Continuous chromatin state feature annotation of the human epigenome

Habib Daneshpajouh†, Bowen Chen†, Neda Shokraneh, Shohre Masoumi,
Kay C Wiese, Maxwell W Libbrecht

School of Computing Science, Simon Fraser University, Burnaby BC, Canada

†These authors contributed equally to this work.

1 Introduction

Segmentation and genome annotation (SAGA) methods are widely used to understand genome activity and gene regulation [5, 9, 14, 17, 26, 21, 30, 4]. These methods take as input a set of sequencing-based assays of epigenomic activity and output an annotation of the genome that assigns a chromatin state label to each genomic position. Existing SAGA methods have several limitations caused by the discrete annotation framework: such annotations cannot easily represent varying strengths of genomic elements, and they cannot easily represent combinatorial elements that simultaneously exhibit multiple types of activity.

In this work, we propose a continuous genome annotation strategy and a method, epigenome-ssm, that uses a non-negative Kalman filter state space model to efficiently annotate the genome. That is, our method outputs a vector of real-valued *chromatin state features* for each genomic position, where each chromatin state feature putatively represents a different type of activity. Continuous chromatin state features have a number of benefits over discrete labels. First, chromatin state features preserve the underlying continuous nature of the input signal tracks, so they preserve more of the information present in the raw data. Second, in contrast to discrete labels, continuous features can easily capture the strength of a given element. Third, chromatin state features can easily handle positions with combinatorial activity by assigning a high weight to multiple features. Fourth, chromatin state features lend themselves to expressive visualizations because they project complex data sets onto a small number of dimensions that can be shown in a plot.

The idea of the SSM method was presented at MLCB 2019. This manuscript presents the following contributions relative to that presentation:

- We present a version of SSM that can be applied to multiple cell types and annotations of eight cell types.
- We present a new comprehensive evaluation of SSM relative to genes and gene expression, in comparison to existing SAGA methods (ChromHMM, Segway).

2 Methods

2.1 State space model

We developed a Kalman filter state space model (SSM) [6] for annotating the genome with chromatin state features. This model takes as input a vector of E observed genomics data sets for each position, $y_g \in \mathbb{R}^E$, for $g \in 1 \ldots G$. This model assumes that at position g there is a latent vector $\alpha_g \in \mathbb{R}^M$ that encodes the chromatin state features of that position. It assumes that the observed data vector at that position (y_g) is generated as a linear function of α_g plus Gaussian noise:

$$y_g = Z\alpha_g + \epsilon_g \qquad \epsilon_g \sim N(0, I) \tag{1}$$

It further assumes that the latent vector α_{g+1} is generated as a linear function of α_g plus Gaussian noise

$$\alpha_{g+1} = T\alpha_g + v_g \qquad v_g \sim N(0, I) \tag{2}$$

To learn the SSM model, we use the expectation-maximization (EM) algorithm to maximize the log likelihood of the model as a function of its parameters, $Z \in \mathbb{R}^{E\times M}$ and $T \in \mathbb{R}^{M\times M}$. Briefly, this algorithm alternates two steps, the E step and the M step. In the E step, we hold Z and T fixed and use a message-passing algorithm to efficiently estimate $\alpha_{1:g}$ and compute sufficient statistics for updates to Z and T. In the M step, we use these sufficient statistics to update Z and T. We initialized $Z \sim \text{Uniform(0,1)}^{E\times M}$ and $T = I_M$.

To limit the model's capacity to overfit and its sensitivity to local optima, we additionally add several L_2 regularization terms to the optimization's objective function $J(Z, T)$, which encourage Z and T to have small values:

$$J(Z, T) = \log P(\alpha, Y|Z, T) + \lambda_1 \|Z\|_F + \lambda_2 \|T\|_F. \tag{3}$$

2.2 Non-negativity constraint

We developed a version of our model, epigenome-ssm-nonneg, in which the chromatin state features α_g and the emission parameters Z are both constrained to be nonnegative.

