Journal Title Here, 2022, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

L_0 segmentation enables data-driven concise representations of diverse epigenomic data

Ali Tuğrul Balcı[1, 2] and Maria Chikina[1, *]

[1]Computational and Systems Biology, University of Pittsburgh and [2]Joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology
[*]Corresponding author: mchikina@pitt.edu

Abstract

Motivation: Epigenetic assays using next-generation sequencing (NGS) have furthered our understanding of the functional genomic regions and the mechanisms of gene regulation. However, a single assay produces billions of data represented by nucleotide resolution signal tracks. The signal strength at a given nucleotide is subject to numerous sources of technical an biological noise and thus conveys limited information about the underlying biological state. In order to draw biological conclusions, data is typically summarized into higher order patterns. Numerous specialized algorithms for summarizing epigenetic signal have been proposed and include methods for peak calling or finding differentially methylated regions. A key unifying principle underlying these approaches is that they all leverage the strong prior that signal must be locally consistent. **Results:** We propose L_0 segmentation as a universal framework for extracting locally coherent signals for diverse epigenetic sources. L_0 serves to both compress and smooth the input signal by approximating it as piece-wise constant. We implement a highly scalable L_0 segmentation with additional loss functions designed for NGS epigenetic data types including Poisson loss for single tracks and binomial loss for methylation/coverage data. In contrast to the widely used L_1 segmentation problem, known as fused lasso, the L_0 solution does not induce global attenuation and is able to capture the salient features of the data over a wide range of compression values. Finally, we show that L_0 segmentation can be used as an effective prior inside other machine learning models, such as matrix factorization. **Availability:** Our approach is implemented as an R package "l01segmentation" with a C++ backend. Available at https://github.com/boooooogey/l01segmentation.

Key words: epigenetics, optimization, compression

Introduction

Using next-generation sequencing (NGS) to profile the epigenetic states of biological systems has rapidly gained popularity. Large scale efforts to profile the epigenome, such as ENCODE [1] and Roadmap [2], have paved the way for these assays being used routinely to investigate diverse biological problems from cancer biology to immune cell development.

Once the data is aligned, epigenetic assays can be represented as (possibly normalized) read densities along the chromosome. Without further processing, each experiment contains on an order of 3×10^9 values but many downstream analyses make use of representations with a much smaller footprint. Most epigenetic assays work by sequencing a sample that is enriched for a subset of the genome with specific properties such as the presence (transcription factor histone mark ChIPseq) or absence (ATACseq, DNAse) of specific chromatin components. With a few important exceptions, these assays produce signals that have peak-like structure, that is, the sequencing reads pile up in specific chromosomal locations representing a small fraction of the genome. Such assays can be succinctly represented by recording the position and value of the peak. The MACS[3] software, which is widely used by consortium projects, produces peak based representations by delineating regions with statistical enrichment of reads over the local background. Importantly, some histone marks, such as K27me3 and K27me36, do not conform to the peak-like structure but instead cover broad regions on the scale of genes. These are generally difficult to find and the approach taken by MACS is to heuristically merge narrow peaks [4].

Sequence-based methylation is unique among epigenetic assays in that it reads out a covalent modification of the DNA and the output data does not have natural "peaks" that are the hallmark of other epigenetic tracks but is rather dispersed through the genome.

Despite not having the peak structure the methylation levels at adjacent sites show a high level of dependence, as they are likely to be governed by a common regulatory process, such as the binding of a TF or polymerase or gene transcription. For experimental designs that are focused on specific biological contrasts, this local coherence of methylation is exploited by combining differentially methylated sites (DMSs) into stretches of differentially methylated regions (DMRs) [5, 6, 7, 8, 9].

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Altogether, the key property of epigenetic data is that while the raw data is reported at nucleotide resolution nearby chromosomal positions are likely to be produced by the same biological or experimental process and this fact can be exploited to define narrow peaks, broad peaks, or DMRs. While these approaches are popular and effective, they rely on specific assumptions, tunable parameters, or, in the case of DMRs, on knowledge of experimental design. In this work we propose a general purpose framework that uses efficient L_0 segmentation and assay specific loss functions to represent any NGS epigentic experiment as piece-wise constant while making no a priori assumption regarding the expected patterns in the data.

Our Approach

A standard way of enforcing that signals are locally consistent is by penalizing the difference in adjacent values. Given an input vector y_i we seek to approximate it with a locally smoothed vector $\boldsymbol{\beta}$ by solving an optimization problem of the following general type:

$$\boldsymbol{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmax}} \sum_{i=1}^{N} e(\beta_i, y_i) - \lambda \sum_{i=1}^{N-1} \Omega\left(\beta_i - \beta_{i+1}\right) \qquad (1)$$

Here $e(\beta, y)$ is a reconstruction error term and $\Omega(\beta_i - \beta_{i+1})$ is a regularization term that penalizes the differences between consecutive values.

If we take squared error and the penalty to be the L_1 norm this problem is known as fused-Lasso. This penalty drives the difference between successive elements of β to 0 and hence the approximation solution is piece-wise constant. Because of this property, the fused lasso can also be seen as a change-point detection algorithm.

In this work we consider a different penalty function, the L_0 error. Instead of penalizing the sum of the absolute values of the differences, this simply penalizes the total number of change points regardless of their values. The formulation of L_0 has important performance and algorithmic consequences. Because all change points are penalized equally, the L_0 formulation can produce much sharper breaks, which, as we show, allows for more efficient segmented representations. Another attractive property of the L_0 formulation is that for the optimal solution, the value of the fused segment is purely a function of the individual data points in the segment (in the case of squared and Poisson error, it is the mean) and does not depend on the adjacent segments. Thus, the L_0 formulation can be understood as a local clustering approach and it does not show any attenuation towards the global mean as is the case with L_1 or fused lasso.

