

Analysis of *P* element Transposon Sequences in *Aedes aegypti*

Trevor Cickovski¹, Jim Hogan², Ryan Kennedy¹

¹Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, 46556

²Department of Biology, University of Notre Dame, Notre Dame, IN, 46556

***P* elements have been found in several genomes outside of *Drosophila*, where they were initially found. We look inside yet another genome, the *Aedes aegypti*, and search for *P* elements using various *in silico* tests. We feel confident that *P* elements can be found in the *Aedes aegypti* genome, however that they may have ancient lineages and perhaps be inactive in the genome.**

Introduction

In analyzing genomes, locating transposable elements is an important step in the process of finding areas which encode genes. Class II describes a subset of transposons which have specific properties. In eukaryotic genomes, Class II transposable elements are uniquely identified by their inverted repeats on opposing ends and internal coding regions. They also can move through DNA intermediates and particularly Class II transposable elements have been used as gene vectors (1), motivating the understanding of the evolution of these elements.

An example of a previously studied Class II transposable element is the *P* element (2). This transposable element was originally identified in the fruit fly *Drosophila melanogaster* and is 2907 base pairs long and has 31 base pair inverted repeats. It was the first transposon used

as a gene vector (3), but its use has been limited in the past because of mobility constraints in organisms outside of the family *Drosophilidae* (4), however identifying and isolating this element in these other organisms can help us to understand its genetics and evolution.

DNA elements with similarity to the *P* element from *Drosophila melanogaster* have been found in three other species: *Lucila cuprina* (5), *Musca domestica* (6), and most recently the mosquito *Anopheles gambiae* (2). This latter discovery has motivated us to search for *P* elements in the yellow fever mosquito *Aedes aegypti*. Lobo *et al.* (7) have used a different Class II transposable element, the *piggyBac* (8), as a gene vector to transform the white-eyed *khw* strain in the *Aedes aegypti* genome. In studying *P* elements in *Aedes aegypti*, we hope to view its potential for genetic engineering as a long term goal.

We first present our approach to locating *P* elements in the *Aedes aegypti* genome. Next we will present our results and any useful nucleotide sequences that we find, along with a discussion of our results.

Approach

Sequence Alignment

We use the tool CLUSTAL-X (9) to produce a multiple sequence alignment (MSA) for 138 different sequences of *Aedes aegypti*. CLUSTAL-X is an improvement of the CLUSTAL MSA program which provides parameter weights. CLUSTAL-X first performs pairwise alignments of the *Aedes aegypti* sequences, then constructs a phylogenetic tree using alignment scores, and finally uses relationships within this tree to sequentially align the sequences through dynamic programming (10). The beginning of an example CLUSTAL-X output alignment of our *Aedes* sequences is shown below:

CLUSTAL X (1.81) multiple sequence alignment

```
gi|58386693|ref|XP_314978.2| -----
gi|55239599|gb|EAA10370.3| -----
gi|58381516|ref|XP_311297.2| -----
gi|55243097|gb|EAA06886.2| -----
gi|58388308|ref|XP_316192.2| -----
gi|55238944|gb|EAA44169.2| -----
...
gi|28630150|gb|AAM94946.1| -----
gi|23664272|gb|AAN39288.1| -----
gi|38564327|ref|NP_078948.2| -----
gi|10439964|dbj|BAB15609.1| -----
gi|57109260|ref|XP_544956.1| MTHNSEIQIQTKNHKLNKVNQSDPEARVLPASAGPAPLCAQGCECEGLLT
gi|50746565|ref|XP_420555.1| -----
```

Building a Hidden Markov Model

We subsequently use these alignments to construct a Hidden Markov Model (HMM) (11) using the tool HMMER (12). An HMM is an abstract machine which is built using a training set of data, constructing its states and probabilities for transitioning between states based on this data. Given a state, the machine outputs a symbol at a certain probability. In the case of protein alignments, the symbols will be one of the 20 amino acids (*A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y*). The output file from `hmmbuild` shows the length of the model in nodes and for each node, a list of scores - one for each amino acid as an emitted symbol, one for each amino acid as an inserted emitted symbol, and one for each state transition (these are log odds scores which are later converted to probabilities).

