Predicting DNA recombination attachment sites using deep learning

Matthew Durrant, mdurrant@stanford.edu
Josh Wolf, jw1@stanford.edu
Vishal Sriram, vsriram@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 their attachment sites typically contain >100k bacterial genomes to identify recombinases and their predicted attachment sites.

Predicting DNA recombination 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 22,502 sequences that were predicted attP sites, and 77 sequences that were not predicted to be recombinase binding sites.

The second task, predicting if protein binds to a given attachment site using biophysical homology models of the 3-dimensional structure of the protein in an attempt to improve binding prediction.

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 22,502 negative sequences, 1,658 attP, and 1,616 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

The inputs of our algorithms are as follows:
- 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 used 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 (tertiary 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).

Results

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) and the output was whether or not the protein targeted the attachment site (1 or 0). The dataset was 22,502 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

The final layer produces a class activation with three classes. The final layer has a softmax activation, and the standard probability via the sigmoid function.

Results

The model that we have developed for the first classification task will be used to facilitate our second task. 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 guessing of 44%. 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 guessing with our best model.

Future Directions

- The second task, predicting if protein binds to a given attachment site using biophysical homology models of the 3-dimensional structure of the protein in an attempt to improve binding prediction.

References


Smith, Margaret C. M. 2015. “Phage-Encoded Serine Integrases and Other Large Serine Recombinases.” Microbiology Spectrum 3 (4).