Predicting Splice Sites from a DNA Sequence

Group: Down
Hannah Brooks, Meaghan Grogan, Jeff Wang, Jakob Weiss

The Central Dogma

The central dogma in Biology is that all genetic material is found in the nucleus in the form of double stranded DNA, deoxyribonucleic acid. Within the nucleus, that DNA is used as a template to transcribe, or synthesize, pre-mRNA, or pre-messenger ribonucleic acid. The protein that carries out **transcription**, the process of synthesizing pre-mRNA from the DNA sequence, is **RNA Polymerase II**. This sequence of pre-mRNA is a direct copy of a specific part of the DNA, including both gene-encoding sequences, **exons**, and sequences that are unnecessary for the final gene product, **introns**. This part of processing of the pre-mRNA is called **splicing**. Once the introns are spliced, or taken out, of the pre-mRNA sequence and other processing is done to protect the ends of the single-stranded mRNA sequence, the mRNA can be transported from the nucleus to the cytoplasm, where it can be transcribed to a protein sequence of amino acids in the endoplasmic reticulum (Figure 1).

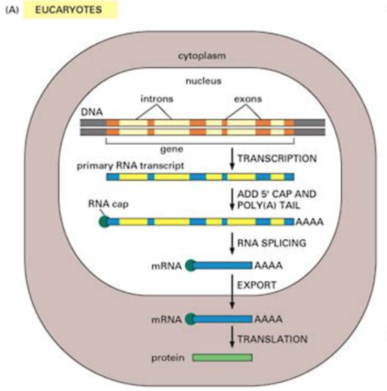

Figure 1: The Central Dogma

The Process of Transcription

For transcription to begin in the nucleus, RNA Polymerase II must bind to that portion of the DNA. RNA Polymerase does this using the DNA binding site found in its amino acid sequence. Each gene, a portion of the DNA, contains a sequence known as the **promoter region**, which contains sequence motifs that RNA Polymerase recognizes. The promoter region contains a **TATA-box**, which is a sequence containing primarily A- and T- nucleotides. A region before the promoter region is referred to as a regulatory region. **Transcription factors**, proteins that bind to the regulatory region, have DNA binding sites. Transcription factors recruit the RNA Polymerase II to increase the chances of transcription at that position.

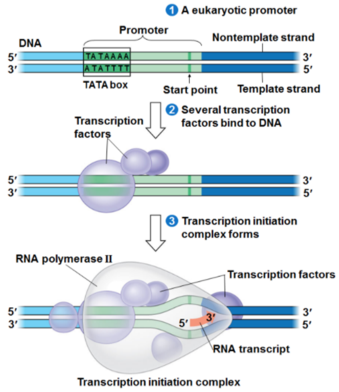

Figure 2: Transcription

Gene Expression and Chromatin Remodeling

Transcription from DNA to RNA and the difference of the type of genes transcribed in each cell is what makes a red blood cell different from a skin cell. This idea is fundamental to biology because all cells contain the same DNA, but different levels and types of mRNA, differentiating them from one another and giving each a different function. This is referred to as **differential gene expression**. A large factor in determining what genes will be transcribed, or expressed, in what cell is based on **chromatin remodeling**. For all of our DNA to fit within the nucleus of a cell, it must be packaged into chromatin. Chromatin packing involves wrapping the DNA strands around

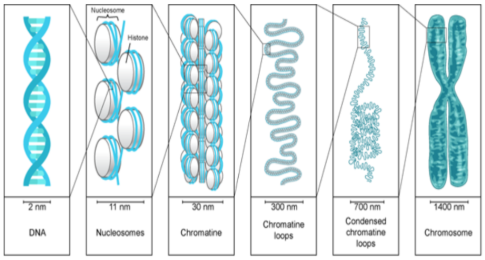

Figure 3: Packaging of DNA into Chromatin

histones, proteins that bind to DNA (Fig. 3). The histone proteins can undergo modifications to their amino acids. These **histone modifications** include **methylation**, **phosphorylation** and **acetylation**. Both phosphorylation and acetylation can decrease the packaging of the chromatin, making the DNA more accessible to machinery such as transcription factors and RNA Polymerase. Phosphorylation of histones is accomplished by histone phosphatase and acetylation is accomplished by histone acetylase. When chromatin packaging is loosened, the chromatin is referred to as **euchromatin**. Histone methylation, done by histone methylase, increases the packaging of the chromatin, making it so that RNA Polymerase and Transcription factors cannot bind to that region of the DNA. Packaging the chromatin more tightly creates **heterochromatin**. Methylation, acetylation, and phosphorylation can all be reversed as well by histone demethylase, deacetylase, and de-phosphatase, respectively. Histone modifications and their ability to take place and be reversed leads to even more complex differential gene expression at different times of development and in response to different environmental factors.

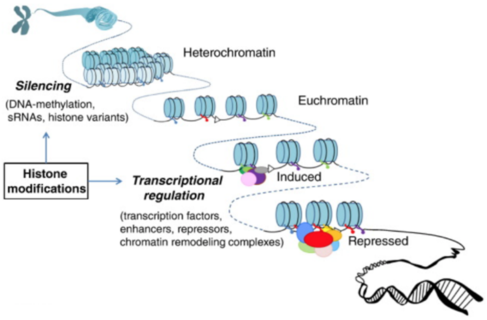
Heterochromatin
Silencing
(DNA-methylation, sRNAs, histone variants)
Euchromatin
Histone modifications
Transcriptional regulation
Induced
(transcription factors, enhancers, repressors, chromatin remodeling complexes)
Repressed

Figure 4: Histone Modifications and DNA Accessibility

DNA Methylation

Another way in which transcription is regulated is through **DNA methylation**. Parts of the promoter region of DNA, aside from the TATA-box, are rich with cytosines. Cytosines, the C-nucleotide in DNA sequences, can be methylated by DNA methylase. When cytosine is methylated, it prevents the RNA polymerase from binding to the region. DNA methylation, therefore, silences the gene and prevents its expression (Figure 5).