We then calibrate this model, running the `hmmcalibrate` tool of HMMER. When searching the *Aedes* genome, our output will consist partly of E-values (expectation values), which is helpful in telling us in addition to the score we receive how many hits expected to achieve that score by chance. It thus improves our ability to discern useful hits.

Searching the Genome

We search the *Aedes aegypti* genome using the tool `hmmsearch` on input genome files in FASTA (13) format. Since we want to use these results to search for P elements, we must take into account all three frames of reference on both strands of the DNA in particular to avoid risks of errors due to frame shifting (*i.e.* if one amino acid is inserted improperly, offsetting subsequent codons). Therefore, for each HMM we run `hmmsearch` six times, one for each frame of reference in the *Aedes* genome. Each time `hmmsearch` is run, it produces an output file which shows all hits in order of significance, including their scores, E-values, and occurrence frequency. Here is an example of one of our outputs:

Sequence	Description	Score	E-value	N
Aedes-0F	aegypti supercontig 1.152	192.0	7.5e-55	1
Aedes-0F	aegypti supercontig 1.309	166.2	4.4e-47	1
Aedes-0F	aegypti supercontig 1.723	148.1	1.2e-41	1
Aedes-0F	aegypti supercontig 1.1056	145.4	7.8e-41	1
Aedes-0F	aegypti supercontig 1.421	53.4	4.1e-13	1
Aedes-0F	aegypti supercontig 1.98	49.7	5.1e-12	1

We consolidated all significant hits (with E-values below $9.9e-06$), and then by extracting flanking amino acids and subsequently their nucleotides, we looked for the inverted repeats which occur at the ends of P elements, intending to find the transposon locations.

Scripting and Tools

Much of the runs described above were helped through shell scripting in BASH (14). All of the HMMER searches were run on an Apple Workgroup Cluster with 64-bit G5 processors running Mac OS X using the Xgrid (15) distributed architecture. Xgrid has a queueing system and is built for scalability with job complexity, which helped on our large searches, some of which took multiple hours.

Figure 1 shows the overall process we used in finding *P* elements in the *Aedes aegypti* genome.