In terms of algorithmic considerations, unlike the L_1 formulation, the L_0 formulation is not convex. However, recent algorithmic advances for the solution to be found in empirical linear time (Johnson 2013).

The algorithm proposed in (Johnson 2013) has theoretical $\mathcal{O}\left(n^2\right)$ time and space complexity. However, in practice, the execution does not reach this theoretical limit and instead scales linearly in both time and space with the length of the input signal. The empirical linear space and time usage of the algorithm make the application to large epigenetic signals feasible.

Our implementation, which extends the approach described to solving L_0 segmentation (Johnson 2013) error functions derived from different probability distributions. We implement the standard Gaussian distribution (squared error) as well as distributions that more accurately model epigenetic count data, such as Poisson for single-track data and binomial for double-track data counts (i. e. methylation and coverage). Empirically, the L_0 segmentation scales linearly in application to biological data. In Fig. 1, we plot the time it takes to solve the L_0 segmentation problem with different probability distributions for various input sizes.

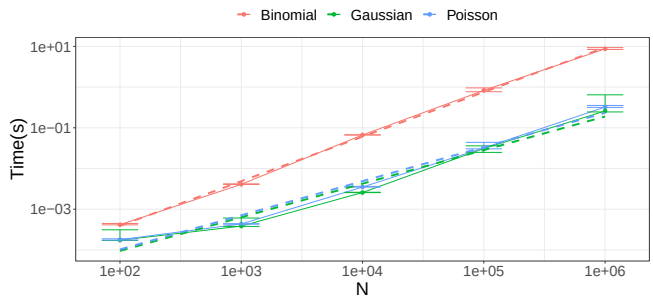

Fig. 1. Binomial, Gaussian, and Poisson L_0 segmentation execution times are plotted in log-space (both x and y) with respect to various segment lengths. The data is simulated using the corresponding probability distributions. Even though the theoretical worst case complexity of the algorithm is $O(n^2)$, in practice, the execution time scales linearly with the length of the input.

For completeness, our implementation also includes functions that solve the L_1-fused approximation problem, another state-of-the-art, ultra-fast algorithm which was also described in (Johnson 2013). However, we will demonstrate that the L_0 formulation is superior in detecting relevant biological signals.

The empirical performance of our implementation both scales linearly and is overall very efficient taking just 0.1 and 1 seconds (for Poisson and Binomial respectively) on 10^6 input. The scalability makes it possible to efficiently compress epigenetic data genome wide. For example, reading in the entire nucleotide-level Chip-seq signal of Pol2 from the BigWig file, collating it into 20 base pair bins, applying L_0 segmentation and storing the segmentation in a BedGraph file takes ~90 seconds with 4 cores. Short range local binning is reasonable for ChIPseq data as the signal is not inherently single nucleotide resolution. On the other hand, methylation data is single nucleotide and binomial error is more computationally demanding. This however is mediated by the fact that most methylation analysis is done on CpGs and there are only ~30 million CpGs in the human genome.

L_0 segmentation retains salient genomic features

To illustrate features of L_0 segmentation we plot segmentation results across tracks with diverse features that represent different signal classes: CTCF and K27ac for narrow peaks, K27me3 and K36me for broad peaks, and Pol2 which show a complex signal pattern composed of high narrow peaks and low broad peaks. Each track was segmented with L_0 Poisson loss with the regularization parameter determined by optimizing "offset" cross-validation, a specialized cross-validation procedure that accounts for frivolous autocorrelations that arise from ChIPseq fragments (see Methods).

We find that while neither the L_0 segmentation nor the cross-validation procedure is optimized for a particular signal type, our approach automatically fits the expected patterns

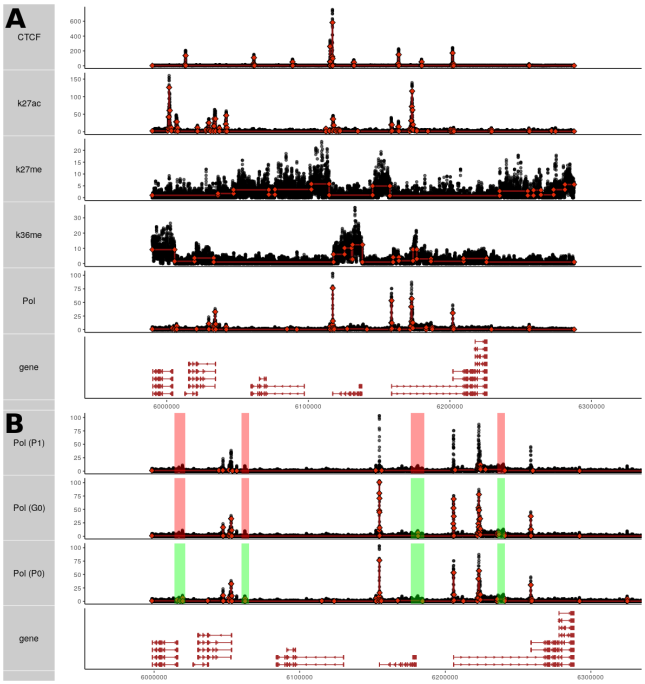

Fig. 2. A Poisson L_0 segmentation on various epigenetic tracks is shown. The hyperparameter (λ) is chosen automatically using cross validation. **B** Comparison of Poisson L_0(P0), Poisson L_1 (P1), and Gaussian L_0(G0) segmentations on RNA polymerase II ChIP-seq (Pol) track. The regions of interest are highlighted with green if the region is captured by the segmentation, otherwise the region is highlighted with red.

to each data type effectively finding narrow peaks and broad peaks.