We used an active set method of Lagrange multipliers to enforce the nonnegativity constraint [12]. Specifically, we add two Lagrange multiplier terms to our objective function

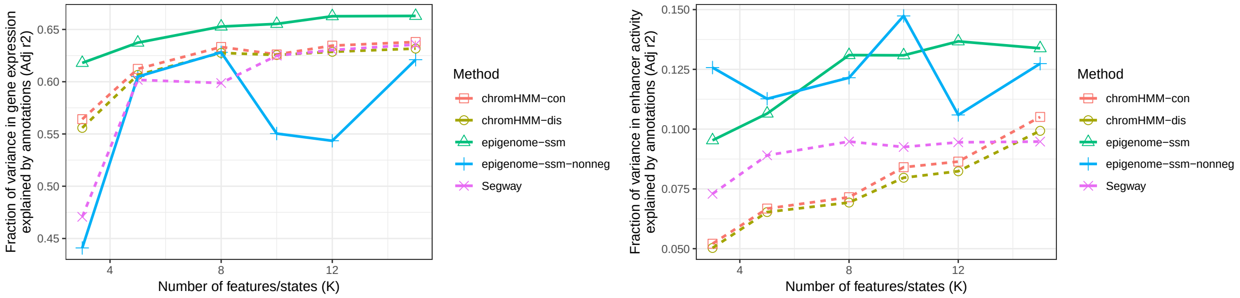
$$J_\Lambda(Z, T, \Lambda_Z, \Lambda_\alpha) = J(Z, T) + \text{tr}(\Lambda_Z^T Z) + \text{tr}(\Lambda_\alpha^T \alpha). \tag{4}$$

2.3 Alternative models

We compared epigenome-ssm with well-known annotation methods including Segway and chromHMM. Segway uses a Dynamic Bayesian Network which assumes that the observed data is generated as a multivariate Gaussian distribution. As with epigenome-ssm, Segway takes as input a vector of m observed genomic data sets for each position, $y_g \in \mathbb{R}^m$, and assigns a discrete label ℓ to each position g. chromHMM on the other hand, uses a hidden markov model (HMM) which assumes that the observed data is generated as a multivariate Bernoulli distribution, and thresholds input data into binary values such that the input data at position g is represented by a binary vector $\bar{y}_g \in \{0, 1\}^m$. We considered two versions of chromHMM, chromHMM-dis and chromHMM-con. For chromHMM-dis, we took the discrete labels (ℓ_g), while for chromHMM-con, we took the vector of posterior probabilities $\alpha_{\ell,g} = P(y_g = \ell)$ generated by chromHMM for each position g.

3 Results

3.1 Chromatin state features at genes are predictive of gene expression



(a) (b)

Figure (1) **Using annotations to predict (a) gene expression and (b) enhancer activity.**