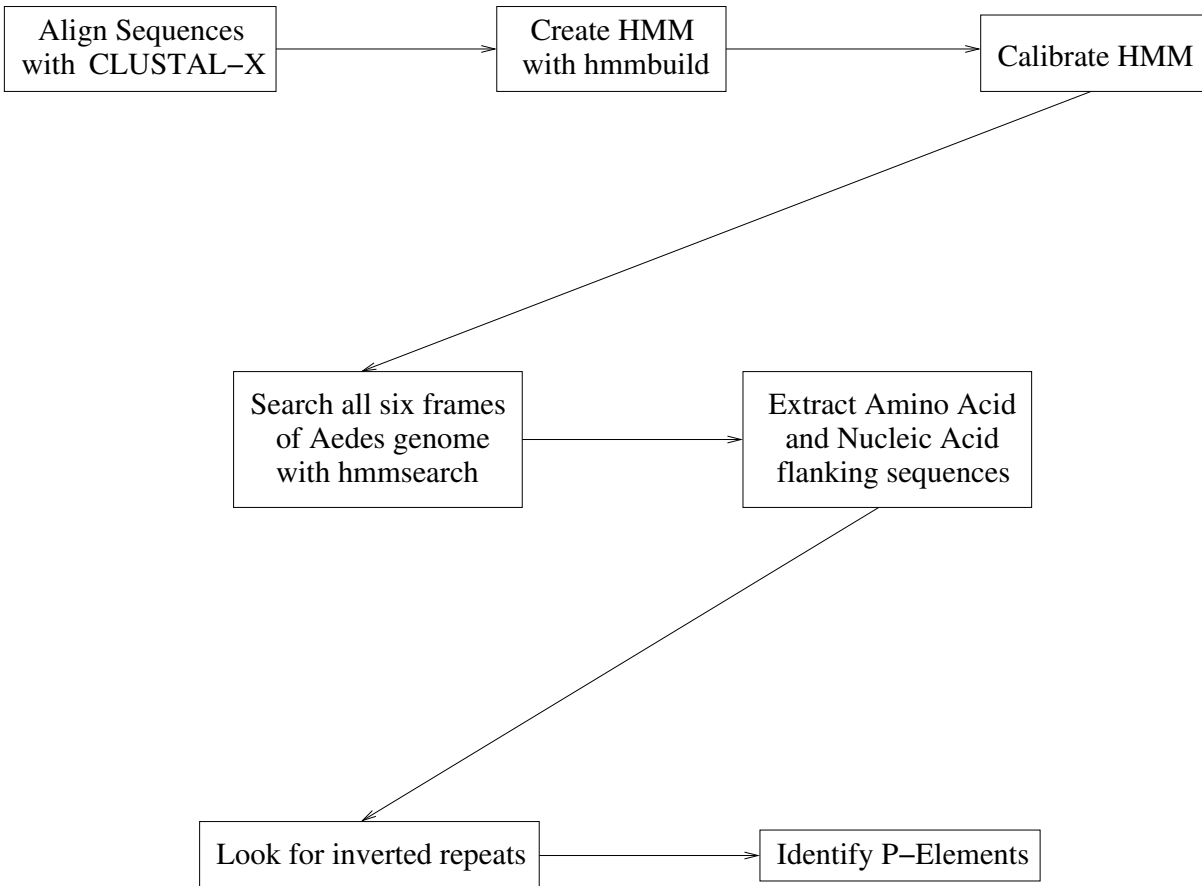


Figure 1: Block diagram of our process of location *P* element sequences in the genome of *Aedes aegypti*.

Results

Based on the output of `hmmsearch` we were able to retrieve the supercontig number for each of our results. Using this, along with the starting and ending nucleotide positions of our sequences, we wrote bash scripts to extract the appropriate regions from the *Aedes* genome using the corresponding frame of reference. While performing this, we realized that frames 4-6 (re-

verse sequences) began their nucleotide counts from the end rather than the beginning, throwing much of our data off. Thus rather than 62 sequences, we analyze in depth the 26 sequences that we retrieved while searching in the forward direction (frames 1-3) due to time constraints and the unexpectedness of this problem. However, immediate future work should consist of analyzing the other 36 sequences as well.

We first illustrate the raw sequence data and further information on each sequence as results of passing them through BLAST. Next, we align all 26 sequences and compare them using CLUSTAL-X, and finally analyze their relationships and time of divergence using a Phylogenetic tree.

Transposon Sequences

After retrieving our supercontigs and extracting the transposon sequences along with enough nucleotides to encompass the inverted repeat length for Class II elements, we ran the results through BLAST. Below we show all resulting sequences with expected values below e^{-12} , ordered by the best hit according to expected value.

