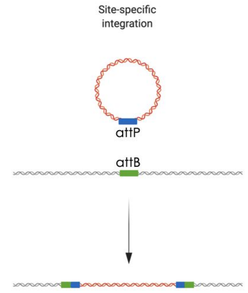
Predicting DNA recombination attachment sites using deep learning

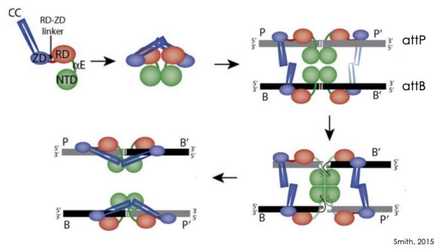
Matthew Durrant, mdurrant@stanford.edu
Josh Wolff, jw1@stanford.edu
Vinay Sriram, vsrirarm@stanford.edu

Background

Integrative mobile genetic elements are ubiquitous in nature. Certain mobile elements use a recombinase to integrate the element (orange circle, right) into the genome (grey line, right) by binding to the attP (blue rectangle, right) and the attB (green rectangle, right) attachment sites to recombine the two strands. In this study, we wanted to use deep learning to understand how recombinases recognize and bind to their target attachment sites.



Large serine recombinases (LSRs) are ~500 amino acids in length, and contain three distinct domains (below) – A catalytic N-terminal resolvase domain (green), a DNA-binding recombinase domain (red), and a zinc-ribbon domain (blue) that likely regulates the directionality of the recombination reaction. These recombinases are useful tools for genome editing, and understanding the mechanism by which they recognize and recombine two strands of DNA will help us to further develop them as tools.



Smith, 2015

Predicting

The inputs of our algorithms are as follow.

- DNA sequences varying from 101-150 nucleotides in length (4 nucleotide characters {A, T, G, C} encoded as a one-hot vector).
- Amino acid sequences varying from 200-600 amino acids in length (20 amino acid characters encoded as a one-hot-vector).

We use neural networks for two different classification tasks. The output for our first classification task (task I) is a label that predicts the type of attachment site for an input DNA sequence (ternary output). The output for the second classification task (task II) is a binary label indicating whether an input DNA sequence can bind an input amino acid sequence (binary output).

Data and Features

Matt Durrant has recently developed a computational pipeline to mine public databases containing >100k bacterial genomes to identify recombinases and their predicted attachment sites (Durrant et al., 2019). This pipeline identified ~9,000 recombinases, along with predicted attachment sites. Negative examples were synthesized using various techniques, and we found that removing low-confidence predictions dramatically improved performance.

Below is an example of the data:

attB TCCCTGTAATCATCCCAGGTCAGGACTACCTGGTTGTGTAAGTACATTGGCACCAGTAAATTAGCAGAAGCATAAGTGGCAGGACGCATAGCCCACACTCAAGAGGA

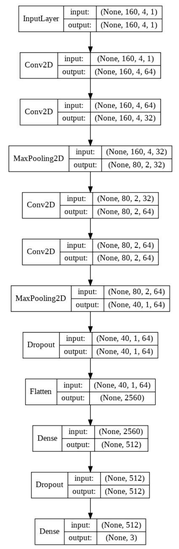
attP GGTAGGCGGCGGTGTCGTCGTCCTCGAAGGACTCGTGTGCCATGACGTGGCACCAGTGGCAGCTGGCGTACCCGACGCTGAGCAGCACGGGCACATCCCGCTTCCTG

