

# Hidden Markov Model Search of PiggyBAC Transposons: Insights and Analysis

J. Scott Breunig  
Dept of Biological Sciences

Philip Little  
Dept of Computer Science and Engineering

Tim Schoenharl  
Dept of Computer Science and Engineering

December 6, 2005

## Abstract

In this paper we describe our approach to tracing the evolution of several piggyBAC transposons, starting from a phylogenetic tree through building a Hidden Markov Model, finally arriving with a group of potentially related sequences that shows the conservation of this group of piggyBAC transposons across a wide range of organisms.

## 1 Introduction

Consider the phylogenetic tree of PiggyBac transposons in Figure 1 (taken from [7]). Given this tree, we are interested in finding related transposon sequences in currently sequenced genomes. The approach presented here is to build a Hidden Markov Model that can adequately represent the creation of the sequence, then use this Hidden Markov Model to search the NCBI GenBank database[4] for related sequences.

Using the phylogenetic tree of piggyBAC transposons from [7], we selected a subset of transposons that appeared to be closely related. Out of three candidate subsets, we chose the subset with the smallest inter-transposon distance that had at least 7 members.

## 2 Background

Multiple sequence alignment is done using the clustalw program[3]. Clustalw uses the Wunsch-Needleman algorithm [5] with a weighting factor that more accurately models the way DNA sequences evolve over time.

Hidden Markov Models for the search are built using the hmmer package[6]. The hmmer package takes a multiple sequence alignment and uses it to build a corresponding Hidden Markov Model.

The BLAST package [1] performs local sequence alignment on a pair of nucleotide sequences. Variations of BLAST have been created to allow sequence alignment on the resulting amino acid sequences[2].

NCBI [4] provides access to a database of known nucleotide sequences at its GenBank website, allowing researchers to compare the similarity of new sequences with archived data. This makes it possible to trace the evolution of genes throughout various organisms and can provide hints as to the function of newly discovered sequences.

## 3 Methods

Sarkar et al [7] used sequence alignments to construct a rooted phylogenetic tree of 50 piggyBAC, with weighted edges to common ancestors proportional to sequence differences. Using this tree, we

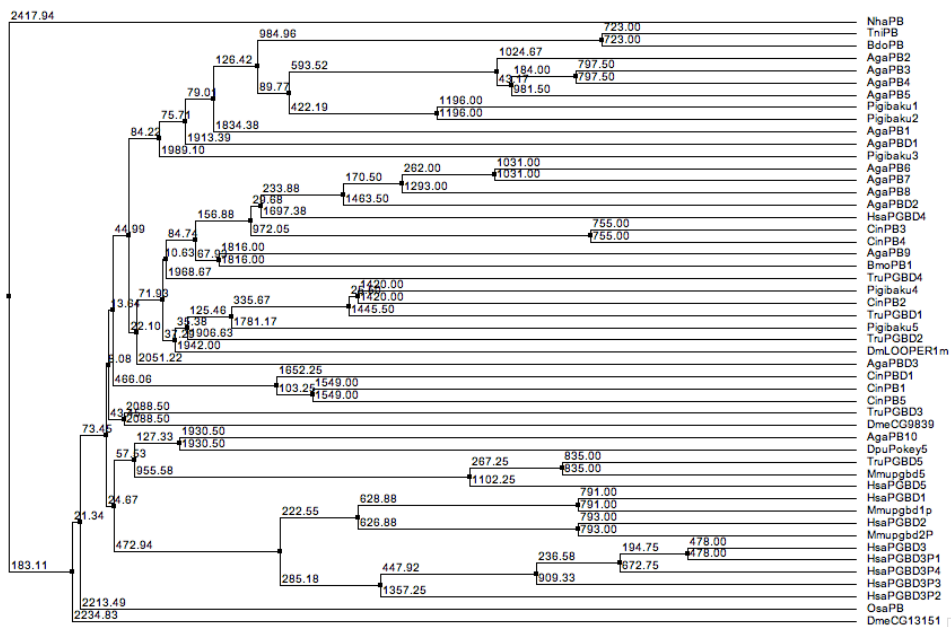


Figure 1: The phylogenetic tree of 50 PiggyBac transposons. The 3 candidate groups were hand picked using this tree.

used two methods to create sets of closely related sequences. Each set was used to build a Hidden Markov Model (HMM).