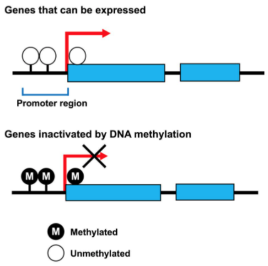
Genes that can be expressed
Promoter region
Genes inactivated by DNA methylation
M
Methylated
Unmethylated

Figure 5: DNA Methylation

Splicing and the Spliceosome

Splicing is a result of removing introns from the pre-mRNA so that the gene product that will be translated into protein will only contain the gene-coding, exon, parts of the sequences. The machinery that carries out splicing is the **spliceosome** (Figure 6). The spliceosome, like RNA Polymerase, is a protein that contains a binding site that recognizes specific motifs in the mRNA, conserved regions that typically flank the intron sequences, or **splice sites** (Figure 7). The spliceosome binds to these specific sequences to determine what should be spliced, or cut, out and what should remain. The spliceosome also has a protein that can cut mRNA, removing that piece from the mRNA so that the final product is accomplished.

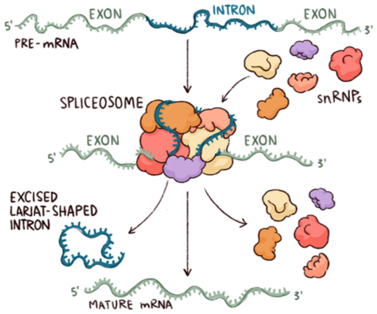

Figure 6: Spliceosome and Splicing

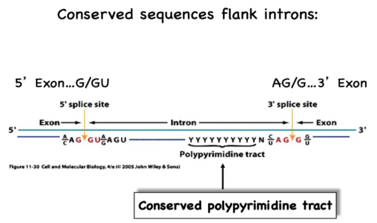

Figure 11-30 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

Figure 7: Splicing Motif Sequences

Janggu

Dataset preparation:

We had to create roi_train.bed and roi_test.bed. The train file contains 9,383 genes, each listed by their chromosome and their start position. The test file contains 3,641 genes, each listed by their chromosome and their start position. Because of the sizes of the genomes and therefore the files needed, we used the GPU server as opposed to JupyterHub as JupyterHub did not offer enough space for this project to work, especially if we were able to train a more complex neural network.

We had to create a sites_all.bed file that contained the site of every exon that we retrieved from the dataset found in *Predicting Splicing from Primary Sequence with Deep Learning* (Jaganthan et. al), its corresponding chromosome, and whether it is a start or end site of an exon, labelled 1 or 2, respectively. We used the training and testing datasets broken up by chromosome number.

We used the same reference genome from the Jaganthan et. al paper because we were using their training and test dataset and had to ensure that the positions lined up correctly. The reference genome was used to retrieve the entire DNA sequence of the genes listed in the roi_train.bed and roi_test.bed using their positions, or coordinates.

Dataset that can be interpreted by the model:

In order to get the data shaped in a way that it could be read by the models, we first used the Bioseq dataset from Janggu, which can be used to load nucleotide or protein sequences along with defining regions of interest. Additionally we used the dataset wrappers ReduceDim, which takes 4D data into a 2D table-like object, and SqueezeDim, which does the same thing. Using both of these dataset wrappers allowed us to make the DNA data into the correct size of (387,64). We conducted a similar process with the Labels data, using ReduceDim and SqueezeDim to size the data to (4281,). The DNA and Labels set were then able to be fed into the logistic regression model.

The model:

First model: Using Logistic Regression and Random Forest to predict for every position the probability that that position is a splice junction. Predicts whether that nucleotide at that position, while accounting for the flanking region of 50 nucleotides, is a splice junction or not.

Second model: We adapted a neural network found on the Janggu documentation which they used to predict DNA regulatory sequences. We used a conv1d layer to extract features from sequential DNA data and apply a MLP layer to predict the output. We ran the neural network twice for 100 epochs each, first with a flanking region of 50 nucleotides and second with a flanking region of 150 nucleotides.

Janggu Documentation:
https://janggu.readthedocs.io/en/latest/tutorial.html#part-i-introduction-to-genomic-datasets

Janggu Neural Network Example:

https://nbviewer.org/github/BIMSBbioinfo/janggu/blob/master/src/examples/pytorch_convnet_example.ipynb

Project Environment Necessary for Janggu:
final-proj-env.yml file in the .zip folder

Deep learning for genomics using Janggu:
https://www.nature.com/articles/s41467-020-17155-y

Predicting Splicing from Primary Sequence with Deep Learning:
https://www.sciencedirect.com/science/article/pii/S0092867418316295#sec4.5

Link for Reference Genome:
https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz

Results:

When we trained the initial models on only one gene, our results for the logistic regression and the random forest models were really good, about 90% accuracy for each. However when we increased the number of genes we trained these models on, the accuracy for these models decreased to about 50%, which can be expected with this many genes. The neural network worked better, with 72% accuracy with the most basic neural network we tried, using 50 flanking nucleotides and 100 epochs. We then ran the neural network again with 150 flanking nucleotides to increase the information that the network has available to train on. However the accuracy was slightly lower at 69%, which indicates that our current neural network is too shallow to accurately predict the splice sites with this amount of data. Additionally, the training was only done with the splice sites from one gene. Given more time and additional resources, we would have trained a deeper neural network on more genes, however due to memory constraints, we were only able to train one model at a time on one computer. Therefore the current neural network is working fairly well for being trained on one gene for a best accuracy of 72%.