TCF21 Binding Sites Characterization using Latent Dirichlet Allocation

Tyler Davis, Eric Kofman, Sean Scott

Abstract

Transcription factors play multiple roles in cell activity and gene expression, and discovering these roles often requires experimentation in a wet lab. We hope to bypass this process computationally by using topic modeling to infer the myriad of functions of a given transcription factor. Specifically, we apply Latent Dirichlet Allocation (LDA) to all peaks derived from running ChIP-seq on TCF21 in fetal heart cells, and model each peak as a mixture of topics composed of known binding motifs and cell types. We cluster peaks with similar topic distributions together, and we leverage GREAT to demonstrate that a significant fraction of the resulting clusters have functional biological significance. Functional significance of the peak clusterings implies that we have found functional subgroups of sites that bind TCF21, and thus functions in which TCF21 is involved. Furthermore, we also show that the topics deduced by LDA can associate motifs with cell activity in a variety of human cells and thus also yield useful information. These results are summarized in Figure 1

I. Introduction

I. Biological Background

One of the many fascinating aspects of biology is the way in which an organism's cells are able to to differentiate themselves from one another, such that a cell in the liver knows to perform different functions from a cell in the lung. This is remarkable because each cell in an organism bases its production of proteins on the same genetic blueprint, and the underlying implication is that each cell must somehow repress certain genes and activate others depending on what function it needs to serve. Proteins called transcription factors serve as the switches in this process by binding to certain stretches of DNA. For example, an activating complex of transcription factor proteins might form a scaffold that heightens the probability of RNA polymerase binding and accordingly of transcription occurring, and conversely, a repressing transcription factor might block key RNA polymerase binding sites. A single transcription factor can play roles in different biological pathways depending on its binding environment and the presence of other protein factors with which it may complex. More specifically, examining the sequences in the region surrounding these peaks (other potential binding sites, or "motifs") gives insight into the roles played by nearby transcription factors. ChIP-seq (Chromatin Immunoprecipitation sequencing) is a technology employed to discover these transcription factor binding sites, and its output is in the form of "peaks" representing regions at which transcription factors have a high probability of binding along a genome [4]. More specifically, the region around each ChIP-seq peak is a combination of certain binding motifs. That is, different combinations of the same short sequences will bind to different sets of transcription factors, and these related but different sets of factors imply different biological functions. Thus, we can view this information as a mixed membership model [5]. For this reason, we use Latent Dirichlet Allocation as a an algorithm to both generate topics of motifs and cluster peaks based on their motifs.

II. Applicability of Latent Dirichlet Allocation

LDA's original use case takes a corpus of documents and a vocabulary as inputs and operates within three-layer framework: each document is comprised of a mixture of topics, and each topic is a distribution over the words in the vocabulary [1]. LDA generates the word composition of each topic and the topic-composition of each document. Each binding site for a transcription factor may play roles in different biological processes, and assigning a mixture of topics to each binding site provides a more realistic description of a transcription factor's versatility. Consequently, uncovering the structure underlying ChIP-seq data requires a topic modeling approach such as that employed by LDA as opposed to a simpler clustering algorithm such as k-means which does not employ this mixed-membership model. We used ChIP-seq results for transcription factor 21 (TCF21) in fetal heart cells, a dataset of known common binding motifs (short sequences of DNA with which transcription factors often interact) and a dataset of known activity levels of DNA regions in 127 cell types. We set out to use LDA to develop a topic model for the ChIP-seq peaks that would

