SpliceSlice Analysis Pipeline

Splicing is an essential biological function that generates diversity among mRNA in the transcriptome. Splicing is performed by a protein complex called the spliceosome, which is facilitated by specific subunits binding to intronic sequences known as **Branch Points (BP)** and **Polypyrimidine tracts (PPT)**. The sequences that make up these genomic features have been described as “loosely conserved”, but differences in their sequences could logically enhance or diminish their binding affinity for spliceosome subunits. These distinct spliceosome binding affinities could be differentially leveraged by the cell to encourage splicing and the generation of different mRNA isoforms.

The idea with SpliceSlice is to identify branch points and polypyrimidine tracts within sets of transcripts and determine if there are statistically significant differences in their sequence composition. These sets of transcripts can be differentially expressed between certain treatments or represent all sets of transcripts in genes with different functions. It is potentially worth exploring how these sequences differ among transcripts or genes that are expressed in different tissues, in different quantities, or in different organisms entirely, as identifying these differences could help clarify more of their functional role in splicing and their evolution among Eukarya. The goal of this pipeline is exploration and its application is intended to be more general.

**Workflow:**

1. **Extracting Intronic Sequences:** Once sets of transcripts are identified, subsequences of their introns are extracted from the reference genome using the corresponding gene annotation according to a modified version of the method described by Zhang, et. al. (https://academic.oup.com/bioinformatics/article/33/20/3166/3870482). The sequences extracted correspond to:

**1.** the polypyrimidine tracts (3-16 nt upstream from the 3’ splice site),

**2.** a background sequence region (187-200 nt upstream from the 3’ splice site), and

**3.** a 200 nt window ending 10 nt before the 3’ splice site which corresponds to the potential location of the branch point.

1. **Calculating Octanucleotide Frequencies:** While polypyrimidine tracts are relatively easy to identify, branch points remain more elusive. We utilize a tool called BP\_PPT.py that identifies BPs through weighted octanucleotide frequencies of experimentally validated BPs and PPTs (weight is determined by binding affinity for the spliceosome). Since the BP base pair frequencies cannot be directly observed prior to prediction, the octanucleotide frequencies of BPs estimated for humans by BP\_PPT is used while the frequences of the PPTs and background sequences are updated by our observations. This reflects a method used by Kadri et. al. for extending this method to bovine (<https://www.nature.com/articles/s42003-021-02725-7>).
2. **Predicting Branch Point Sequences:** The weighted octanucleotide frequencies are input to BP\_PPT.py along with the potential BP intronic slices from the two groups, resulting in predicted BPs represented by a heptanucleotide sequence. These sequences do not have an associated p-value, but there is a score calculated by the normalized sum of the weights. Naturally, the score with the highest value is reported (maybe this can be improved?).
3. **Permutation Testing:** Once predicted sequences for the BPs and PPTs have been identified, the sequences are duplicated to reflect their quantities in the case these transcripts have different levels of expression (this method needs more work, ie normalization, factoring in replicates, all that good stuff). If expression levels are not included, duplication is skipped and each BP and PPT sequence is unique in the dataset for their respective group. These groups are then subjected to a permutation test where the datasets are combined and randomly shuffled to create groups comprised of sequences from both groups. These groups are then used to create position weight matrices and the Kullbeck-Leibler (KL) divergence is then calculated between the two distributions, yielding our test statistic. This simulation is ran 1000 times to create the distribution for the null hypothesis, which is then compared to the actual KL value from the position weight matrices of the two groups. The fraction of iterations that our KL value is less than or equal to the simulated KL value is our p-value.