Supercontig	Evalue	Genomic Hits	Inverted Repeats	Hit
1.1056	e-41	6	cagcgacattcgtcctctatagtttattacctctatt	PREDICTED: similar to THAP domain containing 9 [Danio rerio]
1.604	e-24	137	gacgtaacaagtgaaaaaacgtaaaatttaggtgatttacacatttttaccacaactt (etc)	ENSANGP00000005847 [Anopheles gambiae str. PEST]
1.893	e-18	119	catgaaaacgactgtttaattcacgcaaggcccttaacaataaaatcagcaacaaactg	hypothetical protein LOC54875 [Homo sapiens]
1.327	e-15	126	cacgtggatatatggacggtcct	No significant similarity found
1.421	e-13	9	aagtggtagagcccgcgctacacagcaaagccatataaaggtgtctggt (etc)	PREDICTED: similar to THAP domain containing 9 [Strongylocentrotus purpuratus]
1.639	e-09	218	tcttcgcgcaactctccattcatagactctg	PREDICTED: similar to THAP domain containing 4 [Strongylocentrotus purpuratus]
1.458	e-09	2	ccgcggtacaagcaaagccatgctgaagtgtctgggttcgagtcggtcggtccag (etc)	PREDICTED: similar to THAP domain containing 4 [Strongylocentrotus purpuratus]
1.5	e-06	386	acatacagcttctcctctcttcggtgattgatgctctgaaata (etc)	reverse transcriptase-like protein
1.927	e-06	150	gagtcggtcatcttttcttctaaccatctttg atgcat	transposase [Drosophila melanogaster]
1.266	e-05	4	gtccacaaattacgtaacgcttaaggggaaggggtaggctcaaacattacgactcata	PREDICTED: similar to THAP domain containing 9 [Strongylocentrotus purpuratus]

As can be seen through the BLAST results with our inverted repeat sequences, half corresponded to a THAP domain of proteins which are found in *P* elements (16). These probably initially came from p-elements but the genome assimilates these sequences accordingly for its own purposes, for example if a repressor or enhancer was needed the zinc finger protein of a *P* element could be used as a binding sequence. Note also that the sequences corresponded

to other sequences found in many different organisms, such as the *Danio rerio* zebrafish, the *Strongylocentrotus purpuratus* sea urchin, humans (*Homo sapien*), as well as the *Drosophila melanogaster* fruit fly.

Also, the various sequences had different numbers of genome hits. While one may suspect that sequences with inverted repeats would be found most often in the genome, that was not always the case, as many sequences were also found without inverted repeats that occurred over 100 times in the genome. For example, supercontig 1.227 contained two sequences, with e-values of e-08 and e-10, which occurred 162 and 136 times in the *Aedes* genome respectively. This suggests the existence of several incomplete transposons caused by mutations and random insertions which over long periods of time result in divergence significant enough to result in non-identifiability through BLAST.

Aligned Sequences

Figure 2 shows a section of our CLUSTAL-X alignment of our 26 sequences.

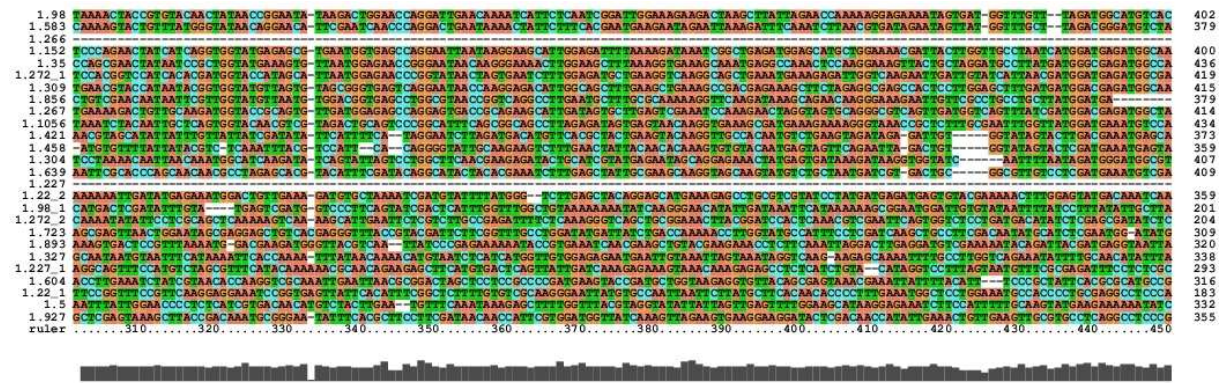


Figure 2: Multiple alignment of our 26 transposon sequences using CLUSTAL-X.

From this figure we can see that the multiple alignment of our 26 transposons produced by CLUSTAL-X is not strong, although some areas (*i.e.* nucleotides 340-350) align decently, others (420-430 for example) aligned poorly. This suggests a historical divergence of the sequences

which occurred a long time ago.