In the algorithm-based approach, we constructed a matrix, shown in Figures 2 and 3, detailing the distance between any two sequences and their common ancestor. Using dynamic programming we calculated the set of sequences with the least total difference by common ancestor. For each set of sequences, this was computed as the sum of distances between each pair of sequences in the set. Common groups of additions formed overlapping subproblems, while the optimal substructure consisted of values for smaller being combined to obtain the values for larger sets. (Overlapping subproblems and optimal substructure are two characteristics of a problem suitable for dynamic programming.) The algorithm was set to terminate after finding the best group of seven sequences. This turned out to have an average distance of 1453.9 between pairs in the set. (To realize the significance of this figure, consider that only 2.7% of all sequence pairs in the tree have a distance below this value. The set discovered by the algorithm has seven sequences, or 42 pairs, averaging this value.) The choice of seven as the number of sequences was intended to maintain the specificity of the HMM. Also, the optimal set of seven compared well to optimal sets of five, six, and eight sequences.

Our other method relied on assigning sequences into groups that seemed to fit into evolutionary groups. We required that each group contain a minimum number of sequences (6), useful for training the HMM to pick up more divergent sequences. In Group 1: HsaPGBD2, Mmupgbd2P, HsaPGBD3, HsaPGBD3P1, HsaPGBD3P4, HsaPGBDP3, HsaPGBDP2; in group 2: TniPB, BdoPB, AgaPB2, AgaPB3, AgaPB4, AgaPB5, Pigibaku1, Pigibaku2; in group 3: AgaPB6, AgaPB7, AgaPB8, AgaPBD2, HsaPGBD4, CinPB3, CinPB4.

In both methods, we then used hmmbuild and hmmscalibrate to generate an HMM based on these sequences. Since the genome of *Aedes aegypti* was provided in nucleotide form, while the transposons were available as peptide chains, it was necessary to create six reading frames of the genome to search against. Our HMM was used along these reading frames by the

program hmmsearch to identify stretches of sequence that were statistically comparable to those used to build the model. By observing the E value associated with each sequence, we isolated four sequences that met a threshold E value of less than  $1e^{-13}$ .

These sequences were subsequently run on the protein protein Basic Local Alignment Search Tool (BLASTp) [2] provided by NCBI [4]. This search provided homologous sequences across multiple genomes which can be seen in the results section. Finally, we used clustalw [3] to align the prominent hmmsearch results and analyze any differences between the sequences.

## 4 Results

The hmmsearch returned a table of possible matches against each reading frame. Anything with an appropriately small E-value (less than  $1e^{-14}$ ) was analyzed by running through the BLASTp program. Figure 4 shows the respective E-values from the hmmsearch as well as the most closely related sequence and its BLASTp E-value. Sequences were picked based on their low E-values relative to other hmmsearch sequence results in their group. The top BLASTp comparison for all four sequences is the same.

The top result for all BLASTp searches was the predicted region of the ZNF261 protein in zebrafish *Danio rerio*. An ExPASy search reveals no known domains in this protein (UniProtKB/TrEMBL entry Q5BJB2). Many other transposable elements share sequence similarity with these sequences, most notably PGBD3 (UniProtKB/TrEMBL entry Q5W0M0) as well as numerous piggyBAC transposable elements in humans (*Homo sapiens*).