We focus on a detailed investigation of the Pol2 track, which is composed of sharp peaks at the transcription start sites and low signal broad regions covering the entire transcript length. We assess how the different components of the segmentation problem contribute to signal discovery in this complex setting by comparing Poisson L_0, with Poisson L_1 and a Gaussian L_0 in Figure 2 . All segmentations are directly comparable as they contain exactly the same number of breakpoints and thus the same representation complexity. We note that this fact may not be obvious visually as L_1 segmentation generates many small segments that do not resolve as individual points on this plot. The L_1 with Poisson loss (top track) is indeed strikingly different from L_0 (bottom track). The segmented values typically fall far below the peak height and it fails to to accurately locate breakpoints at transcript ends (highlighted in green), which are marked by subtle change in signal. We also investigate the difference between using Poisson (P0, bottom track) and Gaussian (G0, middle track) likelihood as the reconstruction loss while using L_0 penalty. This difference is more subtle than the one between L_0 and L_1 but nevertheless reveals some important properties. Under the Poisson loss assumption variance is equal to the mean and thus high values are contribute less to the overall loss. The effect of this can be observed in that the G0 segmented output comes very close to peak max. A consequence of this is that under Gaussian loss more segments are used to approximate large peaks and thus under constant segment number fewer segments are available to capture smaller peaks. Local peaks that are captured by P0 and not by G0 are highlighted in purple.

To demonstrate these qualitative observations in a quantitative evaluation framework we make use of raw and processed dataset form the NIH Roadmap Epigenomics Consortium. The database provides an organized compilation of high quality uniformly processed raw epigenetic data as well as output form standard analysis tools such as MACS or ChromHMM [REF]. Using these well established analysis tools as a standard give us the opportunity to evaluate if we can distill different raw epigenetic tracks into a concise summary using a single principled algorithm. We focus on two cell lines that had all the relevant data types available; fibroblast cells, IMR90 and embryonic stem cells, H1.

The L_0 segmentation provides a reduced representation of a signal under the assumption that it is locally constant. As such, it does not explicitly optimize specific features. Alternative approaches to compressing epigenetic datasets are equal-width local binning or using specialized algorithms to explicitly detect specific features of interest such as "peaks". The MACS software [3] is highly optimized for peak discovery and we use the MACS peak-calls as a silver standard in order to evaluate the performance of L_0 compressed representations, which are in fact signal shape agnostic.

Specifically, we evaluate if reduced representations capture the peak-like features with high fidelity by formalizing a notion of representation efficiency, that is, the trade of between representation size and signal capture which for segmentation methods can be controlled by varying λ. Plotting the signal fidelity (which we define in multiple ways) against representation size for different compressed representations allows us to measure their efficiency independent of specific hyperparameter choices. For our peak based analysis we evaluate L_0 segmentation (our proposed solution), the convex L_1 segmentation, and equal size binning (a baseline that is commonly used to reduce epigenetic tracks in a data agnostic way). We report the representation size in terms of "compression ratio" which is the input length divided but the number of constant segments in the output.

In order to quantify how well the reduced representations retain peak features (signal fidelity) we formulate two different evaluations. The first one is based on computing the median fold change under the MACS peak regions after data compression. At the extreme value of compression ratio being equal to one (no compression) this just corresponds to the median fold change for the raw data. As the compression ratio is increased, this number will eventually decrease as peak regions are merged with adjacent regions. The results of this evaluation are plotted in Figure 3 (A,B). We find that according to this metric the L_0 method is able to maintain accuracy up to 10,000-fold compression while both the binning and L_1 show considerable drop-offs in performance much earlier.

The second evaluation is based on quantifying to what extent the reduced representation approximates the boundaries (as opposed to the values) of the peak calls. For each MACS peak we get the maximum Jaccard index across all possible overlapping segment regions. The distributions are plotted in 3 C,D and 90% quantile trends are plotted in the third row in panels E and F. In this evaluation, we see that the input (compression ratio equals one), where each point corresponds to a segment is actually not optimal. This is expected, as the input each point is a single nucleotide and thus will not have a high Jaccard index with peak regions that are on the order of hundreds of nucleotides long. Instead, the optimal Jaccard index is reached at an intermediate value for all methods, though with widely different compression ratios. Considering

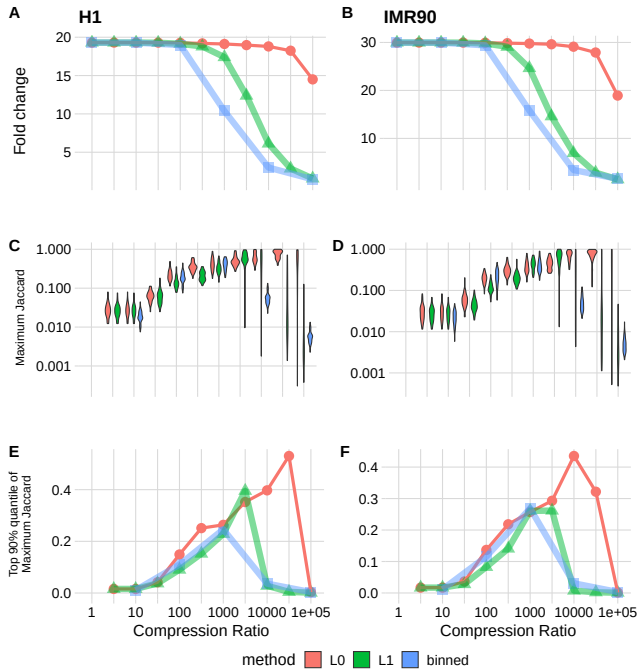

Fig. 3. Comparison of L_0 Poisson (red), L_1 Poisson (green), and fixed size binning (blue) with respect to the various compression ratios on the DNAse track from the H1 and IMR90 cell lines. The methods are applied on a 10M base-pair segment extracted from each cell line's Chromosome 1. We evaluated the methods using 937 peaks for H1 and 1400 peaks for IMR90 discovered by MACS within this segment. The vertical dashed black lines indicate the compression ratio needed to reproduce the exact peak regions. **A and B:** The ratio of the mean signal within peak regions to the mean background signal after segmentation. **C and D:** Maximum Jaccard Index distributions for the methods at different compression ratios. Jaccard indices are calculated between peaks and the segments discovered by the methods. **E and F:** Median of the distributions shown in C and D.