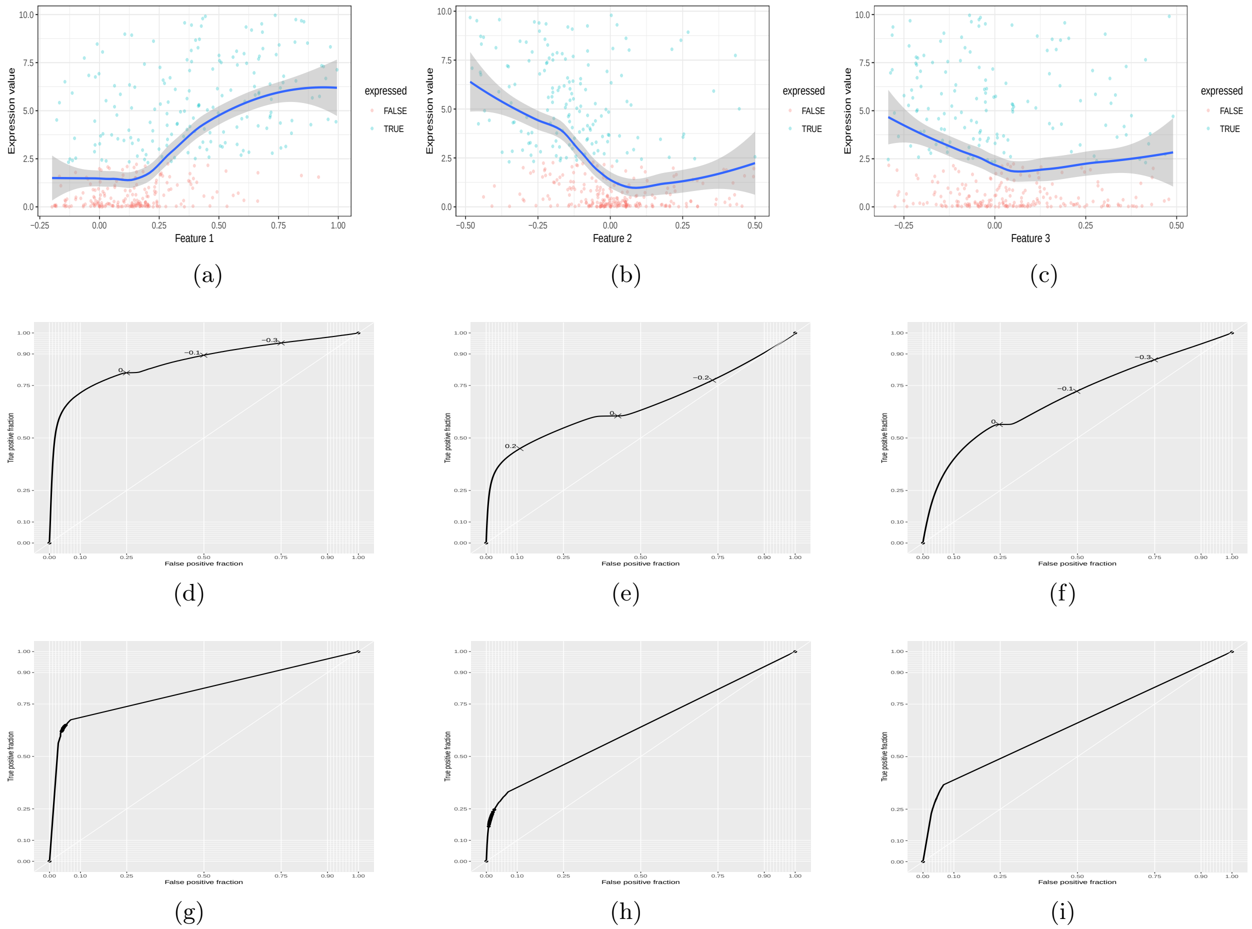
To evaluate our methods, we used the resulting annotations (generated using input data from eight cell types and twelve epigenomic assays) to predict RNA-seq gene expression data, following previous work [29, 18]. Briefly, We used a linear regression model to evaluate the degree to which annotations at a gene region are predictive of gene expression. We computed the average feature vector over the entire gene region [TSS, TTS]. As the regression response value, we used the RNA-seq RPKM (reads per kb per million mapped reads) values. We used the fraction of variance explained (adjusted r^2, also known as the coefficient of determination) to measure the predictive power of a regressor. We found that all methods are predictive of gene expression, but epigenome-ssm clearly outperforms alternatives by this measure (Fig. 1a). For example, with $k = 5$, an SSM annotation explains more variance (Adj $r^2 = 0.64$) than both chromHMM and Segway (0.61, 0.60 and 0.60 for chromHMM-con, chromHMM-dis and Segway respectively). Note that epigenome-ssm performance with a relatively small number of features (e.g. $k = 5$) is better than both chromHMM and Segway in all cases of k. Adding non-negativity constraints to epigenome-ssm reduces the performance slightly due to the model's restricted optimization space; however, the constrained model's performance is still comparable to other methods.

3.2 Chromatin state features at enhancer elements are predictive of enhancer activity

We further evaluated these annotation methods by measuring how predictive each annotation is of experimentally-validated enhancer elements, again following previous work [29]. We used FANTOM5[11] enhancer RNA data as a measurement of the activity of each enhancer element. As illustrated by Fig. 1b, epigenome-ssm and epigenome-ssm-nonneg perform significantly better than both chromHMM and Segway in this task.

3.3 Chromatin state features recapitulate known genome biology

We additionally found that epigenome-ssm features qualitatively recapitulate known genome biology. Fig. 2a-c show that there is a strong correlation between epigenome-ssm features and the expression values of the genes. Moreover, the ROC plots generated using epigenome-ssm features to predict whether a genomic position is TSS, part of a gene, or part of a enhancer (Fig. 2d-f) are clearly better than those generated using chromHMM-con features (Fig. 2g-i).



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure (2) (a-c) Relationship between epigenome-ssm features and gene expression. (d-f) ROC curves using epigenome-ssm features to predict whether a position is TSS, gene or enhancer respectively. (g-i) Same as (d-f) but using chromHMM-con features.

4 Discussion

In this work we explore the utility of chromatin state feature annotation. We propose a non-negative Kalman filter state space model for this problem, epigenome-ssm, that produces the highest-quality continuous annotations of the methods we compared. We also propose several measures of the quality of a chromatin state feature annotation and we compare the performance of several alternative methods according to these quality measures. While continuous features are somewhat more complicated to interpret than discrete labels, we showed that a small number of continuous features outperform even a large number of discrete labels in all of our evaluations. Therefore, a small number of chromatin state features can replace a much larger number of discrete labels, decreasing the overall complexity of the annotation. Moreover, chromatin state features are easy to interpret through visualizations. Because continuous annotations maintain much more of the information in the input data than discrete annotations do, they are more useful for complex downstream applications. For example, a variant effect predictor might take chromatin state features as input in order to predict the functional impact of a given mutation. In the future, we plan to apply this approach to create reference chromatin state feature annotations for all tissues with sufficient available data.

