“Spidey” sequences: Finding transposable elements in the genome sequence of the common house spider, *Parasteatoda tepidariorum*

Renée J. Estephan, Biology

Mentor: Dr. Peter Arensburger

Kellogg Honors College Capstone Project

Abstract

The genome of the common house spider, *Parasteatoda tepidariorum*, has recently become available and little is yet known about the function of individual parts of the genome. The goal of this project is to search for DNA patterns, specifically those of transposable elements, in the genome sequence of the common house spider. Transposable elements are sequences of DNA that have the potential to facilitate their own movement from one genome location to another.

We used a bioinformatics program called RepeatModeler to identify possible transposable elements in the spider genome. Results obtained by running RepeatModeler were in the form of computer files that describe the locations of the transposable elements on the spider genome. One of the flaws of RepeatModeler is that it produces extraneous results that may not actually be transposable elements (i.e. false positives). Therefore, we reanalyzed the results from RepeatModeler to see which sequences were likely real transposable elements and which were not. Much is left to learn about the features of the spider genome. Utilizing results from this project and related projects can lead to practical applications of spider silk in the area of armory development and defense.

Introduction

Transposable elements (TEs) are sequences of DNA that have the potential to facilitate their own movement from one genome location to another. Discovered in the late 1940’s by Barbara McClintock, TE’s contribute about 50% to the size of mammalian genomes (Thornburg et al 2006). TE’s have been found to play a role in regulating gene expression and generating different cell types and structures based on their location in the genome (Britten and Davidson 1969). The large quantity of TE’s in genomes and the role they play in regulation make them a useful tool in learning more about an organism.

The genome of the common house spider, *Parasteatoda tepidariorum,* recently became publicly available (2013) and little remains known about the composition of the genome. Therefore, we decided to look at the genomic structure, specifically TEs, of this organism. Studying the TEs of *P. tepidariorum* could provide insight into the function of various spider genes including the production of spider silk and venom. *P. tepidariorum* is a good model for genome research since it is easy to culture and genetic material can be easily extracted.

Methods

The genome of *P. tepidariorum* was downloaded to the biological sciences department computer cluster (CPPBC) which uses a Linux (Redhat) operating system. Next, the RepeatModeler program was installed on the computer cluster (Smit, A F A 2008). RepeatModeler is a program which identifies repeats of nucleotide sequences in a genome. Before installing RepeatModeler, certain prerequisites had to be met, such as having support programs such as Perl, RepeatMasker, RECON, RepeatScout, and Tandem Repeat Finder installed on the computer cluster beforehand.

Following this installation, the genome of *P. tepidariorum* was run through the RepeatModeler pipeline to identify initial repeats in the genome. The resulting sequences were reviewed using the Seaview program. Any sequences greater than 750 base pairs were kept for further analysis. The narrowed down results were then compared against a reference list of TEs from various invertebrates using the BLAST program. The reference file was obtained from the RepBase database (Jurka et al. 2005). The reference database was converted from a Fasta format to a format usable by BLAST . The comparison of the paired-down RepeatModeler data against the reference TE database was run on the computer cluster using multiple cores to save execution time. The result was an output file that was used to narrow down the results even further based on E-values. E-values, or expectation values, tell how many times we could have expected a result just by chance given a database and the alignment score of a match. The closer the E-value is to zero, the less likely the match was considered to be due to chance.