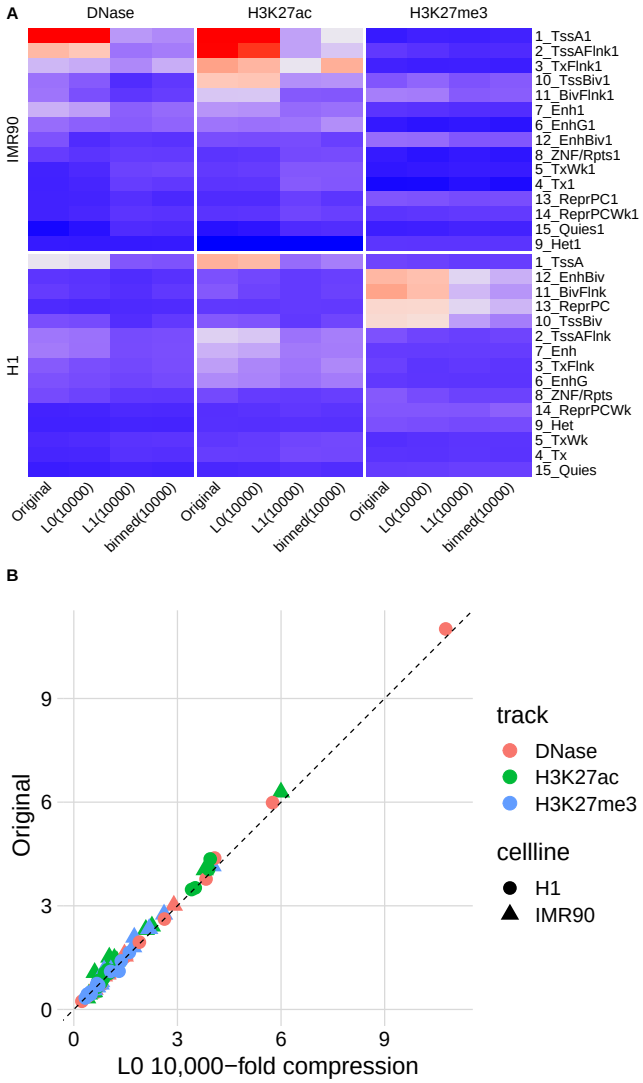

Fig. 4. The ratio of mean signal within ChromHMM clusters to the mean background signal after compressing the signal 10K fold using L_0Poisson (L_0), L_1 Poisson (L_1), and fixed size binning (binned). The results are shown for DNAse, H3K27ac, and H3K27me3 ChIP-seq tracks from H1 and IMR90 cell lines. For all tracks and cell lines, L_0Poisson retains the information content of the signal while L_1 Poisson and binning lose either the structure completely or the signal intensity at clusters where the original signal intensifies.

the H1 cell-line (C,E), local binning achieves optimal signal at a compression ratio of about 1000-fold where 90% of peaks have a Jaccard concordance value 0.2. For L_1 the corresponding values are about 3000-fold and 0.4, a considerable improvement. On the other hand L_0 achieves both the best quantile value of 0.5 and does so at a substantially higher compression rate of 3×10^4.

We also compare the compressed representation quality for three different epigenetic tracks (DNAse, H3K27ac, and H3K27me3) the three different methods with respect to enrichment from ChromHMM states at a compression ratio of 10^4, Figure 4. Importantly, we include a H3K27me3, which is a repressive mark that does not have peak structure but rather covers long genomics regions on the scale of genes. For the evaluation, the ChromHMM regions provided by the Roadmap project are fixed and we evaluate the enrichment of signal within a specific ChromHMM class using either the raw input (as is done in the original ChromHMM paper) or the compressed representation. We find that across the three compression methods, L_0 retains ChromHMM enrichment that is nearly identical to the input data 4 B.

In summary, we demonstrate that the L_0 Poisson compression is highly effective in retaining the salient signals from a diverse signal track types and a variety of different evaluations. Using L_0 compression, it is possible to reduce the complexity of the epigenetic signal by 10,000 fold while retaining features important for downstream interpretation. Thus, L_0 compression is a viable hypothesis free alternative to learning meaningful reduced representations of epigenetic data.

L_0 compression for methylation data

DNA methylation is a widely used assay type to profile the epigenetic state of cells and tissues. DNA methylation is unique among epigenetic signals in terms of biology, assay techniques, and analysis. Methylation is a covalent modification of the DNA which can be measured directly with bisulfite conversion and sequencing and does not require complex biochemical steps needed to perform chromatin based assays. Bisulfite conversion

can be used to add a methylation readout to other sequencing based assays as is done with single cell HiC [10].

Assays used to measure methylation in high throughput via probe hybridization have been in wide use for over a decade and have been applied to diverse biological problems from cancer characterization to epidemiological studies [11, 12]. The wide adoption of array based technologies has revealed how DNA methylation reports not just the state, but the developmental, and environmental history of the cell and organism as exemplified by the various methylation clock models (see [13] for review).

Because methylation is a covalent modification that is relatively stable and simple to assay, methylation models are reproducible across cohorts, tissues, and assay platforms [13, 14, 15] and can be constructed for large numbers of phenotypes [16]. The properties of methylation make it an attractive front-line epigenetic assay. Though sequencing cost is a major limitation, this issue is being rapidly addressed through technological innovation [17].

Unlike other epigenetic tracks methylation is truly single nucleotide resolution as it directly measures the methylation status of individual cytosines. Methylation data is reported as a count pair of methylated and unmethylated reads. The probability that a base is methylated (the ratio of methylated to total reads) is referred to as β. Importantly, the methylation signal doesn't have natural "peaks" that are the hallmark of other epigenetic tracks such as ATACseq and ChIPseq but is rather dispersed through the genome. Because the read coverage is scattered across the genome, the coverage at any single nucleotide is generally low but variable.

