Deep modeling of gene expression regulation in an Erythropoiesis model

Olgert Denas ODENAS@EMORY.EDU
Department of Mathematics and Computer Science, 400 Dowman Dr. Atlanta, GA 30322, USA

James Taylor JAMES.TAYLOR@EMORY.EDU
Department of Mathematics and Computer Science, 400 Dowman Dr. Atlanta, GA 30022, USA

Department of Biology, 1510 Clifton Road NE, Atlanta, GA 30322

Abstract

The fate of differentiation of G1E cells is determined, among other things, by a handful of transcription factors (TFs) binding the neighborhood of appropriate gene targets. The problem of understanding the dynamics of gene expression regulation is a feature learning problem on high dimensional space determined by the sizes of gene neighborhoods, but that can be projected on a much lower dimensional manifold whose space depends on the number of TFs and the number of ways they interact. To learn this manifold, we train a deep convolutional network on the activity of TF binding on 20Kb gene neighborhoods labeled by binarized levels of target gene expression. After supervised training of the model we achieve 77% accuracy as estimated by 10-fold CV.

We discuss methods for the representation of the model knowledge back into the input space. We use this representation to highlight important patterns and genome locations with biological importance.

1. Introduction

ChIP-Seq is a genome-wide *in vivo* measurement of transcription factor occupancy sites (TFos) (Robertson et al., 2007). This is the primary technique for genome-wide annotation of transcription factor binding. Notably, it has been widely used by the human and mouse ENCODE projects(ENCODE Project Consortium, 2011; Mouse ENCODE Consortium et al., 2012) for the generation of hundreds of assays targeting specific transcription factors (TFs) on a variety of cell lines. ChIP-Seq derived TFos correlate well with the locations of functional genome elements, however the resolution is low and the data lacks statistical power being limited to a single cell-TF pair. It is a challenge today to effectively use the data from this technology for accurate prediction of functional elements. Data noise is bound to the technology, but more context can be used to improve accuracy if prediction models combined data from several experiments.

Here, we propose a deep convolutional architecture as candidate.

Motivated initially by the visual cortex (Hubel & Wiesel, 1965; 1968), deep convolutional architectures(LeCun et al., 1989) have been very successful predictive systems in digit classification, and image and object recognition(Bengio & LeCun, 2007; LeCun et al., 2004) and natural language processing (Collobert & Weston, 2008). A convolutional neural network (CNN) replicates feature detectors across all connections between two layers. Thus, sharing the weights amongst all the connections resulting in lower model complexity and in equi-variant activities (i.e., translated input features result in translated activities). Coupled with pooling, the model is further reduced in size and is invariant to small translations in the input features. Another fundamental property of CNNs is that they are immune to the problem of diminishing gradients and thus even deep architectures can be trained with standard stochastic gradient descent (Rumelhart et al., 1986). Lately, these ideas have been implemented in other systems and applied to bigger

Appeared in the 30th International Conference on Machine Learning workshop on Representation Learning, Atlanta, Georgia, USA, 2013.

images(Le et al., 2011; Lee et al., 2009).

Along with discriminative models, there has been significant recent progress on learning deep generative models, which are concerned with computing new low dimensional representations of the input that will hopefully be linearly separable (Bengio, 2009). Since 2006, when it became possible to efficiently train these models, they have been extensive used both as a representation learning (reviewed in (Bengio, 2009)) paradigm and as a pre-training technique to initialize weights of discriminative models, including CNNs(Claudiu Ciresan et al., 2010). The former strives to learn input transformations that will reveal few important features and will hopefully be linearly separable, the latter is generally considered a good practice as the input conveys much more information to the network than a label in discovering new features. The training stage instead uses labels for fine-tuning the category boundaries. Lately, this intuition has been supported by training large discriminative networks with completely unlabeled data(Le et al., 2011). As a side effect, pre-training helps the network also generalize better.

In this work, we try to bring the above notions together into a model for the activity profiles of TFs and histone modifications during E2-induced G1e differentiation.

In the next section, we briefly describe the G1E biological model and the data. Next, we describe the model and its performance. In the following section, we describe the representation of the features learned from the model and finally we provide a biological interpretation of the features.

