Computational approaches to identify DNA motifs for genes expressed in the mushroom body brain structure of *Drosophila melanogaster*

Raymond Anthony Ford,[1*] Carolina Guerra,[2*] Ming-Ying Leung,[1,2] Kyung-An Han[3]

[1]Department of Mathematical Sciences, [2]Bioinformatics Program, [3]Department of Biological Sciences,
The University of Texas at El Paso
*These authors equally contributed to this study

ABSTRACT

Through the use of web scraping scripts we obtained data of interest from the HHMI Janelia Farm Research Campus and FlyBase to construct 6,931 FASTA files containing the DNA sequence data for *D. melanogaster*. In an effort to enable accessing the data more efficiently, we developed a Tcl/Tk based GUI application (FlyTrap). The combination of web scraping and FlyTrap spared us the burden of having to visit approximately 21,000 web sites. The Multiple Em Motif Elicitation (MEME) tool was used to identify conserved sequence motifs in *D. melanogaster* DNA that may allow selective expression of genes in the mushroom body gamma lobe, the brain structure responsible for learning and memory. MEME returned the 100 motifs in the DNA segments driving gamma lobe expression, but only seven were selected based on an E-value less than 10. This information can be used to identify the enhancer elements for the genes that are expressed in the neurons under investigation facilitating the study of the mechanisms underlying learning and memory. With these results, we can further identify additional genes expressed in the gamma neurons, which mediate dopamine and octopamine signals. Abnormal dopamine functions are responsible for various neuropsychiatric disorders including ADHD, autism, schizophrenia, Parkinson's disease, and drug abuse/addiction to name a few. This study would ultimately help understand the underlying pathological mechanisms.

AIM AND SIGNIFICANCE

- Help in the study of the pathogenesis mechanisms underlying learning and memory impairments by identification of enhancer elements.
- Find DNA motifs in the gamma neurons critical to the study of neural transmissions which have a potential effect in certain neuropsychiatric disorders.
- Develop open source software applications that will allow researchers to access data related to the *D. melanogaster* more easily.

INTRODUCTION

- *D. melanogaster* is a powerful model organism due to its well characterized genetics, fully-sequenced genome, and sophisticated nervous system comparable to humans.
- The study of mushroom bodies in the insect brain is important to understand learning, memory, and other brain functions. Also, the mushroom bodies are a widely studied and well-defined structure.
- The mushroom bodies in the *D. melanogaster* are composed of three distinct structures, namely alpha, beta and gamma lobes. This project focuses on gamma neurons due to their important function of mediating dopamine and octopamine signals responsible for learning and memory.
- The understanding of such mechanisms can lead to uncovering important pathways associated with abnormal dopamine function, which is responsible for diseases such as schizophrenia, Parkinson's disease, and drug abuse/addiction.

MATERIALS AND METHODS

Figure 1: Gamma lobe expression of stock number 46537. Source: http://flyweb.janelia.org

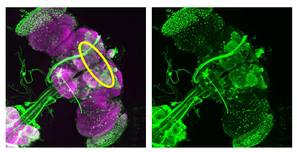


Figure 2: Bloomington *Drosophila* Stock Center at Indiana University. Source: http://flystocks.bio.indiana.edu

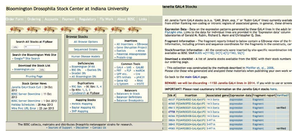
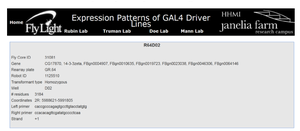


Figure 3: Expression data table for stock number 46537. Source: http://flyweb.janelia.org



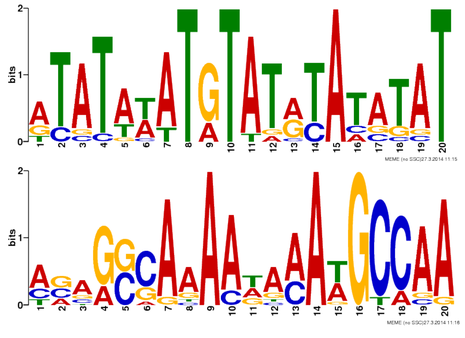
1) The expression data videos were surveyed for the first 600 fly lines out of the 7,060 total Janelia Farm (JF) fly lines listed in the Bloomington Drosophila Stock Center (BDSC) website to determine if gamma lobe expression (demarcated by the yellow circle in Figure 1) was present. We selected the lines with gamma lobe expression to be part of our training set for motif discovery.
2) The Python programming language was used, along with several third party modules, to take information from three different websites and construct FASTA files containing the DNA sequence data for all lines with sequence data accessible. The DNA sequence location in the genome was taken from FlyLight.
3) Using the FASTA files from (2) and information obtained in (1), we constructed a plain text file containing the sequences of the BDSC stock numbers showing gamma lobe expression.
4) The National Biomedical Computation Resource's *Multiple Em for Motif Elicitation* (MEME) tool was used to identify motifs common to DNA sequence (the text file in 3) of the selected lines in (1). Motifs with an E-value less than 10 were considered for further analyses.

