Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks (Review, 2016.10.19)

DeepRegret: Salil Bhate, Bosh Liu, Scott Longwell, Tyler Shimko, Daniel Thirman, Sashwat Udit

Background

There is considerable evidence that noncoding portions of the eukaryotic genome can have significant impact upon expression and phenotype, with corresponding implications for human health and disease. However, the specific mechanisms that govern expression remain poorly understood, and this creates challenges when interpreting individual sequences. There is publicly available information regarding the correlation of certain features with certain phenotypes, but a more powerful model was desired that could interpret and annotate finer and/or rarer variations in sequence motifs. A number of machine learning techniques have been applied to the problem, including support vector machines and random forests. Previous papers have indicated deep convolutional neural networks (CNNs) could produce more accurate results while eliminating the need to engineer features. To maximize use by researchers, the authors felt a GPU trainable, readily deployable solution would be ideal and so created an open-source software package (Basset) to interpret and annotate noncoding sequences of the genome.

Data and Model Summary

The authors merged data on **DNA accessibility** from ENCODE (125 cell types) and the Roadmap Epigenomics Consortium (additional 39 cell types). These data were generated using DNase-seq, which cleaves accessible DNA into smaller fragments that can be detected and mapped through sequencing. Further processing yielded a dataset of 2 million unique DNase hypersensitive sites (DHSs), around each of which a 600 bp window was centered and extracted from the hg19 reference genome.

The authors then encoded each base of these 600 bp sequences as a 4-element binary vectors (one-hot encoding), then supplied the 600x4 sequences as input to a CNN. The first layer of the CNN scans, or convolves, filters across each sequence, with each filter yielding a *feature map* of where the filter is activated. Through training, the weights of these filters effectively represent position-weight matrices (PWMs) for a particular feature or motif. The feature map values are then batch normalized, then rectified through a rectified linear unit (ReLU; negative values set to 0) before being max pooled (subsampled) across strictly adjacent windows to reduce dimensionality. This convolution-pooling process is repeated a total of three times before being fed as input into two layers of fully connected, hidden layers with ReLU nonlinearities and dropout regularization. A final layer maps onto a 164 dimensional vector (one element per cell type), with a sigmoid nonlinearity to scale the values between 0 and 1. During training, these values are compared against the binary ground truth vector via a cross-entropy loss function. The authors used RMSprop to update the weights of the network, and stopped raining after 12 epochs of unimproved validation loss.

Result Summary

The authors used Basset to predict chromatin accessibility. The 300 filters of the network were able to recover a large number of sequence binding motifs, including, for instance, those for Jun and Fos in the AP-1 complex. Subsequently, the authors estimated the importance of each filter using two methods. First, they removed the effect of each filter by keeping the filter's output constant, then used the sum of squares change in predicted accessibility across 164 cell types as the importance of the filter. They also inserted known protein binding motifs into the center of random sequences, and measured the sum of squares change in accessibility as the importance the motif. The authors found that certain features influence the CNN's prediction. For instance, a leading T (5') or trailing A (3') reduced the probability of AP-1 motif according to the model. In addition, ~100nt flanking sequence also influences the predictions. One example is the A-tract, which is known to narrow the minor groove and reduce accessibility.

Interpreting the prediction results of CNNs has been challenging. The authors used *in silico* mutagenesis to score the gain and loss of predicted chromatin accessibility for each variant. They found that potential GWAS causal SNPs are more likely to alter accessibility. For instance, the variant rs4409785 is associated with multiple autoimmune diseases and is predicted to be a causal variant. This SNP is hypothesized to regulate the expression of TYR, which is more than 6 Mb away, via an obscure mechanism. The Basset prediction found that rs4409785 changes the binding affinity of CTCF and potentially disrupts the topologically associating domains.

Training the full multi-task model on all 164 cell types was computationally expensive, requiring 85 hrs of GPU time. To demonstrate that a pre-trained model could be rapidly adapted to new, cell types, the authors pre-trained the model on 149 of the cell types, then further trained the model on each of the remaining 15 cell types (separately). Adapting the pre-trained model to a new cell type required only a single pass through the training set to yield high-accuracy. Since adapting the pre-trained model required only 18 min of GPU time (or 6.5 hours of CPU time), this strategy makes Basset less cumbersome and more accessible to researchers.

Weaknesses

Though the authors aggregated a large dataset of 2 million examples, they had a curious train:val:test split of roughly 93:3:3 (60:10:30 is common). Canonically, small validation and test sets put a model at risk of overfitting, leading to high variance and an inability to generalize. Despite this, the authors mention in Figure 6 that their "full model may benefit from increased capacity or decreased regularization", based on the observation that, for a given cell type, the AUC of the seeded single-task model was often better than the AUC of the full multi-task model. The simplest explanation for this improvement is that when the powerful full model, which was optimized for the complex multi-task problem, was adapted to a simple single-task, it was unsurprisingly able to fit the new data very quickly. Indeed, the authors note that the pre-trained model would overfit with more than one training epoch (notably, it is not clear from the methods if a train-test split was performed on the datasets for each of the 15 withheld cell types - without a dataset split and with a powerful model, a high AUC is to be expected).