2. Experiments

2.1. The G1E biological model and data

GATA1 *null* erythroid cells (G1e) are derived from mouse embryonic stem cells that can be induced to further differentiation into the G1e-ER4 sub-line (Weiss et al., 1997). G1e-ER4 cells resemble normal erythroid progenitor cells, with the exception of an estrogen activated GATA1 receptor (GATA1-ER). This allows the controlling of GATA1-ER expression by treating G1e-ER4 cells with estradiol (E2), which in turn will un pause differentiation in G1e-ER4+E2 cells. Resemblance of G1e-ER4 and G1e-ER4+E2 cells with normal erythroid progenitors and differentiating erythroblasts (Wu et al., 2011) respectively makes the G1e-ER4 differentiation a very good model for normal protheoblast differentiation in mouse. Because Erythroid differentiation depends heavily on the GATA1 transcription factor(Weiss et al., 1994), GATA1-ER release on G1e-ER cells is an important event that causes changes in gene expression and alterations on the TF binding locations and chromatin structure as the cell differentiates into G1-ER4+E2 (Figure 1). Whether the former induce the latter or vice-versa is still an open question, however there is strong belief that GATA1 binding drives changes in histone modification. On this assumption, we consider here, in addition to GATA1, the activity profiles of three other TFs GATA2, TAL1, and CTCF which are important players in the G1e model. The GATA2 TF, a protein similar to GATA1 that recognizes similar motifs (WGATAR) and plays an important role as a regulator of the differentiation process(Yamamoto et al., 1990). The TAL1 protein which is known to form multi protein complexes with both GATA1 and GATA2(Wadman et al., 1997). The CTCF protein, a highly conserved zinc finger protein implicated in diverse regulatory functions, including transcriptional activation/repression, insulation, imprinting, and X chromosome inactivation(Phillips & Corces, 2009).

We process raw data from genome-wide ChIP-Seq *in vivo* detection of TF binding(Cheng et al., 2009) into a continuous signal over a 20Kb window around the TSS of each gene (Figure 1). We further reduce the size of the input by binning the signal into 20bp bins. Each example has 8 gene centered TF activity profiles (4 in G1E and 4 in G1E-ER4+E2. The GATA1 profile, which is absent on the G1E cells, is set to zero) represented as real vectors of size 2000, one for each epigenetic feature. So, a value X_{ijk} represents the activity (i.e. peak enrichment) of protein j, k bins from the TSS of gene i and each sample can be thought as a 2-dimensional matrix with each row corresponding to a TF signal profile.

After standard data preprocessing we were able to extract 406 gene neighborhoods that showed differential feature enrichment in at least one of the tracks and a two fold change in gene expression.

2.2. The model

The model consists of a stack of 3 convolutional layers followed by a fully connected sigmoid layer and a softmax layer for the output labels. The top layer has two outputs representing the state of the gene: *induced* or *repressed*. We search the parameter space for optimal number of kernels per layer, kernel size and layer size for the top two layers, using time and cross-entropy as optimization criteria. The resulting model is in Figure 2.

We train the model using momentum and standard batch gradient descent. Plots of kernel weights show

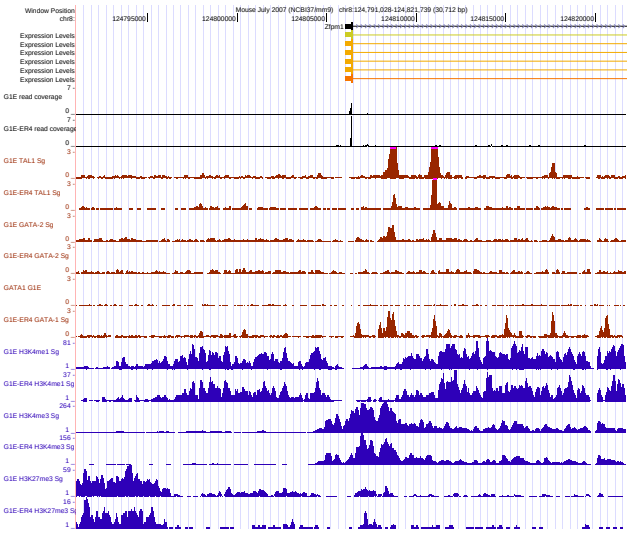

Figure 1. Gene expression levels of the Zfpm1 gene during G1E differentiation and profiles of TF binding and histone modification patterns.

