

Identification of piggyBac Transposons in *Aedes aegypti*

Danielle Cisler
Univ. of Notre Dame
Dept. of Psychology
dcisler@nd.edu

Sarah Frost
Univ. of Notre Dame
Dept. of Comp. Sci. and Eng.
sfrost@cse.nd.edu

Dongyoung Shin
Univ. of Notre Dame
Dept. of Biological Sciences
dshin@nd.edu

Abstract

This study uses a joint hidden markov model and blast search approach to identify potential piggyBac type transposons in the newly assembled aedes aegypti genome. We identify several areas of interest that merit future research.

1 Tools and Procedures

The most important tools used in this study are discussed briefly. Their place in this study and why they were chosen are discussed in the context of the procedures.

1.1 Hmmer

Hmmer [2] is a software package using hidden markov models to identify search protein sequences for areas similar to models built by the user. Using hmmbuild, a tool in the suite, a model is built based on sequences given to the program. This model is then calibrated using another hmmer tool, hmmscalibrate. Finally, this calibrated model is used by hmmsearch to search the desired sequence, in our case the *aedes aegypti* genome.

1.2 BLAST

BLAST[4], the Basic Alignment Search Tool, is well known to biologists. In short, given two nucleotide or protein sequences, blast will return the best local alignment. In practice, when given a short sequence and a long sequence such as a genome, blast searches the genome for the best alignments. In our study, once a potential transposon was found, blast was used to search the genome for other sequences similar to the potential transposon.

1.3 Procedures

To begin exploring for transposons in the *aedes aegypti* genome, the sequences were aligned by CLUSTAL. Aligning the sequences is necessary for creating Hidden Markov Models (HMM) using Hmmer software.

The second step in searching for transposons in the *aedes aegypti* genome was using the phylogenetic tree to divide the piggybac sequences into related groups of sequences. Since the phylogenetic tree provides knowledge of which sequences are closely related, as according to their evolutionary history, the sequences were divided into five groups (The phylogenetic tree used was provided in class on October 13, 2005). The groups chosen were roughly the most related in their evolutionary history. The groups are as follows:

- Group 1: NhaPB;
- Group2: TniPB, PdoPB, AgaPB2, AgaPB3, AgaPB4, AgaPB5, Pigibaku1, Pigibaku2, AgaPB1, AgaPBD1 and Pigibaku3;
- Group3: AgaPB6, AgaPB7, AgaPB8, AgaPBD2, HsaPGBD4, CinPB3, CinPB4, AgaPB9, BmoPB1, TruPGBD4, CinPB2, TruPGBD1, TruPGBD2, DmLOOPER1m, AgaPBD 3, CinPBD1, CinPB1, CinPB5, TruPGBD3 and DmeCG9839;
- Group4: AgaPB10, DpuPOKEY05, TruPGBD5, M mupgbd5, HsaPGBD5, HsaPGBD1, Mmupgbd1p, HsaPGBD2, Mmupgbd2P, HsaPGBD3, HsaPGBD3P1, HsaPGBD 3P4, HsaPGBD3P3 and HsaPGBD3P2;
- Group5: OsaPB and DmeCG13151.

It is important to note that Pigibaku4 and Pigibaku5 were not found in the aligned file.

After creating fasta files containing the aligned sequences of each group, HMMs were created. The advantage of HMMs is that they utilize sound statistical theories to provide sequences taking into account all possibilities in regards to matches, mismatches and gaps [1]. To find the best sequences from hmmer, the models were created from each group of sequences. These models were then calibrated [2]. Calibration in hmmer is an important part of the process. By calibrating the model, sensitivity is increased providing more accurate E values. The calibrated models were then searched for in each of the six translated windows of the *aedes aegypti* genome. The searches provide consensus sequences for each model and window of the *aedes aegypti* genome with the respective E value to help determine biological significance.

After the hmmer models were used to search the aedes aegypti genome, the best results from each model and translation window were used as search sequences in blast. What constituted the “best” results was determined by hand. In each hmmer output, there was a band of results with very similar e-values followed by a significant drop off in e-value. The results in the first band were considered to be the “best” and worth blasting against the aedes aegypti genome.