Recombinase MYTARSQEGASSTTEDETGPSDLDLRGRPAGLATPDELARLHPDAVFLIAYSRISLDWRKHSRKKAATSSWSAGKGVANQHRRNDKNAARHG AIIVHRYTDNDLSASKRDVVRPDFQAMLRDLRRGHTPEGYPVHGAICVDQDRVQRTDRDWEDFVDALTLDPTRRFYTPSGPMDLTEESEIIK TGVMAVVNKAESLKKKRRIRDWHQDRILDGLPHSGPRPFGWEEDRENLRPAEATYLAWAMDERIKGRAMKTLCLEAKRRGLTGTRGGEIVPQ TLTQWMTAPRVCGYRANRGDLVLDDHGEPIVGIWKTICSPEKWIAVCATFGDGSTYLARGSGTPRITGIPKTIKYMGSTLLRCQYRFGEDTE DTSLIGRVCNEPMGGSKASSKKSPYVYTCAACSRNAISGPMVDWQIQGLLLAKLDQAQATFIPPDLAWPKEDELRQRADKLARLEKEKWEADT ISSEMPYRLAPKIEKEVKVLRRERAQFELVSQAERERAPGDVARKWLAGEYDLAQKRKVLFEAFAAIQVRPGRKGNKTPDPKRLTPVWHQ

Task I - Classifying attachment sites

The first task that we addressed was predicting if a given DNA sequence was an attP site, an attB site, or a negative, random sequence. Our dataset was 2/3rd negative sequences, 1/6th attP, and 1/6th attB sites. We used a 70-10-20 train-dev-test split. The raw dataset contained 177,426 examples, and the cleaned dataset contained 90,078 examples.

Model Design



Total Parameters: **1.3 Million**

The final layer has a softmax activation with three classes. The softmax function outputs a vector of probabilities *p*.

$$p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

Each probability is computed using the following formula.

$$p_i = \frac{e^{x_i}}{\sum_{j=1}^{3} e^{x_j}}$$

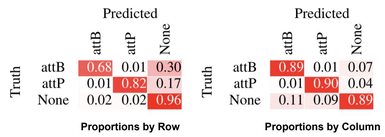
Finally, we use the following generalization of logistic loss.

$$J = -\frac{1}{N}\sum_{i}^{N} \mathbf{y}_i \log(\hat{y}_i)$$

Results

Architecture	Dataset	Training Acc.	Dev Acc.
Majority Class Guessing	Raw	66.7%	66.7%
Logistic Regression	Raw	68.0%	68.3%
Dense Network	Raw	98.0%	70.1%
3x3 filter CNN	Raw	98.6%	81.2%
4x4 filter CNN	Raw	94.6%	80.9%
Majority Class Guessing	Cleaned	66.7%	66.7%
Logistic Regression	Cleaned	70.4%	70.4%
Dense Network	Cleaned	99.2%	83.4%
3x3 filter CNN	**Cleaned**	**94.0%**	**89.8%**
4x4 CNN	Cleaned	93.8%	88.6%

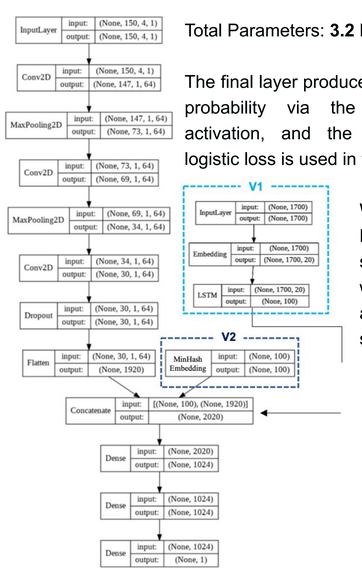
The final selected model had a test accuracy of **89.2%**, but this single metric doesn't fully explain its performance. The confusion matrix (below) reveals that the predictions are quite precise, but the sensitivity is lower for the attX sites, as attB sites are often confused with negatives (non-sites).



Task II - Predicting recombinases that bind attachment sites

The second task was to predict which protein amino acid sequence binds to which DNA sequence. The input was an (attachment site, protein) pair and the output was whether or not the protein targeted the attachment site (1 or 0). The dataset was 2/3rd negatives (no binding), and 1/3rd positives (binding). We used a 70-10-20 train-dev-test split. The raw dataset contained 231,171 examples, and the cleaned dataset contained 107,442 examples.