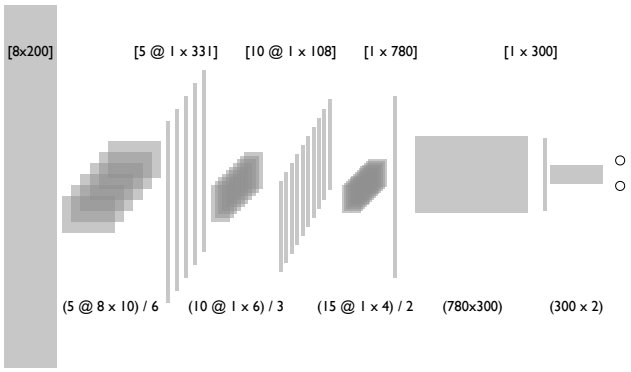

Figure 2. Model architecture kernel annotations are in the format (nr. of kernels @ height × width)/pool size.

Figure 3. Learned filters of the first ConvPool layer of the network.

(Figure 3 for first conv-pool layer) that the model is able to learn features along the tracks as well as combinatorial TF interactions. For example, the first kernel from the left puts much of the weight in the last two tracks and almost completely ignores the first two ones. The bottom rows of the middle kernel (in the first layer kernel rows correspond to input tracks) shows an increasing weight along the track.

3. Model representation

3.1. Representation and model weight representation

The simplest way to represent what the model has learned in the input space is to show the best scoring examples. However, this technique has the problem of running into sampling problems, so below we propose two alternative methods.

One way is to fix the output label at the top layer and run the model backwards: a (hidden) representation is obtained by fixing the previous representation and inverting the function represented by the layer. For a logistic regression layer, let f be the activation function that transforms an input state S into a representation S'. Given S',

$$S = (W^T)^{-1} \times f^{-1}(S')$$

Applying the inverse of the activation function to S and multiplying with the pseudo-inverse of S', give us the most likely the best fit for state S that produced S'. In the convolutional-pooling layer each value of S', $S'_{i,j}$, is one element of the element-wise product of the weights with a region of S, $S_{l,k}$, where $l = [i * p_w, (i + 1) * p_w]$ and $k = [j * p_h, (j + 1) * p_h]$. All values of this are are set to $S'_{i,j}/(|W| * |P|)$, where W is the receptive field and $P = (p_w, p_h)$ is the pooling range.

We can now run the model top-down and obtain a representation of what the network has learned. This representation gives us a way of interpreting epigenetic features that are important in up or down regulation of gene expression.

The input signal obtained above is of course not unique and involves approximation. An alternative method is to initialize the model with learned weights and random input, then run gradient descent with respect to the input maintaining the weights fixed. This method has the problem of getting stuck into local optima. To avoid this we initialize the input to the average signal of the n best scoring examples, for an arbitrary choice of n.

3.2. Biological interpretation

We represent the features learned by the model in the input space by optimizing w.r.t to the input initialized as the average signal of the best 10 examples (Figure 4).

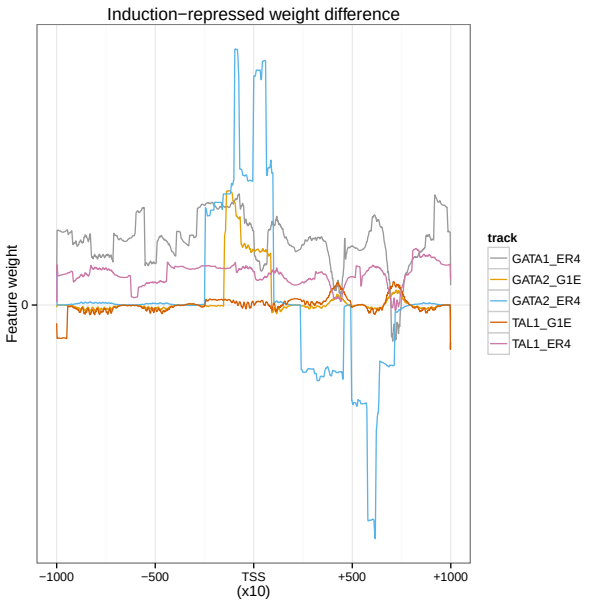

Figure 4. Hypothetical input tracks that maximize model score for each label. For each track we plot the difference between the G1EER4_E2 and G1E signals. Positive signal in this plot corresponds to high enrichment contributing to induced genes and vice versa for negative signal.