Following the hmmer and blast searches, the results were compiled, identifying which sequences of the aedes aegypti genome were found by the searches and which search or searches identified them.

2 Results

Each hmmer model was used to search each of the six translation frames of the aedes aegypti genome. The best results from each of the hmmer model searches was used as the search sequence for a blast against the six translation frames. Tables 1, 2, 3, 4, 5, and 6 contain summaries of the sequences in the genome identified by the hmmer and blast searches. In the evidence column, the search or searches that identified that sequence is listed. The leading letter of each entry indicates what kind of search was conducted. “H” indicates a hmmer search, and “B” indicates a blast search. The first number indicates the model that was used. The second number indicates which translation frame was used to create the results. For blast searches, the third number indicates which hmmer “hit” was blasted. The blast results in parentheses follow the hmmer search from which the sequence used for the blast search was taken. The blast search is included for completeness, but it is not an additional piece of evidence since it was based on a sequence that the hmmer search already found in the genome. The results and their significance are discussed in the following section.

3 Discussion

The biological significance of our results has prompted debate among our group. The hmmer results had significant e-values, but limited sequence matches between the models and the aedes aegypti sequences as identified by the hmmer search. Similarly, the best of the blast results had very limited sequence overlap. However, because several locations of the aedes aegypti sequences were identified by multiple models and search approaches, we feel further investigation is warranted.

The hmmer searches based on the models identified thirteen segments of the aedes aegypti genome as being possible transposons. These results are included in the general results tables and can be viewed in table 7. As you can see in table 7, three of the hmmer-identified sequences were found only by the hmmer model. The ten other sequences were found by multiple models and by blast searches from multiple models and sequence frames. While the blast results were weak, the results from the hmmer searches had very

Table 1. Search results for the first protein translation frame of aedes aegypti genome

Frame	seq-f	seq-t	Evidence
Aedes-0F	199	428	B1-2-2
Aedes-0F	977	1150	B1-2-2, B2-4-1
Aedes-0F	5142	5661	H4-1 (B4-1-1)
Aedes-0F	5316	5665	B4-2-1, B4-4-1, B4-6-1
Aedes-0F	18215	18408	B1-2-1
Aedes-0F	18648	18805	B1-2-2
Aedes-0F	35519	35712	B1-2-2
Aedes-0F	42841	43070	B1-2-2
Aedes-0F	62543	62809	B1-2-2, B2-5-1, B2-4-1
Aedes-0F	95535	95715	B1-2-2
Aedes-0F	99445	99722	B1-2-2
Aedes-0F	117530	117666	B1-2-2
Aedes-0F	245744	246343	H2-1 (B2-1-1), B1-2-2
Aedes-0F	591749	592001	B4-1-1, B4-2-1, B4-4-1, B4-6-1
Aedes-0F	831520	832109	H2-1 (B2-1-2), B1-2-2, B2-1-2, B2-2-1, B2-2-2, B2-1-1 B2-4-1, 2-3-1
Aedes-0F	862977	863309	B1-2-2

Table 2. Search results for the second protein translation frame of aedes aegypti genome

Frame	seq-f	seq-t	Evidence
Aedes-1F	4476	4777	B2-5-1, B1-2-2
Aedes-1F	19243	19575	B1-2-2
Aedes-1F	21645	22145	H4-2 (B4-2-1), B4-4-1
Aedes-1F	21721	22037	B4-6-1
Aedes-1F	53002	53293	B1-2-2, B2-5-1
Aedes-1F	53915	54113	B4-6-1
Aedes-1F	115680	116258	H1-2 (B1-2-2), H2-2 (B2-2-1), B2-1-2, B2-1-1
Aedes-1F	309058	309558	H4-2 (B4-2-1), B4-4-1, B4-6-1
Aedes-1F	324963	325247	B1-2-2, B2-5-1
Aedes-1F	327481	327723	B4-6-1, B4-1-1
Aedes-1F	470225	470803	H1-2 (B1-2-2), H2-2 (B2-2-1) (B2-2-2), B2-1-1, B2-4-1, B2-3-1, B2-5-1
Aedes-1F	474449	474563	B1-2-2
Aedes-1F	502482	503917	H4-2 (B4-2-1), B4-6-1
Aedes-1F	1347265	1347415	B1-2-2

