

Members of the *piggyBac* transposon family in *Aedes Aegypti*

Amy McHenry¹, Deborah Thomas² and Scott Christley²

¹Dept. of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556

²Dept. of Computer Science, University of Notre Dame, Notre Dame, IN 46556

Abstract

piggyBac is a Class-II transposable element with 13-bp inverted terminal repeats. It was originally isolated from the cabbage looper moth, *Trichoplusia ni*. In this article, we analyze the *Aedes aegypti* genome to discover possible members of the *piggyBac* family of transposons using a combination of techniques. Specifically, we construct a set of Hidden Markov Models based upon multiple sequence alignments of different subsets of known protein sequences from the *piggyBac* transposon family; those models are then used to search the genome for high probability matches. We also perform BLAST searches against the genome using the BLOSUM62 scoring matrix with highly conserved protein sequences of various transposons; this provided secondary evidence for matches found by the HMMs. Results indicate that a number of transposable elements exist in the *A. aegypti* genome.

Introduction

piggyBac is a Class-II transposable element first isolated from the cabbage looper moth, *Trichoplusia ni*. Its single ORF codes for a 594-residue transposase that mediates cut-and-paste movement of the element between TTAA target sites (Cary *et al.*, 1989). It has been shown to be mobile in a large range of organisms and for this reason is used extensively as a gene vector. Members of the *piggyBac* family of elements have been shown to be present in the *Anopheles gambiae* genome, as well as in such divergent organisms as *Takifugu rubripes* (pufferfish), *Daphnia pulex*, and humans (Holt *et al.*, 2002, Sarkar *et al.*, 2003). A group of 50 *piggyBac*-like sequences have been analyzed and organized phylogenetically (Sarkar *et al.*, 2003). (For a comprehensive review of transposable elements in mosquitos, see Tu *et al.*, 2004.)

Aedes aegypti is the primary arthropod vector for both dengue viruses and yellow fever worldwide. More than 2.5 billion people worldwide are at risk for contracting dengue viruses. Annual incidence is in the tens of millions and approximately 24,000 people die each year (Severson *et al.*, 2004). There is no specific pharmaceutical treatment for dengue fever (WHO, 2005). Greater understanding of the genetics of this organism are vital for prevention of these diseases.

For our analysis techniques, we constructed a set of Hidden Markov Models based upon multiple sequence alignments of different subsets of known protein sequences from the *piggyBac* transposon family; those models are then used to search the genome for high

probability matches. We also perform BLAST searches against the genome using the BLOSUM62 scoring matrix with highly conserved protein sequences of various transposons; this provided secondary evidence for matches found by the HMMs.

Hidden Markov Models (HMMs) are probabilistic models where a single random variable represents the states of the system. There are probabilities associated with transitioning from one state to another, and each state has a set of emission probabilities that characterizes the output produced by the system while in that state. For biological sequence analysis (Durbin *et al.*, 1998), the general form of an HMM is specialized for the specific type of analysis being performed. In particular, we use the HMMER software package for constructing HMMs and for searching the genome.

BLAST stands for Basic Local Alignment Search Tool. It performs local alignment between sequences, and the sequences can be either proteins or nucleotides. The algorithm uses a dynamic programming approach. It can use either a PAM matrix or a BLOSUM matrix. These matrices allow for points to be gained for correct nucleotide matches while losing points for gaps in the sequence or mismatches. Once the algorithm is run to completion, it gives the top matches along with an “expectation value”. This value corresponds to the likelihood that the number of matches were found by chance. Hence the lower this value, the better the match is.

Materials and Methods

Figure 1 shows the workflow process we performed for analyzing the *A. aegypti* genome for transposons. Specifically, we performed two independent lines of analysis then combined them together for the final results as well as some pre-processing on the genome data set.