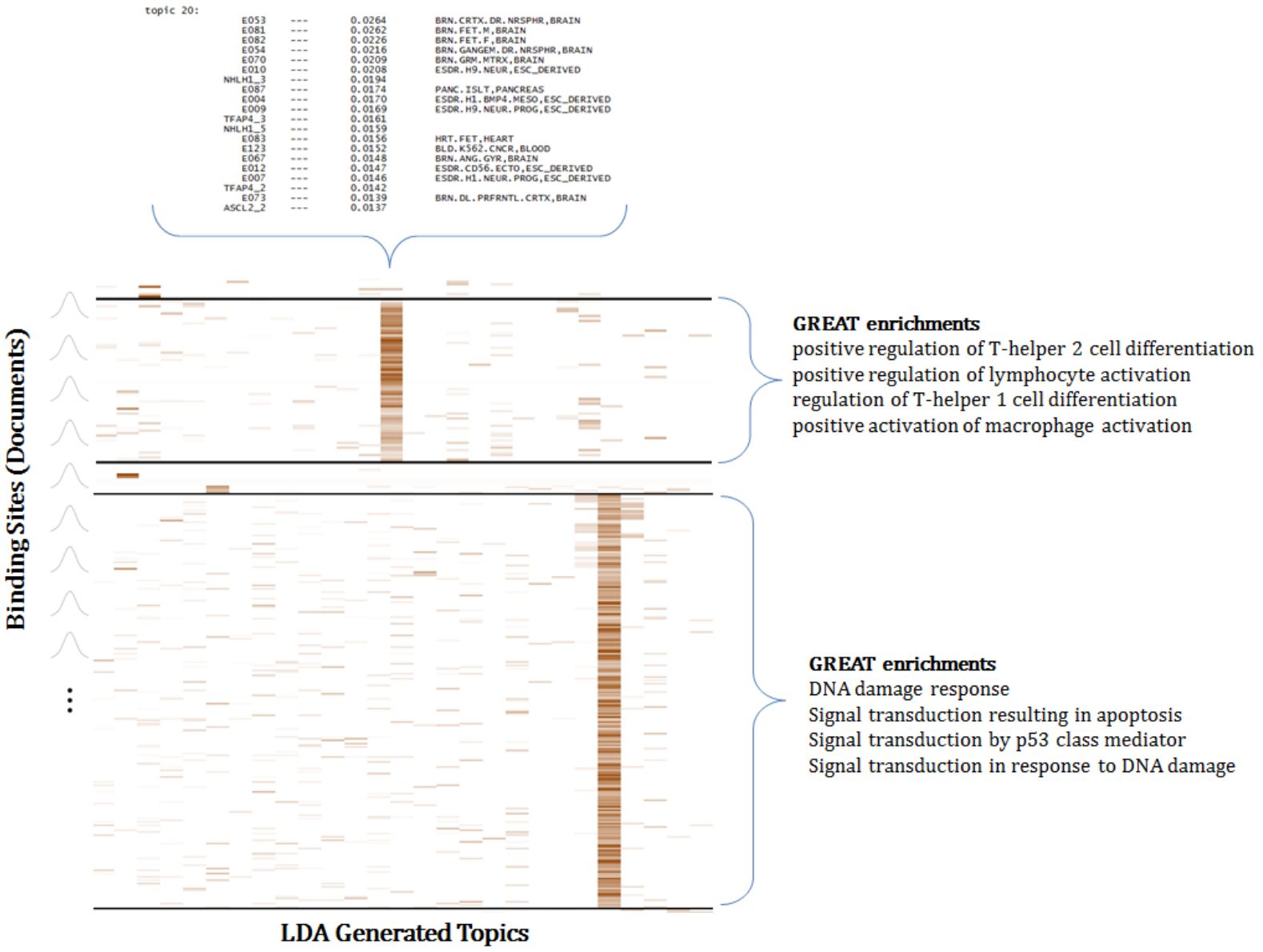

Figure 1: *This is a visualization of the output of running LDA on the combined cell activity/motif matrix for the peaks generated by ChIP-Seq. Each row represents a binding site, and each column represents one of the 50 topics generated by LDA. The darker an intersection is, the larger the assignment of the referenced topic to the referenced binding site. Binding sites are clustered based on similarity of distributions over topics, and topics have been grouped via hierarchical clustering. Sample GREAT outputs are shown for two of the binding site clusters on the right side of the table, and one of topics – a mixture of motifs and cell types – is described at the top of the diagram.*

allow us to split them into functional subgroups representing the different biological functions played by TCF21. TCF21 binding sites serve as the "documents" in our model, and we tested three different "vocabularies" on various runs of LDA: the motifs present at each site, the cell type in which the sites are active, and both simultaneously. In this context, the topics generated by LDA are mixtures of cell types and motifs, which can in turn be related to biological function. A useful byproduct of LDA with this combined vocabulary is the association of certain cell types with certain motifs.