The results of the multiple alignment of the two highest scoring sequences for each approach is shown in Figure 5. A multiple alignment of the four sequences show great similarity (scores > 99) between the sequences despite their different reading frames. Two amino substitutions are present between the reading frames, though similarity is retained between the groups.



```

TruBGRF PigibakuSTruBGRF Dmi GNP AgaPB03 CinPBD1 CinPB1 CinPB5 TruBGRF DmeCG9 AgaPB10 DnuDok TruBGRF Mmuinh HeaBGRF HeaBGRF Mmuinh HeaBGRF Mmuinh HeaBGRF HsaPGBD3P1HeaBGRF HeaBGRF HeaBGRF OsaPB DmeCG13151

1781.17
1906.63
1942 1942 1942
2051.22 2051.22 2051.22 2051.22
2118.31 2118.31 2118.31 2118.31 2118.31
2118.31 2118.31 2118.31 2118.31 2118.31 1652.25
2118.31 2118.31 2118.31 2118.31 2118.31 1652.25 1549
2131.95 2131.95 2131.95 2131.95 2131.95 2131.95 2131.95
2131.95 2131.95 2131.95 2131.95 2131.95 2131.95 2131.95 2088.5
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 1930.5
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2057.83 2057.83
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2057.83 2057.83 835
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2057.83 2057.83 1102.25 1102.25
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 2115.36
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 791
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 2115.36 1419.88 1419.88
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 2115.36 1419.88 1419.88 793
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 2115.36 1642.43 1642.43 1642.43 1642.43
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 2115.36 1642.43 1642.43 1642.43 1642.43 478
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 2115.36 1642.43 1642.43 1642.43 672.75
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 2115.36 1642.43 1642.43 1642.43 909.33
2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2140.03 2115.36 2115.36 2115.36 2115.36 2115.36 1642.43 1642.43 1642.43 1357.25
2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 2213.48 1357.25 1357.25 1357.25
2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82 2234.82

```

Figure 3: Computed Distance between two sequences, continued.

Group	Hmmsearch E-value	BLASTp E-value of best sequence	Sequence description from BLASTp search
Hand-grouped Reading Frame 2	3 e -16	3 e -80	Predicted: similar to ZNF261 protein in zebrafish ( <i>Danio rerio</i> )
Hand-grouped RF 4	4.8 e -15	2 e -80	Predicted: similar to ZNF261 protein in zebrafish ( <i>Danio rerio</i> )
Algorithm-grouped RF 2	6.4 e -19	3 e -80	Predicted: similar to ZNF261 protein in zebrafish ( <i>Danio rerio</i> )
Algorithm-grouped RF 4	1 e -18	2 e -80	Predicted: similar to ZNF261 protein in zebrafish ( <i>Danio rerio</i> )

Figure 4: **hmmsearch and BLASTp results of selected sequences.** Sequences were picked based on their low E-values relative to other hmmsearch sequence results in their group. The top BLASTp comparison for all four sequences is the same

