Scott Teresi
Modeling/Machine Learning Research Plan
Wednesday 25th July, 2018

Introduction:

Tranpososable elements (known as TEs or tranposons) are genomic elements that previously used to be only considered "parasitic" elements of a genome; they were thought to be parasitic in the sense that they proliferate throughout a genome (through a variety of cut-and-paste or copy-and-paste mechanisms) but serve no purpose. However, a growing body of research demonstrates that TEs can modulate gene expression and genome evolution. To my knowledge, this can happen through a variety of mechanisms owing to the typical silencing and locking down of TEs to prevent them from proliferating and thus costing the organism energy; additionally a TE may also paste itself into an existing gene and disrupt the function of that gene by breaking the progression of sequences.

The idea is that when transposons are methylated or silenced through other means, this methylation can also affect nearby genes; I like to think of this as collateral damage. Previous research has observed the general trend of transposon presence and/or density being negatively correlated with gene expression, however this trend has not been explored with respect to certain transposon classifications (hereafter types and the sub-groupings of families) or the distances at which a transposon may can occur in relation to a gene.

To give a short example of what I mean, no one knows if transposon X can actually affect gene expression if it is say more than 500 base-pairs (bp) away from the gene. Is it useless to worry about transposon X if it is of family Y if it is such and such distance away? Do we observe the same trends for each transposon type and family or do these trends differ? Similarly, could gene expression only be affected if transposon density crosses a certain threshold, and again, does this characteristic differ between TE types?

Could it be that certain metabolimic gene families display different levels of TE density? Perhaps it is the case that for a given biosynthetic pathway, say sugar synthesis, if a TE inserts near a gene of that pathway and modulates expression, that there will be disasterous effects on the phenotype of the organism such that there will be selection to never have a TE there. I have reason to believe that this trend is not set in stone and is context dependent. I would like to see if gene function and network connectivity drive TE density.

For certain genes it appears as though a high TE presence has been selected for. Remembering the fact that transposons can modulate gene expression in a number of ways (not always negatively) I have observed that sugar synthase genes in the octoploid strawberry (*Fragaria x ananassa* the farmed strawberry that we eat) are in the top 1% most transposon dense genes in the genome. This seems odd because my gut reaction would be to say that this would be bad, because the expression of sugar genes would be knocked down; you would have a less sugary strawberry and this wouldn't be a good trait. I think that the TE's either upregulated the sugar genes (by bringing in a novel promoter or upregulator of some sort) or that the TE's have inserted themselves and have remained unmethylated due to the selection that if they ever are methylated, the sugar production will drop like a brick, and this wouldn't be selected for. So there is actually selection to maintain high TE presence.

Objective:

Create a predictive model with machine learning to predict gene expression. What transposons and/or methylation data matter for predicting gene expression? What is necessary and sufficient for predicting gene expression?