Confounding the author's use of AUC as a benchmark is the skewness of the underlying data: within each cell type, 4-19% (median 8%) of the 2 million DHSs were accessible, so there were lots of negative examples. In these situations, precision-recall (PR) curves are often more informative than ROC curves. The authors admit that their PR curves (in the supplemental) do

not inspire confidence for *de novo* genome annotation, but they nonetheless state repeatedly that Basset and CNNs “easily handle imbalanced datasets”. Although authors demonstrate that their model is state-of-the-art, it would be interesting to have trained the model with regard to a metric that acknowledges the skewness of the data. A thorough discussion of the skewness with respect to cell-types (other than the range and median) would also have been helpful.

In their section on interpretation of the filters, the authors do not give consideration to the correlations between individual filters. While there are hundreds of filters, some of which produce useful insights, many are likely correlated. The authors present this data as a tool for mining sites, but there is no indication as to the true amount of information present.

In their in silico saturation mutagenesis, they take extremal values to compute a score. This seems inappropriate - their AUC shows that the extreme values are precisely where the algorithm is inaccurate. Therefore, it requires further justification to interpret these values, versus perhaps the median or some other quantile. The relationship with PhyloP is noisy at best and it is visible that there is a mostly uncorrelated distribution with the correlation generated by a bimodal PhyloP score.

In their SNP analysis, aside from for the nonparametric test of the group means, the authors present no statistical tests. Even in their test comparing high PICS and low PICS groups, the effect size seems to be small. Moreover, their threshold for 'high' and 'low' groups appears to be arbitrary. Including one particular example of biological interest is unjustified without any indication of statistical significance or multiple testing correction. In general, whether or not DNAse hypersensitivity confers interpretation for noncoding GWAS hits is a question altogether beyond the scope of the work. In the absence of robust statistical testing, the authors convolute their prediction accuracy with interpretation of population scale studies.

It isn't clear how the dynamics of the model's learned filters relate combinatorially to the 164 cell types, especially considering that they have different genetic backgrounds. Is it the case that the model is learning whether a specific sequence is accessible in all cell types? They include ES cells, which have typically very accessible chromatin. A biased model could succeed in this case. The authors need to describe in detail the performance in each cell type. Without an understanding of the relationship between different cell types' results, the relevance of the model is questionable since DNA accessibility is not a static feature of the genome, but something regulated dynamically in each cell type by processes not necessarily encoded in DNA.

In general, while the authors develop a model capable of predicting DNAse hypersensitivity, their applications are statistically unsound and somewhat contrived. They do not provide a consideration of how their methodology might bias their results.

Extensions

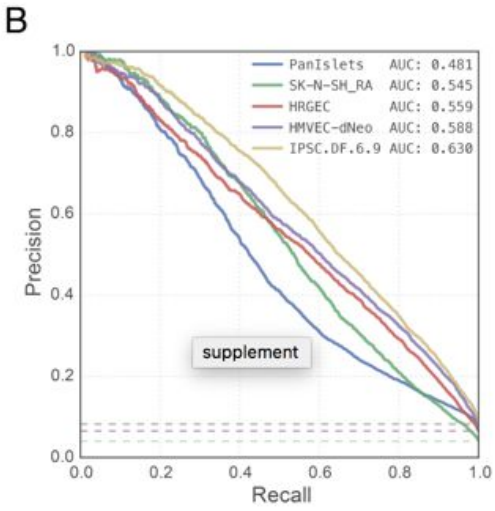
The Basset package retrieves interpretable position weight matrices (PWMs) for many known and unknown DNA binding proteins. In this paper, the authors make efforts to identify these interacting proteins by comparing the learned PWMs to a PWM database. The authors are able to uncover motifs and sequence features associated with open or closed chromatin regions. However, the authors did not extend this further to examine the spatial relationships between the different motifs. Using a technique such as that implemented in the DeepLIFT framework[1], the authors may be able to glean more insight into how the binding motifs of different factors

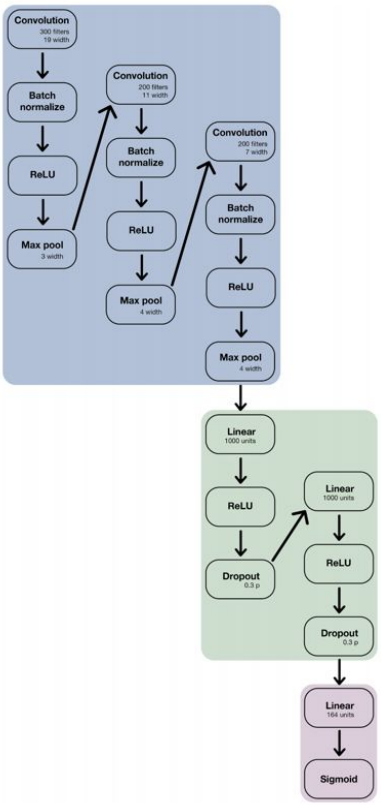
must be arranged in order to predict chromatin state. Similarly, the authors could generate maximally activating inputs, as performed by Simonyan, *et al.*[2], in an effort to better understand these spatial relationships.

Additionally the authors indicate that their model is capable of predicting the effects of mutagenesis on chromatin state for any genomic sequence. This capability represents a marked advance in *in silico* mutagenesis techniques and has the potential to be applied to a wide range of problems. Combined with CRISPR/Cas9-based genome editing technologies, it is possible to test the accuracy of these predictions in immortal cell lines.

References

1. Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016, May 5). Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. arXiv.org.
2. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013, December 20). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv.org.





dataset split:

total	**2,071,886**	**100%**
train	1,930,000	93.2%
val	70,000	3.4%
test	71,886	3.5%