Extending the L_0 segmentation to binomial loss

NGS methylation data can be modeled as a series of draws from a binomial distribution with an unknown success probability. Considering the dual track input where y_i^1 is the number of methylated reads, and y_i^2 is the number of total reads can be derived as follows.

$$\underset{\alpha\in\mathbb{R}^n}{\operatorname{argmax}} \sum_{i=1}^{N} w_i \overbrace{\left(y_i^1 \alpha_i - y_i^2 \log\left(1+e^{\alpha_i}\right)\right)}^{\text{Binomial Error}} - \lambda \sum_{i=1}^{N-1} \underbrace{1\left\{\alpha_i - \alpha_{i+1}\right\}}_{L_0 \text{ Norm}} \tag{2}$$

Explicitly, modeling the binomial error accurately accounts for read coverage at each location, so that low-coverage segments are more likely to be merged with their neighbors. Secondly, because the L_0 does not depend on the size of the breakpoints this formulation readily creates single CpG segments if the methylation level is different from that of their neighbors and coverage is sufficient. We demonstrate these two properties of our formulation on a simulated dataset in Figure 5. In panel A we simulate a region that has a dramatically different β from its neighborhood but the value is not reliable because the coverage is low. Segmenting the β values directly with using squared error assigns this segment its own region. On the other hand, using the explicit two-track explicit binomial formulation, this region is merged with its neighbors (both segmentations have exactly three segments). In panel B we simulated a constant methylation region with two single CpG outlier segments which are readily discovered by L_0but not by L_1/fused-lasso.

We further illustrate the properties of binomial L_0 segmentation on real WGBS data in H1 and IMR90 cell lines. We consider a region near the POU5F1 locus which is an important pleuripotency gene and a major differentially methylated region between these two cell-lines highlighted in the original WGBS study[18]. In Fig. 6 we consider different segmentations of this region. Fixing chromosome level fold-reduction, we contrast L_0 and L_1. For reference, the differentially methylated region (DMR) of the POU5F1 promoter, discovered with supervised analysis, is shown in purple. Contrasting the L_0 and L_1 formulations we see that the unsupervised L_0 segmentation is able to identify the promoter DMR in the POU5F1 expressing cell-lie, H1, at up to 1000-fold compression. The promoter DMR is identified as a feature up to 100-fold reduction in the non-expressing cell-line IMR90.

The region around POU5F1 shows a common methylation pattern of expressed genes – the promoter tends to be hypomethylated while the gene body is hypermethylated. We thus evaluated how well these patterns are preserved globally

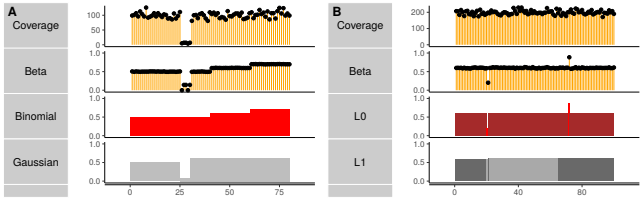

Fig. 5. Binomial L_0 segmentation accurately accounts for read coverage and discovers short regions with distinct methylation rates. We compare segmentation formulations on simulated data. Methylation and coverage are shown in yellow. **A:** Squared (gray) and binomial error segmentation (red) with the same number of segments. Binomial error merges low coverage regions with their neighbors. **B:** Comparing L_0 (red) and L_1 (gray). The L_0 penalty is more sensitive to local structure discovering the 2 CpGs simulated with different β

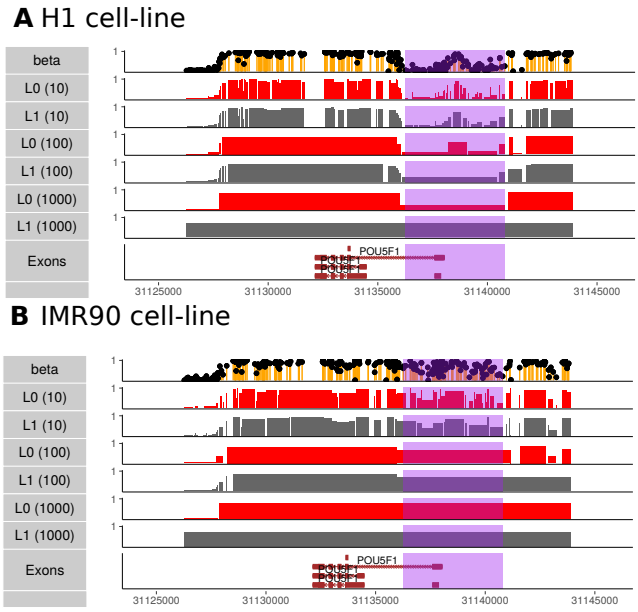

Fig. 6. Comparing L_0 to L_1 with binomial error on real WGBS data with different fold compression computed on the chromosome level. The reported differentially methylated region is highlighted in purple. A: POU5F1 expressing cell line, H1. B: The non-expressing contrasting cell-line, IMR90. Note that unlike Fig. 5, fold compression is fixed at the chromosome level and consequently the number of segments in this region is not constant across L_0 and L_1

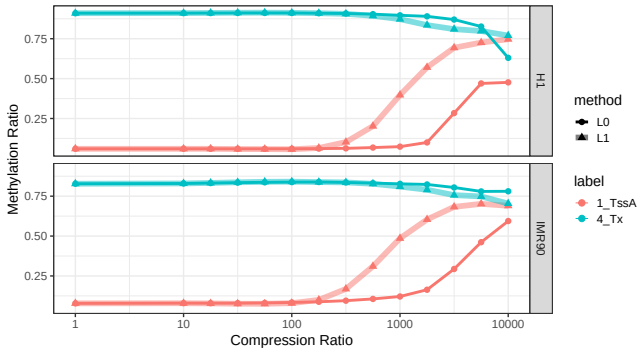

Fig. 7. Comparing L_0 to L_1 with binomial error on real WGBS data with different fold reduction. The reported differentially methylated region is highlighted in purple. A: POU5F1 expressing cell line, H1. B: The non-expressing contrasting cell-line, IMR90.