Table 3. Search results for the third protein translation frame of aedes aegypti genome

Frame	seq-f	seq-t	Evidence
Aedes-2F	267	559800	H2-3
Aedes-2F	52622	52808	B1-2-2
Aedes-2F	53743	53914	B4-6-1
Aedes-2F	62286	62543	B2-5-1
Aedes-2F	468561	468741	B1-2-2
Aedes-2F	503846	503947	B4-1-1
Aedes-2F	559219	559800	B2-4-1, B2-3-1, B1-2-2, B2-1-1, B2-1-2, B2-2-1, B2-2-2
Aedes-2F	572522	572781	B4-2-1, B4-4-1, B4-1-1, B4-6-1
Aedes-2F	599925	600143	B4-2-1, B4-4-1, B4-1-1
Aedes-2F	1347737	1348037	B1-2-2, B2-1-1, B2-2-1, B2-3-1, B2-4-1

Table 4. Search results for the fourth protein translation frame of aedes aegypti genome

Frame	seq-f	seq-t	Evidence
Aedes-0R	38884	39462	H2-4 (B2-4-1), B2-3-1, B1-2-2, B2-1-2, B2-2-1, B2-2-2, B2-1-1
Aedes-0R	70042	70351	B2-4-1
Aedes-0R	110344	110702	B4-2-1, B4-4-1
Aedes-0R	221823	222323	H4-4 (B4-4-1), B4-2-1, B4-1-1
Aedes-0R	305028	305469	B1-2-2, B2-4-1, B2-5-1
Aedes-0R	436836	436896	B1-2-2
Aedes-0R	597626	597973	B4-4-1, B4-6-1, B4-1-1
Aedes-0R	700037	70351	B2-5-1
Aedes-0R	1260519	1260748	B2-5-1

Table 5. Search results for the fifth protein translation frame of aedes aegypti genome

Frame	seq-f	seq-t	Evidence
Aedes-1R	1488	1835	B4-6-1, B4-4-1, B4-2-1, B4-1-1
Aedes-1R	42042	42344	B2-2-2
Aedes-1R	132762	133262	B4-2-1, B4-4-1
Aedes-1R	158253	158486	B2-3-1, B2-4-1
Aedes-1R	262561	263291	H2-5 (B2-5-1)
Aedes-1R	309913	310053	B1-2-2, B2-1-2, B2-2-1, B2-2-2,
Aedes-1R	436675	436822	B1-2-2
Aedes-1R	1342571	1342833	B2-1-1

Table 6. Search results for the sixth protein translation frame of aedes aegypti genome

Frame	seq-f	seq-t	Evidence
Aedes-2R	2203	2505	B2-1-1
Aedes-2R	16402	16578	B4-6-1
Aedes-2R	24673	24951	B1-2-2
Aedes-2R	27588	28100	H4-6 (B4-6-1), B4-2-1, B4-4-1, B4-1-1
Aedes-2R	42042	42344	B1-2-2, B2-2-1, B2-1-2, B2-4-1, B2-3-1
Aedes-2R	42658	42810	B1-2-2, B2-1-2, B2-3-1, B2-4-1
Aedes-2R	44933	45047	B1-2-2, B2-1-2
Aedes-2R	115909	116098	B1-2-2
Aedes-2R	485272	485464	B4-2-1
Aedes-2R	560171	560471	B2-1-1
Aedes-2R	1032960	1033094	B4-2-1, B4-4-4
Aedes-2R	1258648	1258791	B2-5-1

strong e-values, and indicate the segments of the genome that our approach deems of the most interest for future research.