CLUSTAL W (1.82) multiple sequence alignment

```

Han2      --LIYYESIIYIMIHFTD TDDL LLLALTFEGDVSEVEGFDVSDDEDEIEAFHANTESPKC 58
Han4      --LIYYESIIYIMIHFTD TDDL LLLALTFEGDVSEVEGFDVSDDEDEIEAFHANTESPKC 58
Alg2      QKLIYYESIIYIMIHFTD TDDL LLLALTFEGDVSEVEGFDVSDDEDEIEAFHANTESPKC 60
Alg4      QKLIYYESIIYIMIHFTD TDDL LLLALTFEGDVSEVEGFDVSDDEDEIEAFHANTESPKC 60
          . * * * * *

Han2      VKPQK DTPVKSGTKVRRNQNKTSAKQAASKQGGKKGVSTRSKSKKGTTPPKWHEDESVC 118
Han4      VKPQK DTPVKSGTKVRRNQNKTSAKQAASKQGGKKGVSTRSKSKKGTTPPKWHEDESVC 118
Alg2      VKPQK DTPVKSGTKVRRNQNKTSAKQAASKQGGKKGVSTRSKSKKGTTPPKWHEDESVC 120
Alg4      VKPQK DTPVKSGTKVRRNQNKTSAKQAASKQGGKKGVSTRSKSKKGTTPPKWHEDESVC 120
          * * * * *

Han2      FECIEPVVKS SKLPTKPMTPYQYFGLFIDSEIISDICEINLYSSQODETPIDVTSGDIEQ 178
Han4      FECIEPVVKS SKLPTKPMTPYQYFGLFIDSEIISDICEINLYSSQODETPIDVTSGDIEQ 178
Alg2      FECIEPVVKS SKLPTKPMTPYQYFGLFIDSEIISDICEINLYSSQODETPIDVTSGDIEQ 180
Alg4      FECIEPVVKS SKLPTKPMTPYQYFGLFIDSEIISDICEINLYSSQODETPIDVTSGDIEQ 180
          * * * * *

Han2      HIGQLLLMGVTKVPSYRLHWGSTTRYAPIADVMPRNKFETIKRCLHFNDNIKIKQRDEEG 238
Han4      HIGQLLLMGVTKVPSYRLHWGSTTRYAPIADVMPRNKFETIKRCLHFNDNIKIKQRDEEG 238
Alg2      HIGQLLLMGVTKVPSYRLHWGSTTRYAPIADVMPRNKFETIKRCLHFNDNIKIKQRDEEG 240
Alg4      HIGQLLLMGVTKVPSYRLHWGSTTRYAPIADVMPRNKFETIKRCLHFNDNIKIKQRDEEG 240
          * * * * *

Han2      YDKLFKVRPFIDALRRNYLKI ETPNQSIDEIMIPSKAASPLRQYNKNKPHRFGIKVEGR 298
Han4      YDKLFKVRPFIDALRRNYLKI ETPNQSIDEIMIPSKAASPLRQYNKNKPHRFGIKVEGR 298
Alg2      YDKLFKVRPFIDALRRNYLKI ETPNQSIDEIMIPSKAASPLRQYNKNKPHRFGIKVEGR 300
Alg4      YDKLFKVRPFIDALRRNYLKI ETPNQSIDEIMIPSKAASPLRQYNKNKPHRFGIKVEGR 300
          * * * * *

Han2      ASSDGI LHDFS IYGGKTNKEPSGAGISGDVVIKLDITLSONVPYRIFADNWFSYALVKEM 358
Han4      ASSDGI LHDFS IYGGKTNKEPSGAGISGDVVIKLDITLSONVPYRIFADNWFSYALVKEM 358
Alg2      ASSDGI LHDFS IYGGKTNKEPSGAGISGDVVIKLDITLSONVPYRIFADNWFSYALVKEM 360
Alg4      ASSDGI LHDFS IYGGKTNKEPSGAGISGDVVIKLDITLSONVPYRIFADNWFSYALVKEM 360
          * * * * *

Han2      KSRGLE YTGTVRQNRIPGFEMKKNLKA EGRGSFACNVSNDFV VVTWMDNKPINLISSCF 418
Han4      KSRGLE YTGTVRQNRIPGFEMKKNLKA EGRGSFACNVSNDFV VVTWMDNKPINLISSCF 418
Alg2      KSRGLE YTGTVRQNRIPGFEMKKNLKA EGRGSFACNVSNDFV VVTWMDNKPINLISSCF 420
Alg4      KSRGLE YTGTVRQNRIPGFEMKKNLKA EGRGSFACNVSNDFV VVTWMDNKPINLISSCF 420
          * * * * *

Han2      GVQPIDEVKRWSVSDKIYKSI PRPLVVREYNCYMGGIDLNDFLVALYRTQPGTKKYYMRI 478
Han4      GVQPIDEVKRWSVSDKIYKSI PRPLVVREYNCYMGGIDLNDFLVALYRTQPGTKKYYMRI 478
Alg2      GVQPIDEVKRWSVSDKIYKSI PRPLVVREYNCYMGGIDLNDFLVALYRTQPGTKKYYMRI 480
Alg4      GVQPIDEVKRWSVSDKIYKSI PRPLVVREYNCYMGGIDLNDFLVALYRTQPGTKKYYMRI 480
          * * * * *

Han2      FYHLLDVSVVNAWLLYRRHMKQTGQEHMTLLNFRIDVANS LIYQGKIVRRRGRPTNQPI 538
Han4      FYHLLDVSVVNAWLLYRRHMKQTGQEHMTLLNFRIDVANS LIYQGKIVRRRGRPTNQPI 538
Alg2      FYHLLDVSVVNAWLLYRRHMKQTGQEHMTLLNFRIDVANS LIYQGKIVRRRGRPTNQPI 540
Alg4      FYHLLDVSVVNAWLLYRRHMKQTGQEHMTLLNFRIDVANS LIYQGKIVRRRGRPTNQPI 540
          * * * * *

Han2      QKRQRVVASVGPSLEARYDGLEHWPVEDRRERCVLCKDRGSFNSFKCTKCDVCLCIKKGK 598
Han4      QKRQRVVASVGPSLEARYDGLEHWPVEDRRERCVLCKDRGSFNSFKCTKCDVCLCIKKGK 598
Alg2      QKRQRVVASVGPSLEARYDGLEHWPVEDRRERCVLCKDRGSFNSFKCTKCDVCLCIKKGK 593
Alg4      QKRQRVVASVGPSLEARYDGLEHWPVEDRRERCVLCKDRGSFNSFKCTKCDVCLCIKKGK 593
          * * * * *

Han2      NC 600
Han4      NC 600
Alg2      --
Alg4      --