References

[1] Haley M Amemiya, Anshul Kundaje, and Alan P Boyle. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". In: Scientific Reports 9.1 (2019), p. 9354.

[2] Bradley E Bernstein et al. "The NIH roadmap epigenomics mapping consortium". In: Nature Biotechnology 28.10 (2010), p. 1045.

[3] Jacob Biesinger, Yuanfeng Wang, and Xiaohui Xie. "Discovering and mapping chromatin states using a tree hidden Markov model". In: BMC Bioinformatics 14.5 (2013), S4.

[4] Simon G Coetzee et al. "StateHub-StatePaintR: rapid and reproducible chromatin state evaluation for custom genome annotation". In: F1000Research 7 (2018).

[5] Nathan Day et al. "Unsupervised segmentation of continuous genomic data". In: Bioinformatics 23.11 (2007), pp. 1424–1426.

[6] James Durbin and Siem Jan Koopman. Time series analysis by state space methods. Vol. 38. Oxford University Press, 2012.

[7] Timothy J Durham et al. "PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition". In: Nature Communications 9.1 (2018), p. 1402.

[8] Jason Ernst et al. "Mapping and analysis of chromatin state dynamics in nine human cell types". In: Nature 473.7345 (2011), p. 43.

[9] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: Nature Methods 9.3 (2012), p. 215.

[10] Guillaume J Filion et al. "Systematic protein location mapping reveals five principal chromatin types in Drosophila cells". In: Cell 143.2 (2010), pp. 212–224.

[11] Alistair RR Forrest et al. "A promoter-level mammalian expression atlas". In: Nature 507.7493 (2014), p. 462.

[12] Nachi Gupta and Raphael Hauser. "Kalman filtering with equality and inequality state constraints". In: arXiv preprint arXiv:0709.2791 (2007).

[13] Michael M Hoffman et al. "Integrative annotation of chromatin elements from ENCODE data". In: Nucleic Acids Research 41.2 (2012), pp. 827–841.

[14] Michael M Hoffman et al. "Unsupervised pattern discovery in human chromatin structure through genomic segmentation". In: Nature Methods 9.5 (2012), p. 473.

[15] Anshul Kundaje et al. "Integrative analysis of 111 reference human epigenomes". In: Nature 518.7539 (2015), p. 317.

[16] Thomas K Landauer et al. Handbook of latent semantic analysis. Psychology Press, 2013.

[17] Jessica L Larson et al. "A tiered hidden Markov model characterizes multi-scale chromatin states". In: Genomics 102.1 (2013), pp. 1–7.

[18] Maxwell Wing Libbrecht et al. "A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types". In: BioRxiv (2016), p. 086025.

[19] Maxwell W Libbrecht et al. "Entropic graph-based posterior regularization". In: Proceedings of the Internationa 2015.

[20] Maxwell W Libbrecht et al. "Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression". In: Genome Research (2015).

[21] Alessandro Mammana and Ho-Ryun Chung. "Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome". In: Genome Biology 16.1 (2015), p. 151.

[22] Karl Pearson. "On lines and planes of closest fit to systems of points in space". In: The London, Edinburgh, and 2.11 (1901), pp. 559–572.

[23] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: Journal of Machine Learning Research 12 (2011), pp. 2825–2830.

[24] Jacob Schreiber. "Pomegranate: fast and flexible probabilistic modeling in Python". In: The Journal of Machine Learning Research 18.1 (2017), pp. 5992–5997.

[25] Jacob Schreiber et al. "Multi-scale deep tensor factorization learns a latent representation of the human epigenome". In: bioRxiv (2018), p. 364976.

[26] Kyung-Ah Sohn et al. "hiHMM: Bayesian non-parametric joint inference of chromatin state maps". In: Bioinformatics 31.13 (2015), pp. 2066–2074.

[27] Suvrit Sra and Inderjit S Dhillon. "Generalized nonnegative matrix approximations with Bregman divergences". In: Advances in Neural Information Processing Systems. 2006, pp. 283–290.

[28] Yu Zhang and Ross C Hardison. "Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation". In: Nucleic Acids Research 45.17 (2017), pp. 9823–9836.

[29] Yu Zhang et al. "Jointly characterizing epigenetic dynamics across multiple human cell types". In: Nucleic Acids Research 44.14 (2016), pp. 6721–6731.

[30] Jian Zhou and Olga G Troyanskaya. "Probabilistic modelling of chromatin code landscape reveals functional diversity of enhancer-like chromatin states". In: Nature Communications 7 (2016), p. 10528.