by using compressed representations. Similarly to our previous evaluations of compression for other epigenetic tracks we evaluate the trade-off between compression level and signal retention focusing on the chromoHMM states denoted as TssA (TSS) and Tx (transcript). As these are derived from other epigenetic tracks from the same cell-line they mark only the genes that are expressed in each. We find that using raw methylation input we observe the expected hypomethylation for TssA, and hypermethylation for Tx. The pattern is nearly unaltered as we increase the compression level up to 1000-fold for L_0. In contrast, for L_1/fused-lasso the TssA enrichment begins to decrease at a compression ratio of 100. In summary, we find that the binomial error L_0 segmentation we propose is an effective unsupervised feature reduction method for NGS methylation data. It naturally accounts for read coverage, can extract very small segments with distinct methylation patterns, and discover salient epigenetic features such as promoters and DMRs in a fully unsupervised manner.

L_0 compression is an effective general regularizer for epigenetic data

Beyond simply approximating its input as piece wise constant, segmentation methods can be used to enforce this structure inside more complex optimization problems. For example, the fused lasso constraint has been effectively used to find piece-wise constant solutions for matrix factorization problem [19].

We evaluate the performance of L_0 in a similar framework. We consider a rich dataset of 30 epigenetic tracks, originally used to evaluate an HMM based epigenetic segmentation algorithm Segway [20]. The L_0 and L_1 segmentations also solve a special type of HMM model [21] however they are not multivariate and extensions to multivariate formulations are not known to admit fast solutions. We thus formulate an L_0/L_1 regularized NMF problem which can simultaneously model multivariate data and provide piece-wise constant solutions. Formally, given a chromosome-region by epigenetic track matrix $\mathbf{A}$ we wish to solve the following NMF problem.

$$\underset{\mathbf{W}\geq 0,\mathbf{H}\geq 0}{\arg\min} \overbrace{D\left(\mathbf{A}\|\mathbf{W}\mathbf{H}^T\right)}^{\text{NMF objective}} + \underbrace{\lambda_1 \Omega\left(W\right)}_{\text{segmentation}} + \underbrace{\lambda_2 \left|\left|W\right|\right|_F + \lambda_3 \left|\left|H\right|\right|_F}_{L_2 \text{ penalties}} \tag{3}$$

Here the first term represents the KL (Poisson) divergence between $\mathbf{A}$ and the low rank reconstruction $\mathbf{W}\mathbf{H}^{\mathbf{T}}$. The second term represents a L_1, or L_0 fusion inducing penalty, and the last term, a small L_2 penalty, is often employed in standard NMF formulations. We note that the fusion penalties are applied column-wise to $\mathbf{W}$ and whose rows correspond to chromosomal positions. To solve this problem we implement a generalized alternating least square algorithm that can be used with arbitrary beta divergence losses including the Poisson-KL [22]. The core structure of this algorithm is a weighted regression step followed by a non-negative projection. The fusion inducing penalties are easily incorporated as additional projection steps as is done previously [19] allowing us to directly leverage our ultra-fast segmentation implementation. This formulation also allows us to swap L_0 and L_1 keeping all other algorithmic parameters the same.

We evaluate different NMF algorithms on an unsupervised feature discovery task by considering the transcription start site (TSS) binary indicator, defined as 2000 upstream and 500 downstream of each RefSeq gene, as a gold standard. Using the raw data, the maximal correlation with annotated TSS is achieved in the "Pol2_8WG16.Myers" track and is 0.25. After fitting a rank 10 standard NMF model the best performing factor increases this correlation to 0.34 indicating that a multi-track signature is able to improve TSS detection. Using the result of the standard NMF as a starting point, we run the NMF algorithm with fusion inducing regularization for an additional 10 iterations . To characterize the performance independently of regularization strength ($lambda_1$) we use a range of $lambda_1$ values. Results of these NMF runs are visualized in Fig. 8. As the $lambda_1$ value does not have the same meaning for L_0 and L_1, we compare the different penalties in terms of the final complexity measured as the total number of segments in the factor with the best TSS correlation. We find that adding L_0 regularization further improved the TSS discovery performance across a wide range of model complexities, while the addition of L_1 regularization does not.

It may seem unsurprising that forcing a piece wise constant solution improves correlation with ground truth, which is itself piece wise constant. However, regularization could also broadly lead to better disentanglement by forcing changes in track-specific loadings (matrix $\mathbf{H}$). To evaluate this we also plot the performance of the final "raw" $\mathbf{W}$ before the proximal operator is applied. We see that even the unregularized $\mathbf{W}$ is also improved. Visualizing the contributions of different tracks, from the corresponding column of $\mathbf{H}$, we see that while the top track is the same as the one with the best single track performance, adding regularization increases the contribution of H3K9ac and H3K27ac, both known to be associated with active promoters. Thus, the increase in performance is partially attributable to better disentanglement that is induced by introduction of the fusion inducing L_0 penalty.