II. Methods

I. Document Clustering

For each document, LDA outputs a distribution over topics. Thus, the documents (in our case, binding sites) may be clustered according to similarity of topic distributions. We implemented k-means with 50 clusters using Jensen-Shannon Divergence as a distance metric. To validate the use of LDA in this context, we ran the algorithm using various settings and showed that the clusters remain relatively stable (see Model Stability). We also processed the clusters using the Genomic Regions Enrichment of Annotations Tool (GREAT) to see if they represent any enrichment (have any biological significance) over the baseline of all peaks [3].

III. Model Stability

We ran Hoffman's online LDA, setting $\kappa = 0$ (no exponential decay of learning rate) and $S = D = 7798$ (using all documents at once) which recovers batch variational bayes LDA [2]. To compare, we changed some parameters to ensure that our results were stable. In all iterations, we used 15 epochs (passes through the entire dataset) for consistency.

I. Random Seed Variation

LDA uses random numbers to initialize γ and λ. In order to test our algorithm for stability – in other words, to make sure that the distributions of vocabulary within each topic stayed fairly constant between runs of LDA – we seeded the random generator with 10 different numbers. We then compared our baseline distribution to the distributions arising from each of the seeds using Jensen-Shannon Divergence as a distance metric, with a confusion matrix mapping each topic in the first set to one in the second set with the lowest divergence. Theoretically, if each run of LDA with a different seed had produced the exact same set of distributions of words over topics, then 100% of distributions from one run of LDA would be matched with a unique highest similarity distribution from the second run of LDA in a one-to-one pairing. On average, 38.4 of the 50 distributions over topics (76.8%) from each run of LDA matched up well with our baseline distributions. This is significantly larger than the average of 22.1 for comparisons of 50 random distributions over 100 iterations, which would have been the expected value if LDA were outputting completely different distributions on each run. This means that that despite some variability across runs, in general there is a core group of fairly constant topics underlying the dataset.

II. Online LDA

We also used $\gamma \in (0.5, 1]$ and batch sizes between 16 and 1024 as suggested by Hoffman et al [2]. As the perplexity graph shows (Figure 2), this performs similarly to batch LDA, which implies that the algorithm is also robust to changes in parameters. The batch size does not affect the perplexity value, but does affect the amount of noise observed. The value of κ, however, does change the final perplexity, and we see that $\kappa = 0.7$ performs better than 0.6 or 0.8. We also see it should be roughly equivalent to use the online algorithm instead of the full-batch one, for example to process massive or streaming peak data without running into memory issues.

III. Convergence

Using Hoffman et al.'s measure of estimated perplexity as a measure of log-likelihood, we can see the convergence of the model graphically (Figure 2). As mentioned above, the perplexity value stabilizes within 15 epochs for all

