Entropic Graph-based Posterior Regularization for Learning Probabilistic Models

Maxwell W. Libbrecht
Computer Science and Engineering
University of Washington
maxwl@cs.washington.edu

Michael M. Hoffman
Genome Sciences
University of Washington
mmh1@uw.edu

Jeffrey A. Bilmes
Electrical Engineering
University of Washington
bilmes@ee.washington.edu

William S. Noble
Genome Sciences
University of Washington
noble@gs.washington.edu

The advent of high-throughput DNA sequencing methods has led to an explosion in the availability of genome-wide data and a corresponding opportunity to use computational methods to derive insights into cellular function. A predominant form of statistical model used for genomic data has been temporal, such as the hidden Markov model (HMM) (Rabiner, 1989) or dynamic Bayesian network (DBN) (Dean and Kanazawa, 1988). Temporal models consist of a chain of random variables exhibiting the temporal Markov property, which asserts that "future" and "past" variables are independent given some notion of the present. This property enables exact inference to be performed using temporal dynamic programming. Most often, inference takes the form of a forward and then a backward pass along the chain. In genomics, the "temporal" axis is generally position along the genome rather than actual time.

Because the human genome is billions of bases long, applying these inference methods to genomics requires scaling to extremely large models. Although the Markov property often enables tractable inference even on large models (Binder et al., 1997), the strength of dependence between pairs of variables decays rapidly as a function of distance along the chain. By contrast, real data sometimes exhibit strong direct correlations between activity at distant parts of the genome, caused by spatial proximity in the cell or co-regulated processes. In such cases, Markov models are likely to underperform. On the other hand, if we indiscriminately add the ability for very distant parts of the chain to interact directly, then we lose the temporal Markov property, and inference costs can become exponential in the length of the chain. Hence, for genomic applications, we need methods that maintain the tractability and strong backbone dependence of temporal models but allow for interaction between distant variables in a chain.

We propose an approach for expressing long-range dependencies between pairs of variables in probabilistic models in which we add penalty based on the variables' Kullback-Leibler (KL) divergence. We term this combination of probabilistic models and KL penalties *entropic graph-based regularization* (EGBR). The objective function of this model can be optimized efficiently using a three-way alternating minimization algorithm. We apply this model to improve existing methods for annotating the human genome (Hoffman et al., 2012a; Ernst and Kellis, 2010) using data from the NIH-sponsored ENCODE project ((ENCODE Project Consortium, 2012; Hoffman et al., 2012b), http://www.nature.com/encode) and show that this improved model predicts functional elements much more accurately than existing temporal methods. This approach provides a method for functionally annotating the human genome in the hundreds of cell types for which only limited data is available.

1 MODEL AND OBJECTIVE

In a general graphical model, we have a set of observed random variables $X = \bar{x} \in \mathcal{X}$, a set of n hidden variables $Y_1 \dots Y_n = Y_{1:n} = Y \in \mathcal{Y}^n$ and a statistical model parameterized by θ that yields a distribution $\Pr(\bar{x}, Y|\theta)$ over $\mathcal{X} \times \mathcal{Y}^n$. In this work, we require[1] that each Y_i has the same discrete domain $\mathcal{Y}$. We are interested both in choosing θ such that the likelihood of the observations $\Pr(\bar{x}|\theta)$ is maximized (the "training" or "model selection" problem), and in finding the assignment to (or posterior distribution over) the hidden variables $Y_{1:n}$ (the inference problem). One of the most popular ways of doing this is using the expectation-maximization (EM) algorithm.

Optimization formulation of the EM algorithm Neal et al. (Neal and Hinton, 1999) formulated EM as the following optimization problem. Define $\mathcal{C}_{\text{EM}}(r, \theta)$ to be a cost function associated with EM,

$$\mathcal{C}_{\text{EM}}(r, \theta) \triangleq -E_r[\log \Pr(\bar{x}, Y|\theta)] - H(r), \tag{1}$$

where $r \in \Delta_{n \cdot k}$ is a distribution over $\mathcal{Y}^n$ and $H(p)$ is the Shannon entropy $H(p) = -\sum_{y \in \mathcal{Y}^n} p(y) \log p(y)$ of a distribution p. Define the E-step and M-step as:

$$\textbf{E-step:} \quad \text{Compute } r^{(t)} \leftarrow \operatorname{argmin}_r \mathcal{C}_{\text{EM}}(r, \theta^{(t-1)})$$

$$\textbf{M-step:} \quad \text{Compute } \theta^{(t)} \leftarrow \operatorname{argmin}_\theta \mathcal{C}_{\text{EM}}(r^{(t)}, \theta).$$

It can be shown (Neal and Hinton, 1999) that the E-step sets $r^{(t)}(y) \leftarrow \Pr(y \mid \bar{x}, \theta^{(t-1)})$.

