

Using Perl scripts to help identify *P* element Transposon Sequences in *Aedes aegypti*

Ryan C. Kennedy and Jim Hogan

Department of Computer Science and Engineering and Department of Biology, University of Notre Dame

Abstract—We search within the *Aedes aegypti* genome for *P* elements. We utilize a set of tools to accomplish this and rely on Perl scripts to parse some of our data. We are confident that we will identify *P* elements in the newly released *Aedes aegypti* genome.

I. INTRODUCTION AND BACKGROUND

LOCATING transposable elements is an important step involved when analyzing genomes. Transposable elements can be uniquely identified by their inverted repeats and their internal coding regions. Transposons are important to study because they are useful as gene vectors, particularly Class II transposons [1]. This, coupled with the transposon’s ability to move through DNA intermediates, has motivated further understanding of these elements.

The *P* element is a well-known Class II transposable element [2]. It was originally identified in the *Drosophila melanogaster* genome. It has 2907 base pairs and 31 base pair inverted repeats. This *P* element was the first transposon used as a gene vector, but its use has been limited in part because of its mobility constraints in organisms outside the *Drosophilidae* family [3].

DNA elements similar to the *P* element from *Drosophila melanogaster* have been found in three additional species: *Lucila cuprina* [5], *Musca domestica* [6], and *Anopheles gambiae* [2]. Such discoveries have motivated our desire to search for *P* elements in the newly released *Aedes aegypti* genome. We hope that we will contribute to the genetic engineering field by identifying and studying *P* elements from the *Aedes aegypti* genome.

II. METHODS AND RESULTS

The abstract view of the entire project is shown in Figure 1 and explained in depth in *Analysis of P element Transposon Sequences in Aedes aegypti* [6]. Perl scripts were utilized in parsing *Blast* results files. A sample portion of the NCBI *Blast* results file is shown in Figure 2. A major script involved creating two output files. The first output file would contain the supercontig number, the expect value, the start and end of the sequence, and the frame, while the second output file

would contain the supercontig number and the amino acid sequence. This data was to be extracted only if the expect value was less than specified (usually 1e-10). The script utilized regular expression pattern matching to accomplish this. For example, the expression below matches the expect part of the file and then pushes it onto the array if the expect value is appropriate. It is important to note that the current line being read is stored in the variable *\$line*.

```
if($line =~ /. Expect = ([0-9-e\.]*)/ )
  if($1 < 1e-10) {
    push(@expect_list, $1);
    $expect[$index4] = @expect_list;
  }
```

The script iterates through the file and looks at key variables, such as the expect value, and then grabs information as long as the expect value is appropriate. This was not trivial to do, as the same supercontig could have multiple hits within it. This proved difficult to program, but was eventually overcome with the use of a complex array structure and by “looking ahead” in the file. The abstract view of the script is that we have an array with one entry for each of the variables that need to be extracted and for each of these values in the array, we have the actual data. Arrays were used because they can “grow” to accommodate multiple hits within a supercontig. These arrays are all indexed by the supercontig number. Examples of the output files are shown in Figures 3 and 4. Notice that supercontig AAGE01493862.1 had multiple hits.

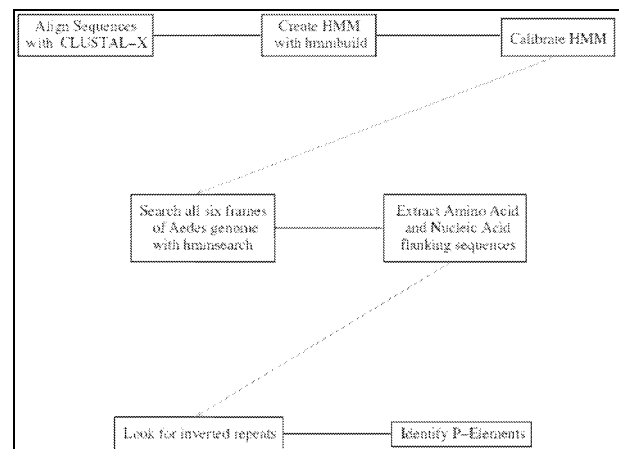


Figure 1. Block diagram of overall process.

```

ALIGNMENTS
>gi|59560372|gb|AAGE01616137.1| Aedes aegypti strain Liverpool ctg_1047197686048, whole genome
  shotgun sequence
  Length=777

Score = 480 bits (1236), Expect = 3e-135
Identities = 234/244 (95%), Positives = 234/244 (95%), Gaps = 10/244 (4%)
Frame = -1

Query 1  MSRWSLQYRLIWFGCKKN-----KAEWIVYNFILYIAYRYHSPAYEILRSEFGY 50
          MSRWSLQYRLIWFGCKKN          KAEWIVYNFILYIAYRYHSPAYEILRSEFGY
Sbjct 765 MSRWSLQYRLIWFGCKKNIKGTY**IHKKAEWIVYNFILYIAYRYHSPAYEILRSEFGY 586

Query 51  PLPCLRSLOEWASCIQMRKGVLEEVLTKIAGTRLSERERVTVLVYDEMSVESTIELDK 110
          PLPCLRSLOEWASCIQMRKGVLEEVLTKIAGTRLSERERVTVLVYDEMSVESTIELDK
Sbjct 585 PLPCLRSLOEWASCIQMRKGVLEEVLTKIAGTRLSERERVTVLVYDEMSVESTIELDK 406

Query 111 RNDVGLGPHSEMQLVMARGLFAQWKQPIFVDFDRQMTKPLEDLASRLHSIGFDVVATVH 170
          RNDVGLGPHSEMQLVMARGLFAQWKQPIFVDFDRQMTKPLEDLASRLHSIGFDVVATVH
Sbjct 405 RNDVGLGPHSEMQLVMARGLFAQWKQPIFVDFDRQMTKPLEDLASRLHSIGFDVVATVH 226

Query 171 DCGGNGMVWKEAGVNHETGQTSMKHPVSGNDIFFFPDAPHLKLFNRWLLDHGFLYK GK 230
          DCGGNGMVWKEAGVNHETGQTSMKHPVSGNDIFFFPDAPHLKLFNRWLLDHGFLYK GK
Sbjct 225 DCGGNGMVWKEAGVNHETGQTSMKHPVSGNDIFFFPDAPHLKLFNRWLLDHGFLYK GK 46

Query 231 SKFN 234
          SKFN
Sbjct 45 SKFN 34