The *A. aegypti* genome is rather large at 1.4 Gbp, and the nucleotide sequences were provided to us in over 4700 supercontigs. Unfortunately, we only had protein sequences for the *piggyBac* transposons, so we needed to translate the genome’s nucleotides into protein sequences. We wrote a Bioperl (Stajich *et al.*, 2002) program that read in each supercontig for the genome, translated it into a protein sequence for all six possible reading frames, and generated six files (one for each reading frame) in FASTA format with the protein sequences. Analysis for the HMM and BLAST searches was then performed on each of these files independently.

CLUSTALW was used to perform multiple sequence alignment of previously described *piggyBac* family sequences (Sarkar *et al.*, 2003). Four separate alignments were performed. All fifty divergent sequences were aligned. Three other alignments were performed using sequences from the three more closely related branches indicated by Sarkar *et al.* in Figure 2 (2003). These four alignments were used as the basis for searching the genome with HMMs and BLAST.

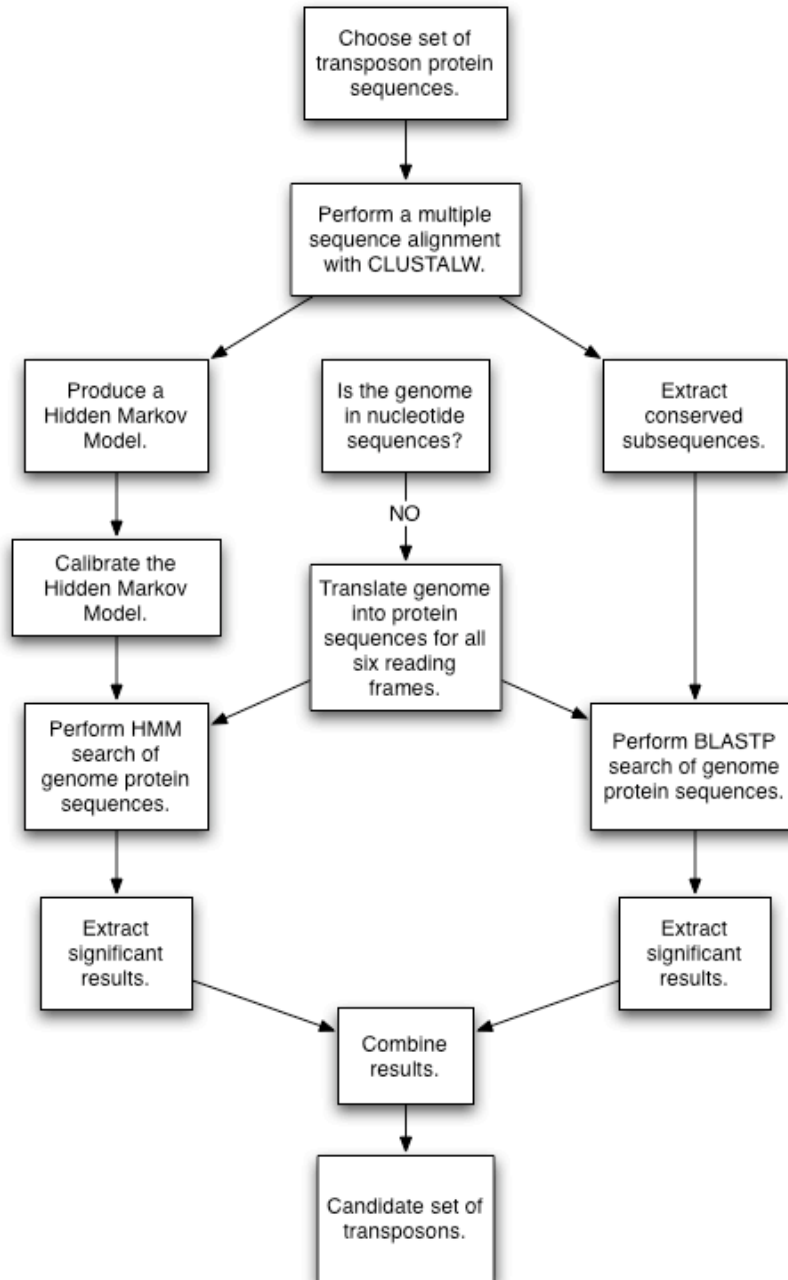


Figure 1. Workflow of transposon analysis for *A. aegypti* genome. HMMER was used for the HMM analysis, and we wrote a Bioperl script to translate the genome nucleotide sequences into protein sequences.