Phylogenetic Tree

Phylogenetic trees are useful for determining common ancestors among a group of nucleic acid sequences (10). Branch length in the tree is an illustration of relation distance; more distantly related sequences have deeper branches, where strongly similar sequences have adjacent branches in the tree and are joined to a common node. We show evolutionary relationships between our transposons using the UPGMA Phylogenetic tree shown in Figure 3. This tree was generated using MEGA version 3.1 (17).

Based on this tree along with our results from BLAST and CLUSTAL-X, there appears to be three clades of sequences that cluster together with spurious sequences intermixed. For example, the top seven sequences from the Phylogenetic tree correspond to THAP sequences in the zebrafish, followed by a section in the tree with many THAP sequences from the sea urchin (1.421, 1.639, 1.458 and 1.266). There are also some miscellaneous sequences, including some from *Anopheles gambiae* (1.227-1 and 1.604).

Conclusions

While *P* elements appear plentiful in the *Aedes aegypti* genome, a genome where they previously have never been described, they appear to be in ancient lineage with several sequences that have been domesticated by the host genome and have been inactivated, based on *in silico* tests. For example, they have little resemblance to one another at least on the nucleotide level. Also, the presence of inverted repeats seems not to be related to the number of copies of sequences in the genome, which may be due to divergence of inverted repeats from one another, or truncation of the transposon sequences such that the inverted repeats are no longer identifiable through BLAST.

However, there still seems to be some similarity at the amino acid level as shown by our BLAST results which found several proteins from the THAP domain. Overall, we found several intact *P* elements in the *Aedes aegypti* genome, some of which occurred more than 100 times. We also may be missing many of the *P* elements in the genome, because *hmmsearch* does not properly account for frame shifts. A future alternative may be to use *Genewise*, although that will be much a much more computationally intensive option.

References and Notes

1. C. Horn, E. A. Wimmer, *Dev Genes Evol* **210**, 630 (2000).
2. A. Sarkar, *et al.*, *Insect Biochemistry and Molecular Biology* **33**, 381 (2003).
3. G. Rubin, A. Spradling, *Science* **218**, 348 (1982).
4. D. A. OBroccha, S. P. Gomez, A. M. Handler, *Molecular and General Genetics* **225**, 387 (1991).
5. H. D. Perkins, A. J. Howells, *Proceedings of the National Academy of Sciences USA* **89**, 10753 (1992).
6. S. H. Lee, J. B. Clark, M. G. Kidwell, *Insect Molecular Biology* **8**, 491 (1999).
7. N. Lobo, A. Hua-Van, X. Li, B. M. Nolen, M. J. Fraser, *Insect Mol Biol* **11**, 133 (2002).
8. A. Sarkar, *et al.*, *Molecular and General Genetics* **270**, 173 (2003).
9. EBML-EBI, Clustalw, <http://www.ebi.ac.uk/clustalw/> (2005).
10. D. M. Mount, *Bioinformatics: Sequence and Genome Analysis, Second Edition* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2004).

11. N. C. Jones, P. A. Pevzner, *An Introduction to Bioinformatics Algorithms* (MIT, Cambridge, MA, 2004).
12. WUSTL, Hmmer, <http://hmmer.wustl.edu> (2005).
13. FASTA, Fasta sequence comparison, <http://fasta.bioch.virginia.edu> (2005).
14. GNU, Bash - gnu project - free software foundation (fsf), <http://www.gnu.org/software/bash/bash.html> (1998).
15. Apple, Apple - mac os x server - xgrid, <http://www.apple.com/server/macosx/features/xgrid.html> (2005).
16. S. E. Hammer, S. Strehl, S. Hagemann, *Molecular Biology and Evolution* **22** (2004).
17. S. Kumar, K. Tamura, M. Nei, *Briefings in Bioinformatics* **5**, 150 (2004).