This objective is easy to optimize as long as $\Pr(\bar{x}, y_{1:n}|\theta)$ has low treewidth, but cannot be exactly optimized tractably if this is not the case. In the present work, we generalize the EM objective by adding terms that create a form of interaction between arbitrary pairs of variables while maintaining tractable inference. This is accomplished by defining a new type of regularization penalty that is inspired by the objective of a recently-developed semi-supervised learning technique (Subramanya and Bilmes, 2011), as described in the next section.

Joint objective We propose a joint objective that combines the EM objective with a regularizer that encourages certain pairs of variables to have similar posterior distributions. Let $W \in \mathbb{R}^{n \times n}$ be an $n \times n$ matrix that defines a weighted, directed graph on n nodes, where w_{ij} represents desired similarity between Y_i and Y_j. We do not constrain W to be symmetric, although it may be.

The objective is:

$$\begin{aligned} \mathcal{C}_{\text{EGBR}}(r, \theta, p) \triangleq & -E_r[\log \Pr(\bar{x}, Y|\theta)] - H(r) \\ & + \gamma D_{\text{KL}}(r \| \Pi_i p_i) + \mu \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(i)} w_{ij} D_{\text{KL}}(p_i \| p_j) \\ & - \nu \sum_{i=1}^{n} H(p_i), \end{aligned} \tag{2}$$

where $p \in \Delta_k^n$ is a fully factorizable distribution over $\mathcal{Y}^n$ (i.e., $p(y) = \prod_{i=1}^n p_i(y_i)$ where each p_i is a distribution over $\mathcal{Y}$), $\mathcal{N}(i)$ is the set of neighbors of node i with $w_{ij} > 0$, and γ, μ and ν are hyperparameters. $D_{\text{KL}}(p\|q)$ is the Kullback-Leibler divergence from p to q, $D_{\text{KL}}(p\|q) = \sum_{y \in \mathcal{Y}^n} p(y) \log(p(y)/q(y))$. We do not require any factorization properties on r. The first two terms in Equation (2) constitute an EM-like objective, the third term encourages r and p to be similar, and the fourth and fifth terms are regularizers that express graph similarity and high-entropy objectives.

We have developed a novel alternating minimization algorithm for efficiently optimizing $\mathcal{C}_{\text{EGBR}}$. This algorithm alternates between optimizing $\mathcal{C}_{\text{EGBR}}$ in r using an algorithm for probabilistic inference such as the junction tree algorithm (which is linear in the size of the model for chain- or tree-structured models) and optimizing $\mathcal{C}_{\text{EGBR}}$ in p by applying an algorithm first developed for the

[1] The results presented here can easily be extended to the case where we have multiple classes of variables such that each variable in the same class has the same domain, and long-range dependencies occur only between pairs of variables of the same class. However, we restrict ourselves to one class here for simplicity.

measure propagation (Subramanya and Bilmes, 2011) method (itself an alternating-minimization algorithm, which applies updates linear in the degree of the graph). Importantly, the efficiency of this algorithm is not influenced by the treewidth of the graph W. This algorithm converges to the global optimum of $\mathcal{C}_{\text{EGBR}}$ in r and p, and maintains monotone convergence when θ are updated using the EM algorithm's M-step.

2 EXPERIMENTS

In this section we evaluate the performance of EGBR. First, we compare EGBR to loopy belief propagation on synthetic data. Second, we apply EGBR to the problem of performing genome annotation in multiple cell types and demonstrate that it improves performance in discovering regulatory elements.

Table 1: Comparison of EGBR and LBP on Synthetic Data.

σ	CHAIN	LBP	EGBR
0.25	0.989 (0.006)	1.000 (0.001)	0.999 (0.001)
0.5	0.896 (0.027)	0.967 (0.012)	**0.981** (0.009)
1.0	0.713 (0.026)	0.739 (0.031)	**0.844** (0.057)
1.5	0.636 (0.036)	0.636 (0.034)	**0.735** (0.050)
2.0	0.603 (0.030)	0.607 (0.037)	**0.680** (0.064)