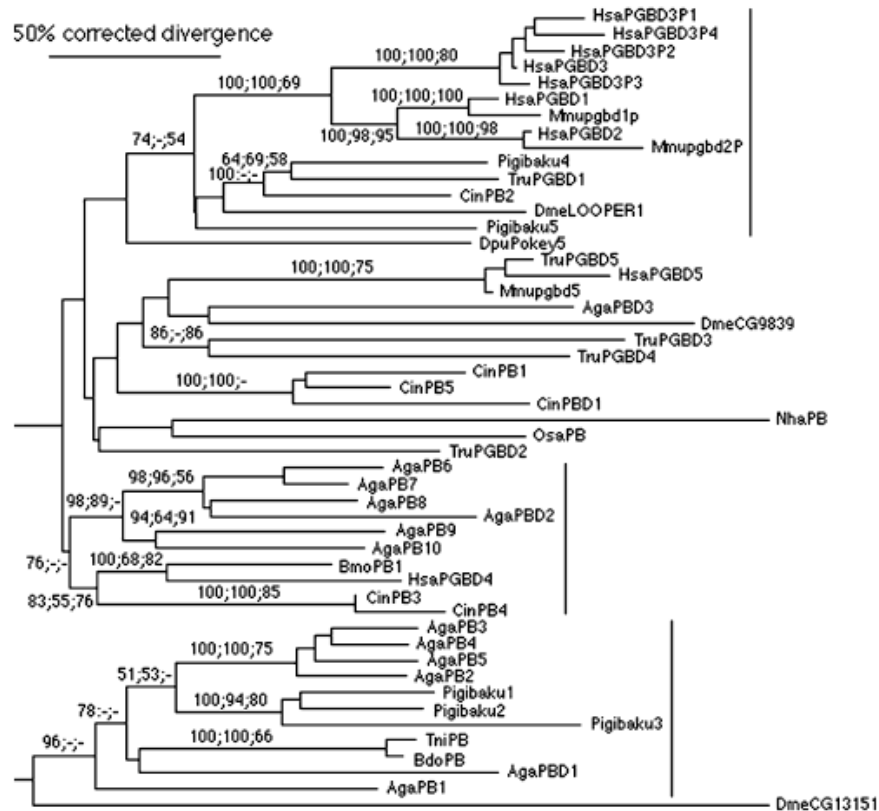


Figure 2. Phylogenetic tree of *piggyBac*-like sequences. CLUSTALW was used to generate multiple sequence alignments of all fifty sequences as well as three alignments of the sequences of each closely related branch, indicated with the vertical lines on the right. (Sarkar *et al.*, 2003)

HMMER (Eddy 2003) builds profile HMMs which are statistical models of a multiple sequence alignment; essentially the profile is a statistical description of the consensus sequence produced by the multiple sequence alignment. HMMER uses the Plan 7 architecture that has states to represent each position (column) in the multiple sequence alignment, insert and delete states for representing insertions and deletions, and a few other special states. Each positional state, called M (match) states, has a vector of emission probabilities for all 20 amino acids determined by the frequency of each amino acid observed in the corresponding column of the multiple sequence alignment.

The multiple sequence alignments produce by CLUSTALW were used to build four profile HMMs using the default parameters of hmmbuild; the model is configured to find one or more non-overlapping alignments with global alignment in respect to the whole model and local alignment in respect to the sequence. The models were calibrated with hmmscalibrate which fits an extreme value distribution to the scores of a large set of synthesized random sequences; this allows the search procedure to produce significance statistics for matches. Lastly, the six translated reading frames of the genome were independently searched with hmmsearch, producing 24 separate result files.