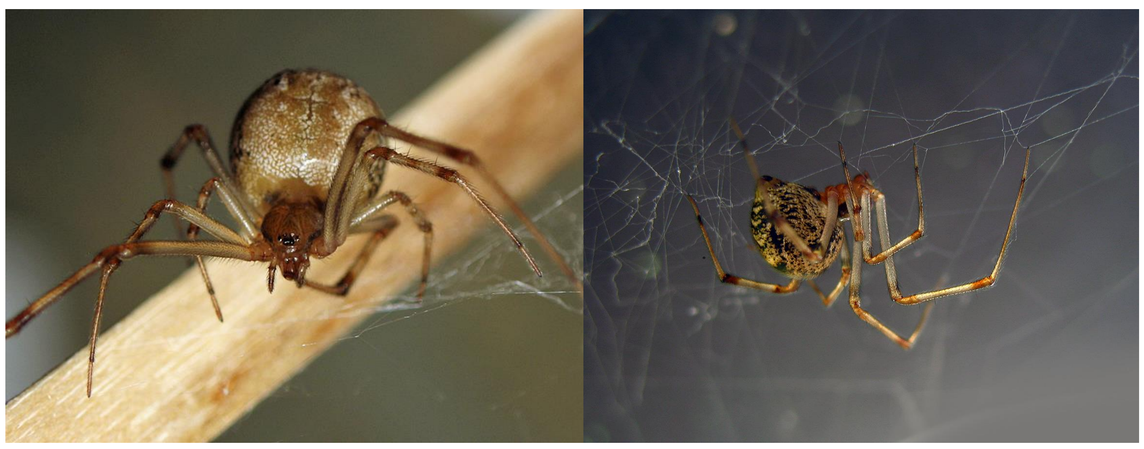

Fig. 1. Images of common house spider, *Parasteatoda tepidariorum*, spinning a web of silk.

Photo credit: http://www.spiders.us/image/parasteatoda-tepidariorum-23/
Photo credit: https://en.wikipedia.org/wiki/Parasteatoda_tepidariorum#/media/File:Achaearanea_tepidariorum.jpg

Results

There were 27 families of repeated sequences found in the *P. tepidariorum* genome that fit our criteria of being possible TEs. A table containing all TEs from the spider genome (ex. rnd-#_family-###), E-values (less than 1), and TEs from organisms in the reference database was constructed. Below is a subset of the TEs found in the spider genome.

Possible Element Name (Family)	Significant Similarities (Repbase Databse Query)	E-value
rnd-5_family-94	CR1-3_DK	0.000000002
rnd-5_family-76	EnSpm-N2_RPr	5.00E-62
rnd-4_family-721	Gypsy19-LTR_Dpse	0.97
rnd-5_family-96	Copia-13_Dpu-I	1.00E-58
rnd-5_family-3506	Mariner-10_AEc	4.00E-88
rnd-5_family-127	Jockey-2_CQ	3.00E-76
rnd-5_family-3506	Lm2C1	0.63
rnd-5_family-94	L2-1_SK	3.00E-16
rnd-5_family-96	Copia-8_CQ-I	2.00E-46
rnd-6_family-3846	Gypsy-90_CQ-I	3.00E-36

Discussion

After looking at the output file, we saw that some transposable elements were likely present in the spider genome. However, there were some TE’s that had relatively high E-values compared to the majority of the other TE’s. Therefore, the likelihood of those TE’s with high E-values (ex. 0.63) actually being TE’s was relatively low. There were several queries that had multiple hits with various families. For example, the query Mariner-10_Aec had 17 hits with 17 TE’s found in the spider genome. It was interesting to find that TE’s in the spider genome are similar in sequence to a variety of genomes in the reference database rather than to a single genome. This further supports the idea that TE’s make up a large part of various genomes and may play an important role in gene regulation. By annotating the newly sequenced genome of the common house spider, we can discover TE’s that may play a role in the regulation of genes that code for the production of spider silk and venom. Learning more about this genome can reveal gene evolution or novel silk and venom genes, which has many applications for defense. The silk of the common house spider is sturdy and tough, which allows for the capture of prey and can play a role in the biomaterial application of silk (Sanggaard et al. 2013). Future directions of this project include finishing annotating the common house spider genome and compiling a complete list of TE’s in the genome that are also found in organisms in the reference database. These TE’s may be added onto an online genome browser for the common house spider, which will provide information to researchers looking at silk or venom genes.

References

Britten, R. J., & Davidson, E. H. Gene regulation for higher cells: A theory. *Science* **165**, 349–357 (1969).

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. “Repbase Update, a Database of Eukaryotic Repetitive Elements.” *Cytogenetic and Genome Research* 110 (1-4): 462–67. doi:10.1159/000084979.

Sanggaard, K.W. et al. Spider genomes provide insight into composition and evolution of venom and silk. *Nature* **5**, 1-12 (2014).

Smit, A F A, Hubley, R. 2008. “RepeatModeler Open-1.0.” <http://www.repeatmasker.org>.

Thornburg, G.B. et al. Transposable elements as a significant source of transcription regulating signals. *Gene* **365**, 104-110 (2006).

Acknowledgements

I would like to thank my mentor, Dr. Peter Arensburger, for his guidance and patience throughout the course of this project. I would also like to thank German Lagunas-Robles and Mark Ellie Alonzo for their assistance and willingness to answer questions pertaining to my project. Thank you to the Kellogg Honors College and the Achieve Scholars Program for their generous funding for materials and supplies.