Comparison with loopy belief propagation A natural experiment is to compare EGBR to approximate inference on a graphical model with the same dependence structure. We compare to loopy belief propagation (LBP) because it is one of the most widely used approximate inference methods. While we would have preferred to perform this comparison using our genomics data sets, it appeared that even our fastest LBP implementation would take months to converge. Therefore, we instead performed this comparison using synthetic data. We generated a chain of length $n = 300$, with $Y_{1:n} \in \{0, 1\}^n$ set to alternating length-5 segments of label 0 and 1 respectively (in other words, $y_i = \text{floor}(i/5) \bmod 2$). We generated observations $X_{1:n}$ as $X_i \sim N(Y_i, \sigma)$, where we vary σ to control the difficulty of the problem—higher σ results in more challenging inference. We defined an HMM over this chain with transition probabilities $\Pr(Y_i = Y_j) = 0.8$ and emission probabilities as above. In addition, we generated a graph W over the vertices of the chain by setting $w_{ij} = 1$ with probability 0.2 if $Y_i = Y_j$, $w_{ij} = 1$ with probability 0.1 if $Y_i \neq Y_j$, and $w_{ij} = 0$ otherwise. This model is meant to simulate the task of labeling a chain (such as a genomic sequence) where we have noisy information about which pairs of positions have the same label.

We compared three methods of inference: 1) inference on the chain alone, without using W, 2) LBP on the chain plus extra factors of $\Pr(Y_i = Y_j) = \text{sigmoid}(\lambda w_{ij})$, where λ controls the strength of these factors; and 3) EGBR using the graph W. For EGBR, we used the hyperparameters $\gamma = 10^{-2}$, $\mu = 10^{-7}$, $\nu = 0$. In order to give LBP the greatest advantage possible, for each value of σ we varied λ from 10^{-10} to 1 and picked the value that produced the best performance. The results are shown in Table 1. In this table, values show the average accuracy over 30 simulations (standard deviation in parentheses) of MAP inference using the model in question. In each row, the highest accuracy is in bold if it is significantly larger than the second-highest accuracy (Wilcoxon test, $p < 0.05$). EGBR has the best performance for all experiments, and significantly outperforms the other methods for large values of σ. These results are likely due to the fact that EGBR finds the global optimum of its objective, while LBP is an approximate method.

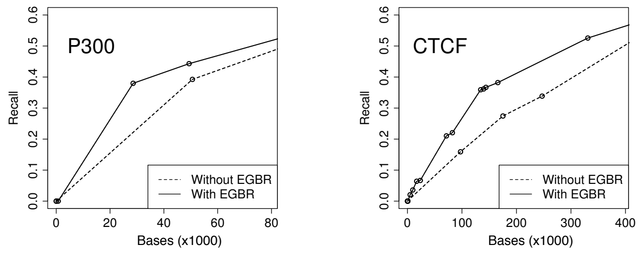

Figure 1: Detecting Enhancers and Insulators

Genome annotation in multiple cell types Next we turn to our primary application, using EGBR to improve existing methods for semi-automated annotation of the human genome. Recently, many methods have been described that partition and label the human genome on the basis of a number of genome-wide real-valued signal tracks, generally employing temporal models such as HMMs (Hoffman et al., 2012a; Ernst and Kellis, 2010; Thurman et al., 2007; Lian et al., 2008; Jaschek and Tanay, 2009). Formally, these methods aim to learn a labeling $Y_{1:n} \in \{1..L\}^n$ which associates each position in the genome with one of L integer labels, such that positions that receive the same label exhibit similar patterns in the signal data. The input is comprised of a feature vector $X_i \in \mathbb{R}^F$ at each position that represents the output of biological assays that measure local properties of the DNA, including its interaction with binding proteins, its local structure, and various types of chemical modifications. The process is "semi-automated" because a human performs a functional interpretation of the integer labels subsequent to the unsupervised learning phase.

Existing methods for semi-automated genome annotation work well on data from a single cell type, but annotating multiple cell types remains an active area of research. There are several possible strategies for performing annotation of multiple cell types. The simplest strategy is to apply the same model to both genomes (Sheffield et al., 2013), but this requires that all cell types have the same set of available data, which is not generally true. Alternatively, one could perform annotation separately on each cell type and find a mapping between the labels (for example, by using the Hungarian algorithm). However, since different cell types generally have different types of activity and different sets of signal tracks, such a mapping is generally very poor. Moreover, neither of these approaches utilize evidence across cell types in the annotation process.

These two problems—requiring a common set of data and failure to integrate evidence—are especially important because, although biologists would like to understand a large number of cell types, very limited numbers of experiments have been performed in most of these cell types due to the cost of genomic experiments. For example, ENCODE has performed 335 experiments in its most-studied cell type, but has performed just 2-10 experiments in more than 100 cell types.

We employ EGBR to perform annotation of multiple cell types in a way that integrates evidence between cell types. In order to remove the requirement of a one-to-one mapping between labels, we take a general approach in which we simply encourage that if two positions got the same label in cell type A, then they should be more likely to get the same label in cell type B. We express this relationship by connecting the two positions with an edge in the EGBR graph used when annotating cell type B.