For the BLAST searches, we took the multiple sequence alignment and picked out four sequences that were highly conserved. These sequences were about 400 proteins long. Before using BLAST to search, the six translated reading frame files had to be formatted into a database using the formatdb program; we used the default settings for the formatting which takes protein sequences as input in FASTA format. The next step was to run blastall using a protein search on each of these translated databases on the four sequences. We used the default search settings of a gapped alignment with a BLOSUM62 scoring matrix and a threshold of eleven proteins for extending the sequence.

Results and Discussion

Both HMM and BLAST returned significant matches for members of the *piggyBac* transposon family. HMM returned 25 matches (score 100 or higher). A number of the matches were achieved with multiple searches. For example, a match with an acceptable score was found on contig 1.316 using two different HMMs. Table 1 compares the HMM matches found using the four different HMMs.

Table 1

Protein Frame	Hidden Markov Model			
	All sequences	Branch 1	Branch 2	Branch 3
1	5	1	0	2
2	6	4	0	2
3	4	1	0	1
4	5	3	0	1
5	4	2	0	1
6	1	1	0	0
Totals	25	12	0	7

BLAST returned 79 matches (score 50 or higher). Table 2 shows the location of the top matches from the BLAST search. A consensus sequence of approximately 100 bp was chosen from highly conserved regions of the multiple sequence alignment of the stated sequences.

It is interesting to note that the HMM and BLAST searches show different patterns. It is clear that the HMM using a multiple sequence alignment returned the most results within the acceptable range (See Table 1). However, the BLAST search using consensus sequences from multiple sequence alignment of branches 2 and 3 returned the most results within the accepted range (See Table 2). The BLAST search returned more matches overall, but with a lower score threshold. We were unable to compare the two sets of results directly because the contig numbering systems were not comparable. However, it appears that using both methods will give a clearer overall picture of the presence of transposable elements in the *A. aegypti* genome.

Table 2

Protein Frame	Consensus Sequence used for BLAST search			
	All sequences	Branch 1	Branch 2	Branch 3
1	2	2	5	10
2	1	2	3	5
3	4	3	7	9
4	3	2	9	8
5	3	2	5	4
6	1	1	6	5
Totals	14	12	35	41

One apparent concern with our analysis was the need to translate the genome from nucleotide into protein sequences and search each reading frame separately. This analysis technique leaves open the possibility of missing transposable elements due to frame shifts, caused by insertion or deletion mutations in the middle of the element, which would generate a completely different protein sequence. In future analysis, we could obtain the nucleotide sequences of the *piggyBac* transposon family, or we could use the GeneWise (Birney *et al.*, 2004) program that searches all six reading frames simultaneously.

It is clear that members of the *piggyBac* transposon family are present in the *A. aegypti* genome. It will be interesting to further characterize these elements and determine their phylogenetic relatedness to other members of the family. The characterization of these elements is also useful in light of the extensive use of *piggyBac* elements as gene vectors for insect transformation. Use of the *piggyBac* element for transposition in *A. aegypti* has previously been reported by Lobo *et al.* (1999).

References

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Research* 14:988-995.

Cary LC, Goebel M, Corsaro BG, Wang HG, Rosen E, Fraser MJ (1989) Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology* 172(1): 156-169.

Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, UK.

Eddy S (2003) HMMER User's Guide. <http://hmmer.wustl.edu/>

Holt RA *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129-149.

Lobo N, Li X, Fraser MJ Jr. (1999) Transposition of the *piggyBac* element in embryos of *Drosophila melanogaster*, *Aedes aegypti* and *Trichoplusia ni*. *Mol Gen Genet* 261(4-5): 803-10.

Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, Collins FH (2003) Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. *Mol Gen Genomics* 270:173-180.

Severson DW, Knudson DL, Soares MB, Loftus BJ (2004)

Stajich JE, et al (2002) The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research* 12(10):1611-1618.

Gu Z, Coates C (2004) Mosquito transposable elements. *Insect Biochem Mol Bio* 34: 631-644.

NCBI Tools for data mining. <http://www.ncbi.nlm.nih.gov/Tools/>