```

Figure 2. NCBI Blast Results File Snippet

Name	Supercontig	Expect	Start	End	Frame
Potential Aedes	AAGE01616137.1	3e-135	765	34	-1
Potential Aedes	AAGE01140521.1	9e-123	1192	461	-3
Potential Aedes	AAGE01362873.1	5e-121	2	631	+2
Potential Aedes	AAGE01260454.1	6e-95	879	1427	+3
Potential Aedes	AAGE01493862.1	3e-79	574	939	+1
Potential Aedes	AAGE01493862.1	3e-79	342	564	+3

Figure 3. Sample Output

```

>AAGE01616137.1
MSRWSLQYRLIWFGCKKNIKGTY**IHKKAEWIVYNFILYIAYRYHSPAYEILRSEFGYPLPCLRSLOEWASCIQMRKGVLEEVLTKIAGTRLSERERVTVLVYD
EMSVESTIELDKRNDVGLGPHSEMQLVMARGLFAQWKQPIFVDFDRQMTKPLEDLASRLHSIGFDVVATVHDCGGGNGMVWKEAGVNHETGQTSMKHPVSGNDIFFF
PDAPHLKLFNRWLLDHGFLYK GKSKFN
>AAGE01140521.1
MSRWSLQYRLIWFGCKKNYQGNILINS*YKNYNNKAEWIVYNFILYIAYRYHSPAYEILRSEFGYPLPCL*SLQEWASCIQMRKGVLEEVLTKIAGTRLSERERV
TVLVYDEMSVESTIELDKLNDVIRPHSEMQLVMARGLFAQWKQPIFVDFDRQMTKPLEDLASRLHSIGFDVVATVHDCGGGNGMVWKEAGVNHETGQTSMKHPVSGNDIFFF
ETGQTSMKHPVSGNDIFFFPDAPHLKLFNRWFLDHGFLYK GKSKFN
>AAGE01362873.1
YNFILYIAYRYHSPAYEILRSEFGYPLPCLRSLOEWASCIQMRKGVLEEVLTKIAGTRLSERERVTVLVYDEMSVESTIELDKRNDVGLGPHSEMQLVMARGLFA
QWKQPIFVDFDRQMTKPLEDLASRLHSIGFDVVATVHDCGGGNGMVWKEAGVNHETGQTSMKHPVSGNDIFFFPDAPHLKLFNRWLLDHGFLYK GKSKFN
>AAGE01260454.1
MSRWSLQYRLIWFGCKKNIKGTY**IHKKAEWIVYNFILYIAYRYHSPAYEILRSEFGYPLPCLRSLOEWASCIQMRKGVLEEVLTKIAGTRLSERERVTVLVYD
EMSVESTIELDKRNDVGLGPHSEMQLVMARGLFAQWKQPIFVDFDRQMTKPLEDLASRLHSIGFDVVATVHDCGGGNGMVWKEAGVNHETGQTSMKHPVSGNDIFFF
PDAPHLKLFNRWLLDHGFLYK GKSKFN
>AAGE01493862.1
MTKKNDEVIGPHSEMQLVMARGLFAQWKQPIFADFVQVTKELLNDLITRLHNIGFSVMANVHDCGAGNRGVWRDCGVDHETKQTTMKHPVTGHDIFFFPDAPHLK
FRWLLDRGFLYK GK
>AAGE01493862.1
ITCRYHSCAYELLRTEFNPLPTLRLSLRNWASHINMRKGVLEDVLTLLQIAGTHMTEREKISVLVYDEMSVQSTTEYDKK

```

Figure 4. Sequence Sample Output

I. FUTURE WORK

Because the *hmmer* [7] suite of tools does not account for frame shifts and cannot run against all the frames at once, we would like to utilize a tool called *Genewise* to help identify additional *P* elements.

ACKNOWLEDGMENT

This project relied on the expertise of a Ph.D. biology graduate student, Jim Hogan, and the guidance of his advisor, Dr. Frank Collins. Contributions to this project were also made possible through our work cited below [6].

REFERENCES

- [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. O'Brotha, 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] Trevor Cickovski, Jim Hogan, and Ryan Kennedy, *Analysis of P element Transposon Sequences in Aedes aegypti*, Departments of Biological Sciences and Computer Science & Engineering, University of Notre Dame, December 2005. Available: http://www.cse.nd.edu/courses/cse60532/www/Student_Papers/CickovskiHoganKennedy.pdf
- [7] hmmer. <http://hmmer.wustl.edu>
- [8] Bioperl. <http://www.bioperl.org>