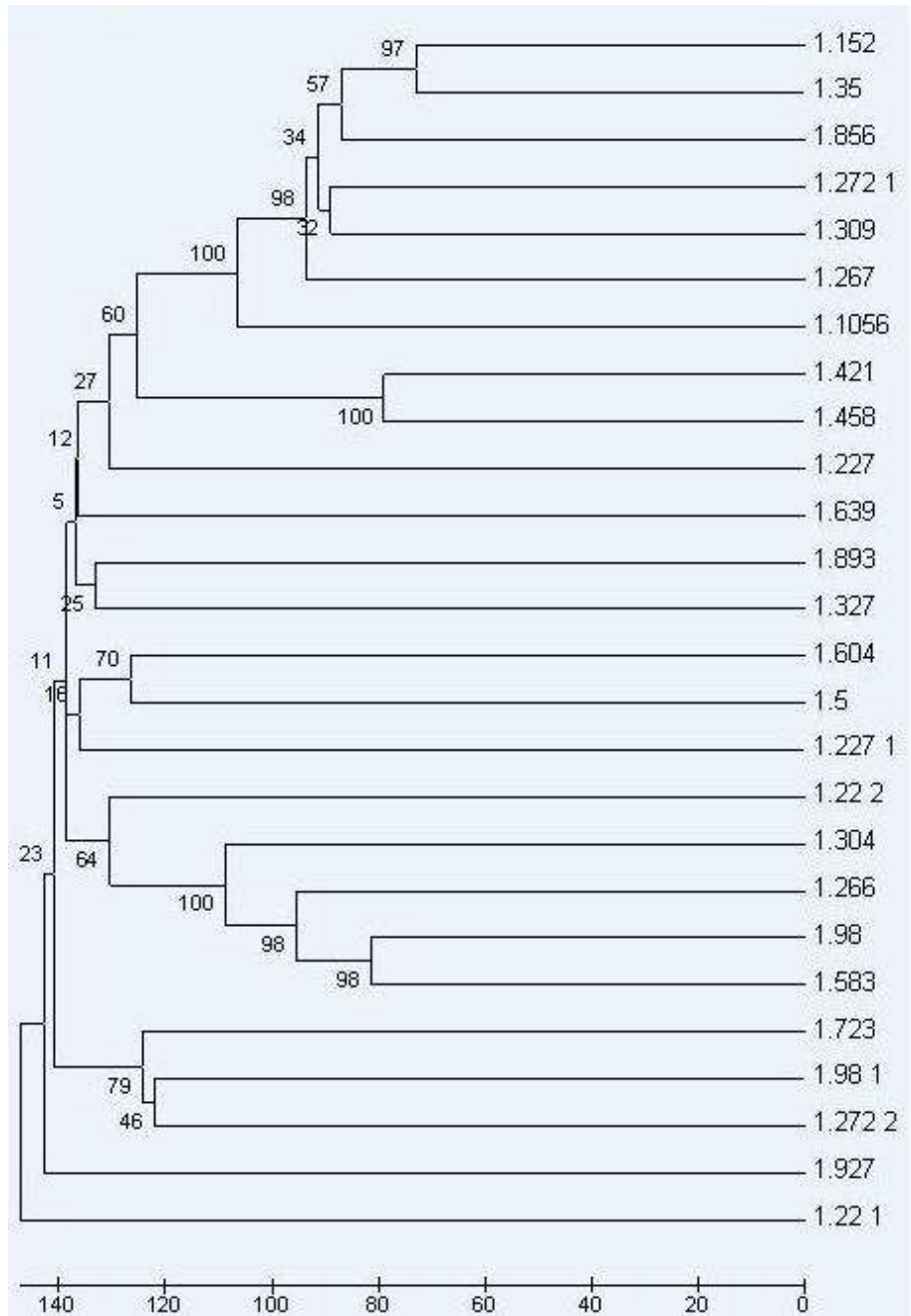


Figure 3: Phylogenetic tree.