```

Figure 5: **Multiple Sequence Alignment of the 2 highest scoring sequences discovered by each approach.** Complete alignment between the Algorithm-assigned (Alg) and the Hand-assigned (Han) groups occurs with small exceptions at the beginning and end of the sequences, and at two spots within the sequence. An amino acid switch from aspartic acid to tyrosine at position 149 (Alg), 147 (Han); from isoleucine to threonine at position 437 (Alg), 435 (Han). In both cases, the difference lies between the reading frames.

## 5 Discussion

The fact that the algorithm-picked group and the hand-picked group resulted in the same gene sequence for analysis, ZNF261, does not mean that the two approaches are equal. Rather, as seen in Figure 4 E-value columns, the algorithm-based approach yields a higher degree of certainty in its results. A lower E-value in the hmmsearch column is representative of a nonrandom sequence conforming to the same parameters as the build conditions. Thus, a  $10^2$  to  $10^4$  difference in E-values for sequences between the two different methods suggests that our algorithm-based approach provides a stronger confidence in identifying corresponding regions.

This transposons similarity to other species versions of piggyBAC is not surprising since multiple species were found to be homologous by Sarkar et al [7]. The presence of such a near-ubiquitous transposon that is so useful for genetic experimentation in various species offers a chance that similar projects may be run on higher order model organisms.

Also worthy of notice is the presence of two nearly identical sequences in separate reading frames in the genome. This demonstrates the flux of the piggyBAC transposon, it has replicated itself almost exactly in a different part of the genome.

The mutations between the reading frames bring significant differences. These mutations seem minor due to their infrequency yet the difference between the alternates could cause a massive upheaval in the structure of the protein and its subsequent function. The amino acid isoleucine is hydrophobic amino acid, with rigid structure, found in the core of its protein. The alternate amino acid at that position for the fourth reading frame is threonine, a hydrophilic, hydrogen-donating peptide which is found on the exterior of its protein. This change may expose different domains of the protein and significantly alter function. Similarly, the earlier difference between the hydrophobic tyrosine which is normally oriented inwards to aspartic acid, one of the most hydrophilic amino acids that is known for bonding positively-charged molecules and ions on the proteins exterior. Tyrosines aromatic side group is often responsible for forming a strong, stable core unit for the protein, so

removing it alone will change the tertiary structure, much less changing it to a strongly hydrophilic element.

To determine the extent of such changes, we suggest running a Northern blot for both sequences in an attempt to determine if either sequence is transcribed. A Western blot for each version would be a logical following step if mRNA had been found that matched this sequence. By measuring levels of each type of protein, it might be possible to determine which version is useful towards cellular activities. An antibody block would then provide the starting point towards determining function.

## References

- [1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- [2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.
- [3] Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G., and Gibson T.J. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.
- [4] National center for biotechnology information - genbank. <http://www.ncbi.nlm.nih.gov/BLAST/>.
- [5] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.
- [6] A. Krogh R. Durbin, S. Eddy and G. Mitchison. *Biological sequence analysis: probabilistic models*

*of proteins and nucleic acids*. Cambridge University Press, 1998.

- [7] A. Sarkar, C. Sim, Y. S. Hong, J. R. Hogan, M. J. Fraser, H. M. Robertson, and F. H. Collins. Molecular evolutionary analysis of the widespread piggybac transposon family and related "domesticated" sequences. *Molecular Genetics and Genomics*, pages 173–180, 2003.