The model suggests a strong enrichment of GATA2 in the G1E cells proximal to the TSS. After differentiation GATA1 becomes prevalent 3-7 Kb downstream the TSS. This interplay suggests the ability of the cells to use GATA2 as a surrogate for GATA1 and to restore its function afterwards.

The GATA1_ER4 and the TAL1_ER4 signals show good agreement downstream of the TSS. This indicates that induced genes show a joint enrichment for both GATA1 and TAL1 in the ER4 cell line, as it has already been observed elsewhere(Cheng et al., 2009). Another analogous signal alignment, is between GATA2_G1E and TAL1_G1E. This is consistent with GATA2 acting in the role of GATA1 in G1E cells and recruiting TAL1. After differentiation, GATA1 replaces GATA2 at specific binding locations, but retains TAL1.

Another signal pattern that stands out is the alignment of GATA1 and GATA2 signals in ER4 cells at the proximity of the TSS. This suggests supportive action of both GATA proteins in gene inducement in ER4 cells.

References

Bengio, Y. and LeCun, Y. Scaling learning algorithms towards ai. *Large-Scale Kernel Machines*, 34, 2007.

Bengio, Yoshua. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009. ISSN 1935-8237. doi: 10.1561/2200000006. URL http://dx.doi.org/10.1561/2200000006.

Cheng, Yong, Wu, Weisheng, Kumar, Swathi Ashok, Yu, Duonan, Deng, Wulan, Tripic, Tamara, King, David C, Chen, Kuan-Bei, Zhang, Ying, Drautz, Daniela, Giardine, Belinda, Schuster, Stephan C, Miller, Webb, Chiaromonte, Francesca, Zhang, Yu, Blobel, Gerd A, Weiss, Mitchell J, and Hardison, Ross C. Erythroid gata1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mrna expression. *Genome Res*, 19(12):2172–84, Dec 2009. doi: 10.1101/gr.098921.109.

Claudiu Ciresan, D., Meier, U., Gambardella, L. M., and Schmidhuber, J. Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition. *ArXiv e-prints*, March 2010.

Collobert, R. and Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167. ACM, 2008.

ENCODE Project Consortium. A user's guide to the encyclopedia of dna elements (encode). *PLoS Biol*, 9(4):e1001046, Apr 2011. doi: 10.1371/journal.pbio.1001046.

Hubel, D.H. and Wiesel, T.N. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology; Journal of Neurophysiology*, 1965.

Hubel, D.H. and Wiesel, T.N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Le, Q.V., Monga, R., Devin, M., Corrado, G., Chen, K., Ranzato, M.A., Dean, J., and Ng, A.Y. Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv:1112.6209*, 2011.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, December 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541. URL http://dx.doi.org/10.1162/neco.1989.1.4.541.

LeCun, Y., Huang, F.J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pp. II–97. IEEE, 2004.

Lee, H., Grosse, R., Ranganath, R., and Ng, A.Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609–616. ACM, 2009.