Model Design



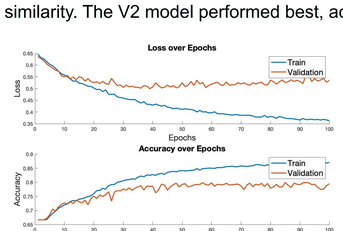
Total Parameters: **3.2 Million**

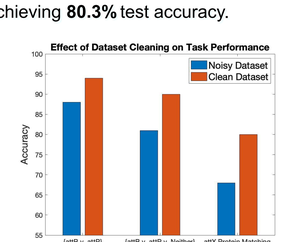
The final layer produces a class probability via the sigmoid activation, and the standard logistic loss is used in training.

Results

Architecture	Dataset	Training Acc.	Dev Acc.
Majority Class Guessing	Raw	66.7%	66.7%
LSTM-CNN	Raw	66.7%	66.7%
Minhash-CNN	Raw	70.4%	68.0%
Majority Class Guessing	Cleaned	66.7%	66.7%
LSTM-CNN	Cleaned	71.8%	70.1%
Minhash-CNN	**Cleaned**	**90.1%**	**80.1%**

We considered two possible encoding schemes for the protein sequence: an LSTM based many-to-one character level encoding (V1) and a min-hash signature based on the k-mers in the protein (V2). The minhash embeddings were 100-dimensional (i.e. we used 100 random permutations over the set of all possible 4-mers to produce the embeddings), and preserve Jaccard similarity. The V2 model performed best, achieving **80.3%** test accuracy.

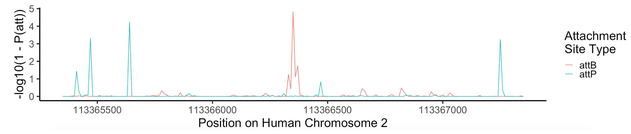




Application

We wanted to try applying the classifier that we trained to an entirely distinct dataset, the human genome. Understanding how often predicted attachment sites exist in the human genome and where these sites are located can help us to develop a recombinase that works effectively on the human genome.

We scanned 1 mbp regions in the human genome selected at random (see below), with a sequence length of 106 base pairs (bp) and a stride of 10 bp. We analyzed 31,795,188 sequences is total. Using a probability cutoff of 0.99999, we identified 22,502 sequences that were predicted attP sites, and 77 sequences that were predicted attB sites. This large disparity between the frequency of the two sites is intriguing and requires further investigation to understand, but it may be due in part to our model's lower sensitivity to attB site detection (see confusion matrix to the left).



An example of our model's predictions when scanning across the human genome.

Discussion

On the whole, we are pleased with the outcome of this project. Most deep learning models that predict whether or not a protein binds to a DNA sequence are trained on data generated from controlled experiments on human cells. Our dataset was generated by analyzing thousands of diverse bacterial genomes, and we were clearly able to identify meaningful signals in both tasks. This is an exciting finding and an encouraging start.

In consulting with an expert in deep learning for genomics, Anshul Kundaje, we were informed that our second task would be considerably more difficult than the first. We found that the second task was indeed more difficult than the first, but we were still able to demonstrate a significant improvement over random with our best model. Another interesting finding of this study was the importance of cleaning the initial dataset - we substantially increased performance despite cutting the size of the dataset in half.

Future Directions

- The model that we have developed for the first classification task will be useful for Matt Durrant's PhD Thesis. It will be used to screen predicted attachment sites going forward, which can then be experimentally validated.

- The second task, predicting if protein binds to a given attachment site using primary sequence alone, is more difficult. Future models could draw upon biophysical homology models of the 3-dimensional structure of the protein in an attempt to improve binding prediction.

References

Durrant, M. Li, M. Siranosian, B., Bhatt, A. S. (in press), A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host & Microbe*, preprint on bioRxiv - doi: https://doi.org/10.1101/527788

Smith, Margaret C. M. 2015. "Phage-Encoded Serine Integrases and Other Large Serine Recombinases." *Microbiology Spectrum* 3 (4). https://doi.org/10.1128/microbiolspec.MDNA3-0059-2014.