Conclusion

Epigenetic data refers to a heterogeneous class of assay types that differ in their signal properties. However, all epigenetic data shares a local coherence structure as assay values at nearby nucleotides are determined by a common biological or technical process. This local structure can be leveraged in different ways via algorithms that are specialized for specific data types and data features. He we propose a general computational approach that can exploit local structure in an entirely data agnostic way. Our approach generalizes a recently proposed highly efficient L_0 segmentation algorithm to probability distributions relevant to

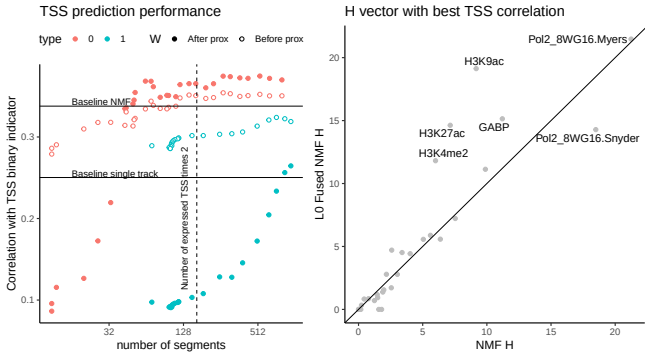

Fig. 8. Comparing L_0 to L_1 regularized NMF on unsupervised TSS discovery. L_0 regularization improves the TSS discovery benchmark across a broad range of regularization strengths while the identical algorithm with L_1 regularization does not. Introducing L_0 induced alterations in the non regularized loadings matrix $\mathbf{H}$ which favors a more disentangled representation

modeling epigentic data. We show that the L_0 segmentation efficiently represents epigenetic tracks while retaining many salient features such as peaks, promoters and ChromHMM states while making no assumptions about the underlying signal structure beyond piece wise constant. This makes our approach suitable for representing diverse track types, including ones that have "broad peak" or even complex structures such as Pol2, in a single unifying framework. Finally, we show that the L_0 regularizer and specifically our scalable implementation is a generally useful tool that can be applied inside more complex machine learning problems such as NMF.

Methods

Data

Roadmap datasets

For our quantitative evaluations of specific epigenetic data we used raw data, MACS peak calls, and ChromHMM results from the H1 and IMR90 cell-lines available in the Roadmap data portal.

Software

L_0 segmentation is implemented as a R package with C++ backend. The embedding of C++ in R is carried out by the Rcpp interface. The implementation is available at `https://github.com/boooooogey/l01segmentation`. Within R, `fusedsegmentation` is the main function that delegates segmentation tasks to various solvers. Through the arguments, users can specify a distribution (Gaussian/Poisson/binomial), a penalty (L_0/L_1) and change the format of the output. The second useful function is `compressBigwig` which is an end-to-end pipeline that takes a BigWig file and outputs a BedGraph with segments. Detailed of the algorithm and our implementation can be found in the appendix.

Binned cross-validation

Cross validation can be used to find optimal regularization parameters for fused lasso and related segmentation problems [19, 23]. In the standard cross-validation procedure individual points are left out of the computation and are estimated using the value of their segment. The best *lambda* is taken to be that which achieves minimal test error. If the data are independently sampled from the piece wise constant signal, this procedure will identify the correct λ that recovers the underlying structure without capturing the sampling noise. In the case of epigenetic assays that depend on immunoprecipitation (ChIP) independent sampling, conditioned on true signal, is not an accurate assumption. Since the protein is pulled down with a DNA fragment that is much larger than its actual occupancy site, regions of the DNA that are not directly bound but are simply near the site are likely to be captured. Moreover the probability of capture depends continuously on the proximity to the true occupied site inducing local dependency across adjacent sites violating independent sampling. Instead, the value at a single position is best predicted from the average of its immediate neighbors and using naive cross-validation results in a very large number of segments. In order to avoid fitting this local structure we propose a binned cross validation strategy. Instead of leaving out random or equally spaced elements the data is divided into windows whose width approximately matches that of expected ChIPseq fragment length (we use 300bp) and the test set is then taken to be equally spaced windows. Furthermore, the test region signal is reconstructed as the average of the left-most point in the adjacent left window and the right-most point in the adjacent right window.

To illustrates consider a region of length 6 and a cross-validation window of 2. One of the cross-validation test train splits will be as follows: train: $X_{\text{train}} = X_1, X_2, X_5, X_6$ and test: $X_{\text{test}} = X_3, X_4$. We perform segmentation on X_{train} and we predict $X_3 = X_4 = (\bar{X}_1 + \bar{X}_6)/2$ where $\bar{X}_1, \bar{X}_6$ are the predicted values.

We note that for methylation data, independent sampling is valid and regular cross-validation produced intuitive results.

NMF

To evaluate NMF we used the dataset from the paper describing Segway, a genome segmentation algorithm [20]. This dataset contains 30 different epigenetic tracks (CTCF.Bernstein, CTCF.Iyer, DNaseI., DNaseI.Stam, FAIRE, GABP, GATA1, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K9ac, H3K9me1, H4K20me1, JunD, Max, NFE2, NRSF, Pol2_8WG16.Myers, Pol2_8WG16.Snyder, Pol2_CTD4H8, Rad21, SRF, XRCC4, ZNF263, cFos, cJun, cMyc.Iyer, cMyc.Snyder). For the NMF problem we considered a 1Mb regions of chromosome 10 (chr10:6,200,000-7,200,000 in hg18 coordinates). The data was binned by summing the values in 20 nucleotide bins. We used the following NMF algorithm to solve for the NMF problem with KL divergence loss. This approach to the NMF problem is described in [22]. For fused NMF we add an additional projection step after the non-negative projection. We first fit a baseline model to the Segway 30 track data to convergence (relative change of $< 10^{-6}$). We then run the algorithm for 10 additional steps with the fusion inducing penalty using a variety of λ_1 values for both L_0 and L_1 regularization.

Competing interests

No competing interest is declared.

Author contributions statement

M.C. and A.T.B. conceived the approach, developed algorithms, analyzed the results, and wrote the paper.