3.1 Analysis of Models

The five different models each yielded different results when the hmmer search was conducted. The first model was comprised of only NhaPB which is from the *Nectria haematococca*. Only when this model was searched against the second translation of the *aedes aegypti* were significant supercontigs found. Eighteen results with decent E values were found. In model 2, each of the six translations found significant results. This model was comprised of sequences from *Trichoplusia*, *Anopheles gambiae* and *piggy-Bac*. The most extreme E values occurred when the second model was searched in the first translation. The supercontigs found when searching the third model were greatest in the second translation. Model three contained sequences found in *Anopheles gambiae*, *Homo sapiens*, *Ciona intestinalis*, *Bombyx mori*, *Takifugu rubripes*, and *Drosophila melanogaster*. Model four was successful in finding significant super contigs against all translations. Windows one and two seem to be the most successful in their findings. Model 5 is comprised up of sequences from *Oryza sativa* and *Drosophila melanogaster* [6]. This model failed to find any super contigs of significance.

4 Limitations and Future Work

Our results do indicate some potential areas of interest. These need to be verified through further study and biological inspection. However, the methods used here to find

Table 7. Results from hmmer searches

Aedes-0F	5142	5661	H4-1 (B4-1-1)
Aedes-0F	245744	246343	H2-1 (B2-1-1), B1-2-2
Aedes-0F	831520	832109	H2-1 (B2-1-2), B1-2-2, B2-1-2, B2-2-1, B2-2-2, B2-1-1 B2-4-1, 2-3-1
Aedes-1F	21645	22145	H4-2 (B4-2-1), B4-4-1
Aedes-1F	115680	116258	H1-2 (B1-2-2), H2-2 (B2-2-1), B2-1-2, B2-1-1
Aedes-1F	309058	309558	H4-2 (B4-2-1), B4-4-1, B4-6-1
Aedes-1F	470225	470803	H1-2 (B1-2-2), H2-2 (B2-2-1) (B2-2-2), B2-1-1, B2-4-1, B2-3-1, B2-5-1
Aedes-1F	502482	503917	H4-2 (B4-2-1), B4-6-1
Aedes-2F	267	559800	H2-3
Aedes-0R	38884	39462	H2-4 (B2-4-1), B2-3-1, B1-2-2, B2-1-2, B2-2-1, B2-2-2, B2-1-1
Aedes-0R	221823	222323	H4-4 (B4-4-1), B4-2-1, B4-1-1
Aedes-1R	262561	263291	H2-5 (B2-5-1)
Aedes-2R	27588	28100	H4-6 (B4-6-1), B4-2-1, B4-4-1, B4-1-1

transposons in the *aedes aegypti* genome do have some limitations. The sequences chosen for the models was done in a relatively arbitrary fashion by dividing the given piggyBac transposons into groups of roughly equal size and relatedness.

Using only transposons from other mosquitoes or insects might lead to clearer results, as would dividing the groups based more closely on relatedness. Another limitation of the models are the limited number of sequences in which they were derived. HMMs are more sound when the number of sequences are increased. Combining these two limitations, having more transposons from insect or mosquito genomes, may strengthen the findings.

To improve the project further, the supercontigs found in from the hmmer searches could have been blasted against nucleotides, instead of the protein translations. Blasting against the nucleotide sequence would allow a natural shifted reading from to occur. More powerful methods also exist. Instead of using hmmer or BLAST, a more advanced method such as neural networks, context free grammars, or graph algorithms could be used to find transposons.

References

- [1] M. David. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2nd edition, 2004.
- [2] S. Eddy. Hmmer user's guide: Biological sequence analysis using profile hidden markov models. 2003.
- [3] M. Fraser, J. Brusca, G. Smith, and M. Summers. Transposon mediated mutagenesis of baculoviruses. *Virology*, 145:356–361, 1985.
- [4] NCBI. Basic local alignment search tool (blast). [http : //www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/).
- [5] H. Robertson. Evolution of dna transposons in eukaryotes. In N. Craig, R. Craigie, M. Gellert, and A. Lambowitz, editors, *Mobile DNA 2*, pages 1093–1110. ASM press Washington, D.C., 2002.
- [6] A. Sarkar, C. Sin, Y. S. Hong, J. Hogan, M. F. aser, H. Robertson, and F. Collins. Title goes here. *Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences*, pages 173–180, 2003.