Detecting transposons in plant-pathogenic fungi using Profile Hidden Markov Models

Parsa Hosseini[1,2], Arianne Tremblay[1], Nadim W. Alkharouf[2], Benjamin F. Matthews[1]
1 – Dept. Computer / Information Science, Towson University, Towson, MD
2 – USDA-ARS Soybean Genomics & Improvement Laboratory (SGIL), Beltsville, MD

Introduction

Very little is known about transposons and their role within plant-pathogenic fungi. What is known however is that with probabilistic models, we can create statistical solutions to predict plant-pathogenic fungal transposable elements.

We present a probabilistic model for the prediction of transposable elements in plant-pathogenic fungi.

Data and literature mining retrieved a set of 35 biotrophic plant fungal transposons, ranging from *A. bisporus* to *T. inflatum*. Homology and sequence analysis was performed for each pathogen, resulting in a multiple-sequence 'profile' model. This profile model, made up of sequence families, represents conserved regions in the sequence alignment and probabilistic states for each nucleotide.

Given a query sequence, calculations can be made as to how probable motifs within this sequence compare to each profile model.

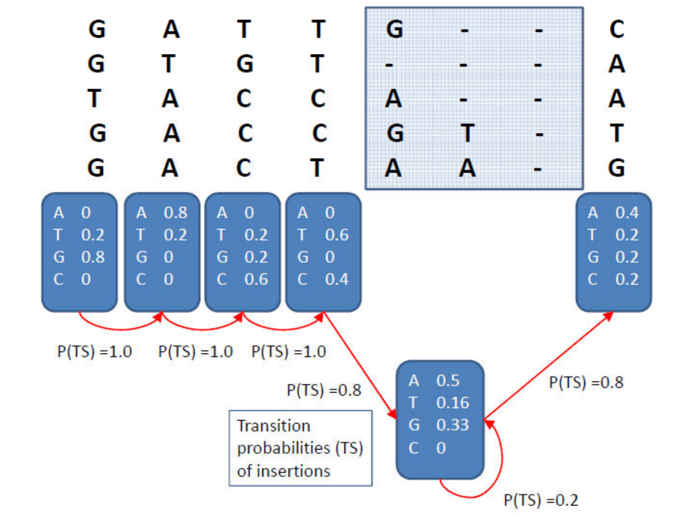

Figure 1a. Multiple sequence alignment states.
Creating a model of an MSA and using its transition-probabilities to move onto consecutive states.

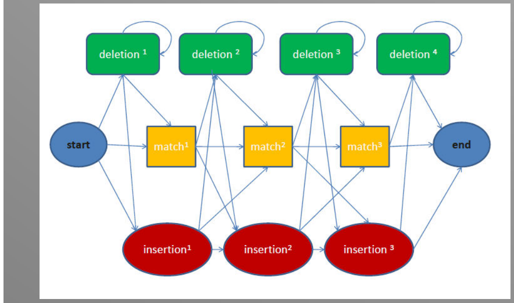

Figure 1b. States in an HMM.
An HMM model makes use of frequency-counts (in this case, amino-acid frequencies) to help guide transition states from state X -> Y

Profile Hidden Markov Models

Profile Hidden Markov Models (PHMM) are statistical techniques for modeling multiple sequence alignments; useful in protein family modeling and homology analysis. Given a multiple sequence alignment (MSA), PHMMs can probabilistically represent an MSA consensus column in the form of 'states' (Figure 1b). The probability of transitioning from a consensus column to the next is driven by the 'transition state' of moving from one state (i.e. insertion) to another (i.e. deletion). (Figure 1a).

Model development

Literature mining was used to create a set of 35 unique plant-pathogenic fungal TEs. For each transposon, tools such as COGEME and RepeatMasker were used to create a family of homologous sequences per TE. Family sizes varied for each TE, from 1 - 50 sequences. Transposons with < 5 sequences (n=25) were filtered to prevent weak model creation. Of the remaining 10 transposons (Table 1), multiple sequence alignments were created using MAFFT. The output CLUSTALW file was converted to Stockholm 1.0 format using custom Python scripts. For each Stockholm MSA, HMMER v3.0, a sequence homology software tool, was used to create a HMM model. Python scripts were made to facilitate iteration of a user-defined fasta file against each model as well as output the best-scoring hit (based on e-value).

ID	Pathogen	TE_Name	#/sequences
1	Agaricus Bisporus	ABR1	7
2	Botryotinia Fuckeliana	FLIPPER	6
3	Botrytis Cinerea	BOTY1	9
4	Fusarium Oxysporum	FOT1	13
5	Fusarium Oxysporum	IMPALA	16
6	Magnaporthe Grisea	MAGGY	22
7	Magnaporthe Grisea	MGR583	32
8	Magnaporthe Grisea	MGR586	20
9	Magnaporthe Grisea	POT2	50
10	Magnaporthe Grisea	SINE	5

Table 1. TE details
For each of the 10 transposon sequence families, multiple sequence alignment and PHMM model generation were executed.

Evaluation and Performance

For evaluation, 74,671 contigs were used from a *Glycine max* illumina mRNA-Seq study 7hrs after infection. Reads making up such contigs did not map to the Soybean genome. All contigs were queried against all 10 models in an average of 6.23 seconds, with 233 contigs mapping to at least 1 model with an e-value threshold 1e-3.

Conclusions

We illustrate a fast probabilistic model for predicting pathogenic fungal TEs. Our models shall help advance plant fungi TE research knowing it to be in its infancy.

We wish to thank the USDA-ARS SGIL and Towson University for their financial assistance and support.