y}$ we computed $\tilde{\mathbf{m}}^1(\mathbf{y}), \dots, \tilde{\mathbf{m}}^{19}(\mathbf{y})$ (6). For the largest profile we considered ($d = 88,509$ data points), the DP algorithm computed the 19 constrained optimal segmentations in 27 minutes.

We compared the proposed PeakSeg algorithm to the two previous state-of-the-art peak detectors, macs and hmcan.broad. For each data set, we randomly divided the annotated windows into half train and half test. We trained the macs and hmcan.broad algorithms by choosing the significance threshold with minimal annotation error on the train set. We trained PeakSeg by learning a penalty function $\hat{f}$ on the train set (8).

We show the percent test error for each algorithm and each data set in Figure 2. As previously described [Hocking et al., 2014], macs had lower test error than hmcan.broad for H3K4me3 data, and hmcan.broad had lower test error than macs for H3K36me3 data. It is clear that our proposed PeakSeg algorithm had lower test error than both macs and hmcan.broad algorithms, for both data types.

6 Discussion, conclusions, and future work

We proposed to use the solution of a constrained optimal segmentation problem (6) as a ChIP-seq peak detector. For d data points and s_{max} segments, we proposed to use dynamic programming to compute the solutions in $O(s_{\text{max}} d^2)$ time. Furthermore, we proposed to use annotated region data as supervision in a penalty learning problem (8). This approach yields state-of-the-art test error rates for peak detection in a benchmark that includes both H3K4me3 and H3K36me3 data sets.

[1] http://cbio.ensmp.fr/~thocking/chip-seq-chunk-db/

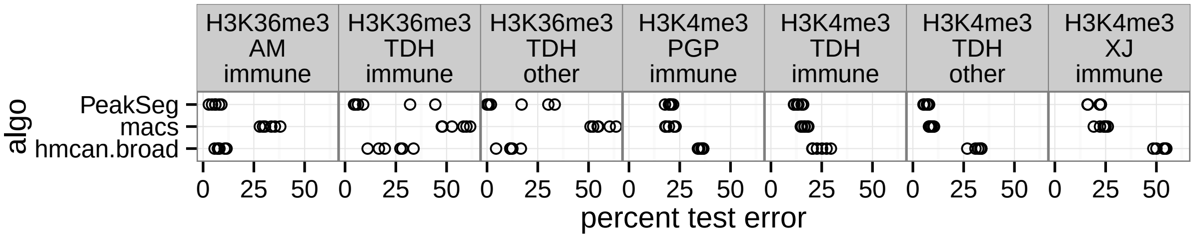

Figure 2: Test error of three algorithms on seven annotated region data sets (each point shows one of six randomly selected train/test splits). Data set names show histone mark type (e.g. H3K36me3), annotator (AM), and cell types (immune). PeakSeg had lower test error than macs and hmcan.broad for both H3K36me3 and H3K4me3 data.

PeakSeg is the first supervised peak detection algorithm. It has been explicitly designed to take advantage of the annotated region data proposed by Hocking et al. [2014]. PeakSeg is also the first algorithm to exhibit state-of-the-art accuracy on both sharp H3K4me3 and broad H3K36me3 ChIP-seq profiles. In the future, even better accuracy may be obtained by engineering better features $\mathbf{x}_i$ for the penalty learning problem (8), perhaps based on Poisson segmentation model selection theory [Cleynen and Lebarbier, 2013].

The current implementation of PeakSeg using dynamic programming has one major limitation. The $O(s_{\max}d^2)$ time complexity has limited its application to subsets of chromosomes with $d \leq 100,000$ data points to segment. In comparison, the largest chromosome subset in the benchmark has $d = 254,451$ data points, and the largest chromosome in hg19 (chr1) has $d = 249,250,621$ base pairs. To apply PeakSeg to these larger data sets, we are investigating a constrained version of pruned dynamic programming [Rigaill, 2010, Cleynen et al., 2014], which has $O(s_{\max}d \log d)$ time complexity.

Finally, we are interested in segmenting multiple samples i at the same time, since peaks are often observed in the same genomic location across several samples of the same cell type. This joint peak detection problem may lead to more accurate peak calls, but it is a considerably more difficult segmentation problem.

References

H. Ashoor, A. Hérault, A. Kamoun, F. Radvanyi, V. B. Bajic, E. Barillot, and V. Boeva. HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics*, 29(23):2979–2986, 2013.

T. Bailey, P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, and J. Zhang. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology*, 9(11):e1003326, 2013.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

R. Bellman. On the approximation of curves by line segments using dynamic programming. *Commun. ACM*, 4(6): 284–, June 1961.

A. Cleynen and E. Lebarbier. Segmentation of the poisson and negative binomial rate models: a penalized estimator. *arXiv:1301.2534*, 2013.

A. Cleynen, M. Koskas, E. Lebarbier, G. Rigaill, and S. Robin. Segmentor3IsBack: an R package for the fast and exact segmentation of Seq-data. *Algorithms for Molecular Biology*, 9:6, 2014.

T. Hocking, G. Rigaill, J.-P. Vert, and F. Bach. Learning sparse penalties for change-point detection using max margin interval regression. In *Proc. 30th ICML*, pages 172–180, 2013.

T. D. Hocking, P. Goerner-Potvin, A. Morin, X. Shao, and G. Bourque. Visual annotations and a supervised learning approach for evaluating and calibrating ChIP-seq peak detectors. *arXiv:1409.6209*, 2014.

G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. arXiv:1004.0887, 2010.

Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137, 2008.