RESULTS

Table 1: The 13 BDSC stock numbers showing gamma lobe expression, along with their DNA sequence location in the genome

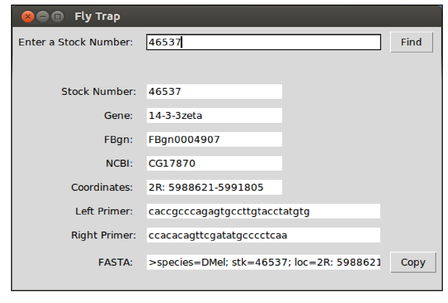
Stock#	Fragment	Gene	Gene location
46537	R65G03	14-3-3zeta	2R: 5988621-5991805
47708	R65G03	14-3-3zeta	2R: 5992501-5993251
50455	R53H09	5-HT1A	2R: 14971817-14975677
39511	R70B01	5-HT7	3R: 26830224-26834165
39900	R75F05	ab	2L: 11227992-11231845
45228	R35B04	Abd-B	3R: 12786057-12789039
39338	R65B07	Appl	X: 428701-431207
46580	R67C11	star1	3L: 18559334-18562946
48463	R11E10	AlCR2	3L: 18502914-18506693
39166	R56H09	Atf3	X: 1142016-1145229
46536	R64D01	Atpalpha	3R: 16780321-16783496
46565	R65H05	Atpalpha	3R: 16783817-16785018
47614	R49C07	beat-IIIc	2L: 17243021-17246573

Figure 4: Two of the seven motifs discovered by MEME. Source: http://meme.ncbr.net/



1) From the 600 JF lines surveyed, we identified 13 fly lines showing gamma lobe expression (Table 1).
2) FASTA files for individual fly lines were created.
3) MEME on the DNA sequences of the selected 13 lines returned 100 motifs where only seven had an E-value less than 10 (Figure 4).
4) To help access the FASTA files, along with other data related to *D. melanogaster*, a Tcl/Tk based GUI application was developed and named FlyTrap. This application facilitates the access of DNA sequence and associated information without having to visit numerous websites (Figure 5).

Figure 5: Screenshot of FlyTrap being used to find information related to BDSC stock number 46537



CONCLUSIONS

- From this study, we found 13 fly lines showing gamma lobe expression out of 600 fly lines surveyed.
- The FlyTrap GUI application made the accessing the data easier to perform the MEME analysis.
- With the known motifs it would be feasible to identify specific binding sites for proteins such as transcription activators and repressors, splicing machinery, and chromatin remodeling components responsible for tissue-specific expression.

FUTURE WORK

- Using the discovered motifs, search all 6,931 FASTA files to determine whether each motif is present in additional fly lines.
- Examine the expression patterns of the additional lines containing the motif under study to substantiate it as a key enhancer element for gamma lobe expression.
- Search for the discovered motifs in other *Drosophila* species to explore whether they are conserved across species.
- Make the source code of Fly Trap, along with all the other data files that it needs to run, publicly available under a GPLv3 license.

REFERENCES

Bailey TL, Elkan C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems Biology*. 28-36.

Bloomington Drosophila Stock Center. (2014). Janelia GAL4 stocks.

Paia TP, Chena CC, Lina HH. (2013). Drosophila ORB protein in two mushroom body output neurons is necessary for long-term memory formation. *PNAS* 110(19):7898-7903.

Pfeiffer BD, Jenett A, Hammonds AS. (2008). Tools for neuroanatomy and neurogenetics in Drosophila. *PNAS* 105(28):9715-9720.

Pfeiffer BD, Jenett A, Hammonds AS, Ngo TT, Misra S. GAL4 driver collection of Rubin Laboratory at Janelia Farm.

St. Pierre SE, Ponting L, Stefancsik R, McQuilton P, the FlyBase Consortium. (2014). FlyBase 102 — advanced approaches to interrogating FlyBase. *Nucleic Acids Research* 42(D1):D780-88.

Tanaka NK, Tanimoto H, Ito K. (2008). Neuronal assemblies of the Drosophila mushroom body. *Journal of Comparative Neurology* 508(5):711-55.

ACKNOWLEDGEMENTS

This work is supported in part by the NSF grant DUE-0926721 and NIH-NIMHD grant 5G12MD007592. C.G. is also supported by a scholarship from the Chihuahua government (Secretaria de Educación, Deporte y Cultura), Chih. Chih. México 2014.