Algorithm 1 L_0/L_1 fused NMF

Require: $A \in \mathbb{R}^{M \times N}$, $0 < K < \min(M, N)$, $0 < \lambda$
Ensure: $y = x^n$
$\boldsymbol{A'} = \boldsymbol{W}\boldsymbol{H}^T$
$\boldsymbol{E} = \boldsymbol{A} - \boldsymbol{A'}$
repeat
  $\boldsymbol{B} = (\boldsymbol{A'})^{-1}$
  for $k = 1, 2, ..., K$ **do**
    $\boldsymbol{A}^{(k)} = \boldsymbol{E} + \boldsymbol{w}_k \boldsymbol{h}_k^T$
    for j=1,2,...,N **do**
      $h_{jk} = \left[\frac{\sum_{i=1}^{M} b_{ij} a_{ij}^{(k)} w_{ik}}{\sum_{i=1}^{M} b_{ij} w_{ik} w_{ik} + \lambda_3} \right]_+$
    end for
    for i=1,2,...,M **do**
      $w_{ik} = \left[\frac{\sum_{j=1}^{N} b_{ij} a_{ij}^{(k)} h_{jk}}{\sum_{j=1}^{N} b_{ij} h_{jk} h_{jk} + \lambda_2} \right]_+$
    end for
    $\boldsymbol{W} = \text{prox}_{\lambda_1}(\boldsymbol{W})$ ▷ projection onto L_0/L_1 fused space.
    $\boldsymbol{E} = \boldsymbol{A}^{(k)} - \boldsymbol{w}_k \boldsymbol{h}_k^T$
  end for
  $\boldsymbol{A'} = \boldsymbol{W}\boldsymbol{H}^T$
until until stopping criterion is reached.

References

1. Ian Dunham, Anshul Kundaje, and et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. Number: 7414 Publisher: Nature Publishing Group.
2. Anshul Kundaje and et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330, February 2015.
3. Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):1–9, November 2008.
4. macs3 project. MACS, December 2022. [Online; accessed 15. Dec. 2022].
5. Kasper D. Hansen, Benjamin Langmead, and Rafael A. Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83, October 2012.
6. Yutaka Saito, Junko Tsuji, and Toutai Mituyama. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic Acids Research*, 42(6):e45, April 2014.
7. Xiaoqing Yu and Shuying Sun. HMM-DM: identifying differentially methylated regions using a hidden Markov model. *Statistical Applications in Genetics and Molecular Biology*, 15(1):69–81, March 2016.
8. Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E. Garrett-Bakelman, Maria E. Figueroa, Ari Melnick, and Christopher E. Mason. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, 13(10):R87, October 2012.
9. Yongseok Park, Maria E. Figueroa, Laura S. Rozek, and Maureen A. Sartor. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422, September 2014.
10. Guoqiang Li, Yaping Liu, Yanxiao Zhang, Naoki Kubo, Miao Yu, Rongxin Fang, Manolis Kellis, and Bing Ren. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat. Methods*, 16(10):991, October 2019.
11. Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, Emilie Calonne, Gianluca Bontempi, and François Fuks. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings Bioinf.*, 15(6):929, November 2014.
12. Zelin Jin and Yun Liu. DNA methylation in human diseases. *Genes & Diseases*, 5(1):1, March 2018.
13. Tessa Bergsma and Ekaterina Rogaeva. DNA Methylation Clocks and Their Predictive Capacity for Aging. *Neuroscience Insights*, 15, 2020.
14. Morgan Levine, Ross A. McDevitt, Margarita Meer, Kathy Perdue, Andrea Di Francesco, Theresa Meade, Colin Farrell, Kyra Thrush, Meng Wang, Christopher Dunn, Matteo Pellegrini, Rafael de Cabo, and Luigi Ferrucci. A rat epigenetic clock recapitulates phenotypic aging and co-localizes with heterochromatin. *eLife*, November 2020.
15. Todd R. Robeck, Zhe Fei, Ake T. Lu, Amin Haghani, Eve Jourdain, Joseph A. Zoller, Caesar Z. Li, Karen J. Steinman, Stacy DiRocco, Todd Schmitt, Steve Osborn, Bill Van Bonn, Etsuko Katsumata, June Mergl, Javier Almunia, Magdalena Rodriguez, Martin Haulena, Christopher Dold, and Steve Horvath. Multi-species and multi-tissue methylation clocks for age estimation in toothed whales and dolphins. *Commun. Biol.*, 4(642):1–11, May 2021.
16. Mike Thompson, Brian L. Hill, Nadav Rakocz, Jeffrey N. Chiang, , Sriram Sankararaman, Ira Hofer, Maxime Cannesson, Noah Zaitlen, and Eran Halperin. Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *medRxiv*, page 2022.02.07.22270047, February 2022.
17. Hayan Lee, Gat Krieger, Tyson Clark, Aziz Khan, Casey Ryan Hanson, Yizhou Zhu, Nasim Bararpour, Aaron M. Horning, Edward D. Esplin, Stephanie Nevins, Annika K. Weimer, Eti Meiri, Shlomit Gilad, Sima Benjamin, Danit Lebanony, Nika Iremadze, Florian Oberstrass, Ariel Jaimovich, William Greenleaf, James M. Ford, Doron Lipson, Zohar Shipony, and Michael P. Snyder. Ultra high-throughput whole-genome methylation sequencing reveals trajectories in precancerous polyps to early colorectal adenocarcinoma. *bioRxiv*, page 2022.05.30.494076, May 2022.
18. Ryan Lister, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A. Harvey Millar, James A. Thomson, Bing Ren, and Joseph R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462:315–322, November 2009.
19. Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, 10(3):515, July 2009.
20. Michael M. Hoffman, Orion J. Buske, Jie Wang, Zhiping Weng, Jeff A. Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*,

9:473–476, May 2012.

21. Nicholas A. Johnson. A Dynamic Programming Algorithm for the Fused Lasso and L0-Segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.
22. Liangda Li, Guy Lebanon, and Haesun Park. Fast bregman divergence NMF using taylor expansion and coordinate descent. In *KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 307–315. Association for Computing Machinery, New York, NY, USA, August 2012.
23. Sean Jewell and Daniela Witten. EXACT SPIKE TRAIN INFERENCE VIA 0 OPTIMIZATION. *annals of applied statistics*, 12(4):2457, December 2018.