To demonstrate this approach, we extended a recent semi-automated genome annotation algorithm, Segway, which uses EM on an HMM-like model to annotate the genome (Hoffman et al., 2012a). We first ran Segway without EGBR on a "reference" cell type *K562* to produce an annotation A_{ref}. Following the protocol used by the ENCODE consortium (ENCODE Project Consortium, 2012), we trained our model on data from 11 signal tracks available through ENCODE: three measurements of DNA accessibility (FAIRE-seq and two protocols for DNase-seq) and eight measurements of covalent modifications of histone proteins . We annotated at 1 base pair resolution 20 million bases ($\sim 1\%$) of the genome with 50 labels. We constructed an EGBR graph from this annotation by connecting each pair of positions that received the same label in A_{ref} with a pair of directed edges of weight 1. To mitigate the problem of quadratic growth in the degree of this graph, we subsampled this graph such that each node had outgoing degree $17 \approx \log_e(2 \times 10^7)$. We chose this graph degree because well-known properties of random graphs ensure that a randomly-subsampled graph with $n \log n$ edges has the same connected components as the full graph with high likelihood, and our experiments on synthetic data showed that the sparse graph performed similarly to a complete graph. We then used Segway with EGBR to annotate a "target" cell type *GM12878*. To simulate the case where the target cell type has limited data available, we omitted the most informative tracks, measures of DNA accessibility, and used just eight tracks measuring histone modification. Following the ENCODE protocol, we used 25 labels for the target annotation.

We evaluate whether EGBR improved the model's ability to identify known biological phenomena in the genome. Since it is hard to rigorously evaluate the unsupervised label choice task, we frame our validation as the problem of semi-supervised detection of two types of regulatory elements: *enhancers* and *insulators*. While these types of regulatory elements have been known about for decades, accurately determining their locations in the genome is an active area of research, for which temporal models represent the state of the art (Hoffman et al., 2012a; ENCODE Project Consortium, 2012; Marsman and Horsfield, 2012). Because there are few known examples of regulatory elements, we used the experimentally-determined binding sites of two proteins as proxies for regulatory element locations: P300 for enhancers and CTCF for insulators. These substitutions are imperfect, but represent a good approximation (Visel et al., 2009; Burgess-Beusse et al., 2002).

We split the annotation into validation and test sets of 10 million bp each. For each type of regulatory element, we ranked the labels by the ratio of elements covered to bases annotated in the validation set, then used this ranking to order all positions in the test set by their predicted likelihood of being the type of regulatory element in question. We chose values for the EGBR hyperparameters according to the number of bases required to reach 50% recall on the validation set. Figure 1 shows our accuracy on the test set with and without EGBR (dotted and sold lines, respectively). Each panel plots the fraction of elements detected as a function of the number of bases annotated. We plot recall up to 1000 bp times the number of elements. For both targets, EGBR significantly outperforms inference without regularization.

References

Binder, J., Murphy, K., and Russell, S. (1997). Space-efficient inference in dynamic probabilistic networks. *Int'l, Joint Conf. on Artificial Intelligence*, 1(5):1292–1296.

Burgess-Beusse, B., Farrell, C., Gaszner, M., Litt, M., Mutskov, V., Recillas-Targa, F., Simpson, M., West, A., and Felsenfeld, G. (2002). The insulation of genes from external enhancers and silencing chromatin. *PNAS*, 99(Suppl 4):16433.

Dean, T. and Kanazawa, K. (1988). Probabilistic temporal reasoning. *AAAI*, pages 524–528.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012a). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476.

Hoffman, M. M. et al. (2012b). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, 41(2):827–41.

Jaschek, R. and Tanay, A. (2009). Spatial clustering of multivariate genomic and epigenomic information. In *Proc. of RECOMB 2009*, volume 5541, pages 170–183.

Lian, H., Thompson, W., Thurman, R. E., Stamatoyannopoulos, J. A., Noble, W. S., and Lawrence, C. (2008). Automated mapping of large-scale chromatin structure in encode. *Bioinformatics*, 24(17):1911–1916. PMC2519158.

Marsman, J. and Horsfield, J. (2012). Long distance relationships: Enhancer-promoter communication and dynamic gene transcription. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*.

Neal, R. and Hinton, G. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. MIT Press.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Sheffield, N. C., Thurman, R. E., Song, L., Safi, A., Stamatoyannopoulos, J. A., Lenhard, B., Crawford, G. E., and Furey, T. S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome research*, 23(5):777–788.

Subramanya, A. and Bilmes, J. (2011). Semi-supervised learning with measure propagation. *Journal of Machine Learning Research*, 12(10):33113370.

Thurman, R. E., Day, N., Noble, W. S., and Stamatoyannopoulos, J. A. (2007). Identification of higher-order functional domains in the human encode regions. *Genome Research*, 17:917–927.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858.