Mouse ENCODE Consortium, Stamatoyannopoulos, John A, Snyder, Michael, Hardison, Ross, Ren, Bing, Gingeras, Thomas, Gilbert, David M, Groudine, Mark, Bender, Michael, Kaul, Rajinder, Canfield, Theresa, Giste, Erica, Johnson, Audra, Zhang, Mia, Balasundaram, Gayathri, Byron, Rachel, Roach, Vaughan, Sabo, Peter J, Sandstrom, Richard, Stehling, A Sandra, Thurman, Robert E, Weissman, Sherman M, Cayting, Philip, Hariharan, Manoj, Lian, Jin, Cheng, Yong, Landt, Stephen G, Ma, Zhihai, Wold, Barbara J, Dekker, Job, Crawford, Gregory E, Keller, Cheryl A, Wu, Weisheng, Morrissey, Christopher, Kumar, Swathi A, Mishra, Tejaswini, Jain, Deepti, Byrska-Bishop, Marta, Blankenberg, Daniel, Lajoie1, Bryan R, Jain, Gaurav, Sanyal, Amartya, Chen, Kaun-Bei, Denas, Olgert, Taylor, James, Blobel, Gerd A, Weiss, Mitchell J, Pimkin, Max, Deng, Wulan, Marinov, Georgi K, Williams, Brian A, Fisher-Aylor, Katherine I, Desalvo, Gilberto, Kiralusha, Anthony, Trout, Diane, Amrhein, Henry, Mortazavi, Ali, Edsall, Lee, McCleary, David, Kuan, Samantha, Shen, Yin, Yue, Feng, Ye, Zhen, Davis, Carrie A, Zaleski, Chris, Jha, Sonali, Xue, Chenghai, Dobin, Alex, Lin, Wei, Fastuca, Meagan, Wang, Huaien, Guigo, Roderic, Djebali, Sarah, Lagarde, Julien, Ryba, Tyrone, Sasaki, Takayo, Malladi, Venkat S, Cline, Melissa S, Kirkup, Vanessa M, Learned, Katrina, Rosenbloom, Kate R, Kent, W James, Feingold, Elise A, Good, Peter J, Pazin, Michael, Lowdon, Rebecca F, and Adams, Leslie B. An encyclopedia of mouse dna elements (mouse encode). *Genome Biol*, 13(8):418, Aug 2012. doi: 10.1186/gb-2012-13-8-418.

Phillips, Jennifer E and Corces, Victor G. Ctcf: master weaver of the genome. *Cell*, 137(7):1194–211, Jun 2009. doi: 10.1016/j.cell.2009.06.001.

Robertson, Gordon, Hirst, Martin, Bainbridge, Matthew, Bilenky, Misha, Zhao, Yongjun, Zeng, Thomas, Euskirchen, Ghia, Bernier, Bridget, Varhol, Richard, Delaney, Allen, Thiessen, Nina, Griffith, Obi L, He, Ann, Marra, Marco, Snyder, Michael, and Jones, Steven. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–7, Aug 2007. doi: 10.1038/nmeth1068.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. Learning representations by back-propagating errors. *Nature*, 323:533 – 536, 1986.

Wadman, I A, Osada, H, Grütz, G G, Agulnick, A D, Westphal, H, Forster, A, and Rabbitts, T H. The lim-only protein lmo2 is a bridging molecule assembling an erythroid, dna-binding complex which includes the tal1, e47, gata-1 and ldb1/nli proteins. *EMBO J*, 16(11):3145–57, Jun 1997. doi: 10.1093/emboj/16.11.3145.

Weiss, M J, Keller, G, and Orkin, S H. Novel insights into erythroid development revealed through in vitro differentiation of gata-1 embryonic stem cells. *Genes Dev*, 8(10):1184–97, May 1994.

Weiss, M J, Yu, C, and Orkin, S H. Erythroid-cell-specific properties of transcription factor gata-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol Cell Biol*, 17(3):1642–51, Mar 1997.

Wu, Weisheng, Cheng, Yong, Keller, Cheryl A, Ernst, Jason, Kumar, Swathi Ashok, Mishra, Tejaswini, Morrissey, Christapher, Dorman, Christine M, Chen, Kuan-Bei, Drautz, Daniela, Giardine, Belinda, Shibata, Yoichiro, Song, Lingyun, Pimkin, Max, Crawford, Gregory E, Furey, Terrence S, Kellis, Manolis, Miller, Webb, Taylor, James, Schuster, Stephan C, Zhang, Yu, Chiaromonte, Francesca, Blobel, Gerd A, Weiss, Mitchell J, and Hardison, Ross C. Dynamics of the epigenetic landscape during erythroid differentiation after gata1 restoration. *Genome Res*, 21(10):1659–71, Oct 2011. doi: 10.1101/gr.125088.111.

Yamamoto, M, Ko, L J, Leonard, M W, Beug, H, Orkin, S H, and Engel, J D. Activity and tissue-specific expression of the transcription factor nf-e1 multigene family. *Genes Dev*, 4(10):1650–62, Oct 1990.