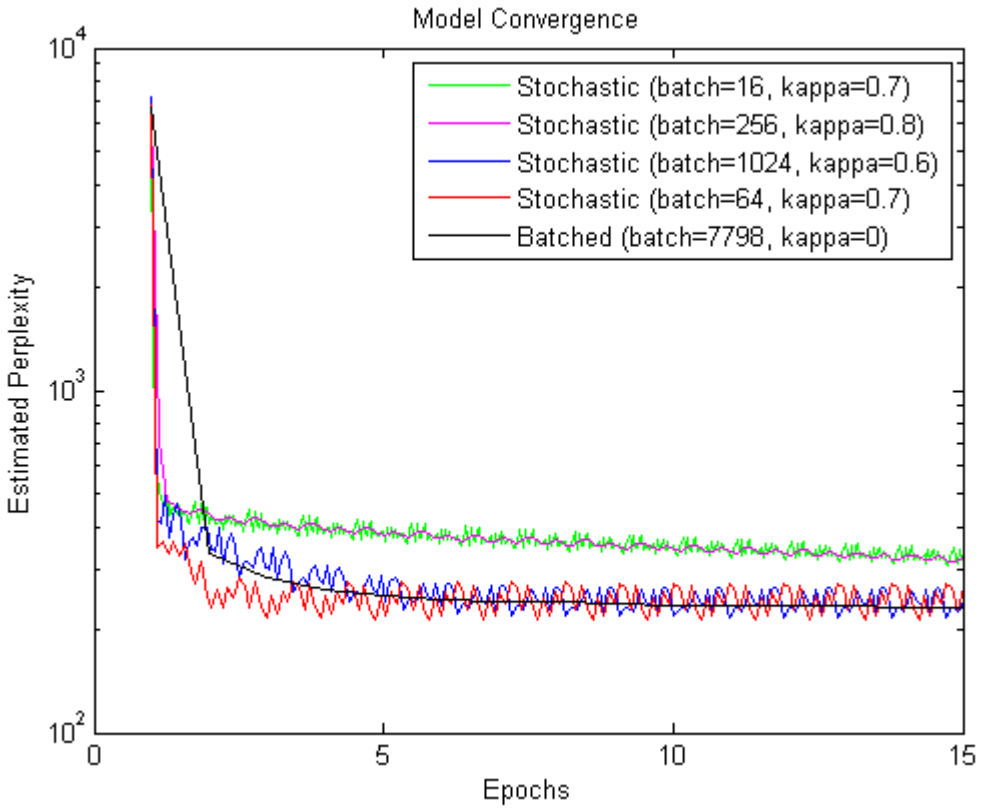

Figure 2: *Looking at both the stochastic (online) and batch versions of LDA, we see that the algorithms converge.*

parameters used.

IV. Biological Validation

I. Cluster Enrichments

In order to determine whether our final clusters of binding sites based on the LDA distributions contained meaningful biological information, GREAT (Genome Regions Enrichment of Annotation Tool) was utilized to test each cluster for biological categorization. 66% of the clusters generated using a combination of cell activity and motif presence as a vocabulary could be related to genes with known biological functions, an improvement over simple k-means clustering (Figure 3).

	Motif Presence	Cell Activity	Motif and Activity
LDA	14%	82%	66%
k-means	12%	90%	32%
Random Assignment	24%		

Figure 3: *We see that the cell activity matrix is the best data set for finding enrichments using GREAT. Adding the motifs seen in the peaks adds noise (and performs worse than a random baseline). However, it should be noted that LDA does much better on the combined matrix, which can be explained by the fact that it weights features differently, namely the cell type activities over the motifs.*

	Motif Presence	Cell Activity	Motif and Activity
Estimated Perplexity	787.4	141.3	232.7

Figure 4: *Here we see that the estimated perplexity for LDA run on the activity matrix only is lower than that run on the combined motif and activity matrix, which is in turn lower than that of the run on the motif matrix only. This corroborates our results above, since the model with the lowest perplexity also performs best in GREAT.*

II. Cell-Motif Associations

DNA binding motifs that often found at sites active in certain cell types are grouped into topics by the algorithm. Thus LDA also yields informative cell-type-motif associations when run on a vocabulary composed of a concatenation of the two.

V. Summary

Useful insight into transcription factor multifunctionality can be garnered from data from just one initial laboratory experiment, using probabilistic methods rather than further laboratory experiments to tease out the details. Figure 1 shows what kind of information can learned from the output matrix.

VI. Acknowledgements

We would like to acknowledge Anshul Kundaje, assistant professor at the Stanford University Department of Genetics and the Department of Computer Science, and Chuan-Sheng Foo, Ph.D student at the Department of Computer Science at Stanford, for their help throughout this project. We would also like to acknowledge Thomas Quertermous and his lab at the Stanford School of Medicine for providing us the fetal heart cell data Finally, we would also like to acknowledge the course staff of CS229, Machine Learning, for their guidance on the project in general.

VII. Data Sources

Region activity levels by cell type: http://www.broadinstitute.org/ anshul/projects/roadmap/segmentations/models/-coreMarks/parallel/set2/final/
Transcription factor binding site motifs: http://compbio.mit.edu/encode-motifs/
Cell type maps: https://sites.google.com/site/epigenomeroadmapawg/project-updates/auxiliarysegmentationcoremarksh3k27ac
Human genome: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/
Fetal heart cell data: Courtesy of the Quertermous lab.

References

[1] D. Blei, A. Ng, M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993-1022, 2003.

[2] M. Hoffman, D. Blei, F. Bach. Online Learning for Latent Dirichlet Allocation In NIPS, 2010.

[3] CY. McLean, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 495-501 (2010).

[4] E.R. Mardis. ChIP-seq: Welcome to the New Frontier. *Nat. Methods* 4, 613-614 (2007).

[5] D. Blei. Mixed Membership Models. *Princeton University* lecture, 10 Oct. 2013.