A bioinformatics approach to identify potential enhancer elements for genes expressed in the mushroom body neurons in *Drosophila*

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

Bioinformatics is an interdisciplinary science that focuses on the management and interpretation of data obtained from complex biological phenomena using mathematical models, computational algorithms, and statistical methodologies. In this project, we combine several existing bioinformatics software for identifying DNA motifs with other computational tools to develop a new approach for finding potential enhancer elements of a set of genes related to learning and memory in *Drosophila.* 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 efficiently, we developed a Tcl/Tk based GUI application (FlyTrap). Using six different DNA motif discovery software, 18 conserved motifs were identified among the DNA sequences driving mushroom body expression. The FASTA files for all Janelia Farm lines were computationally searched to determine if any of the discovered motifs were present in these sequence data. This information can be used to identify potential enhancer elements for the genes expressed in the mushroom body gamma neurons, which are responsible for learning and memory. With these results, we can further identify additional genes expressed in the gamma neurons, which mediate dopamine and octopamine signals that are involved in neuropsychiatric disorders including ADHD, autism, schizophrenia, Parkinson's disease, and drug abuse/addiction. 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.

INTRODUCTION

- *D. melanogaster* is a powerful model organism due to its well characterized genetics, fully-sequenced genome, and sophisticated nervous system comparable to humans.

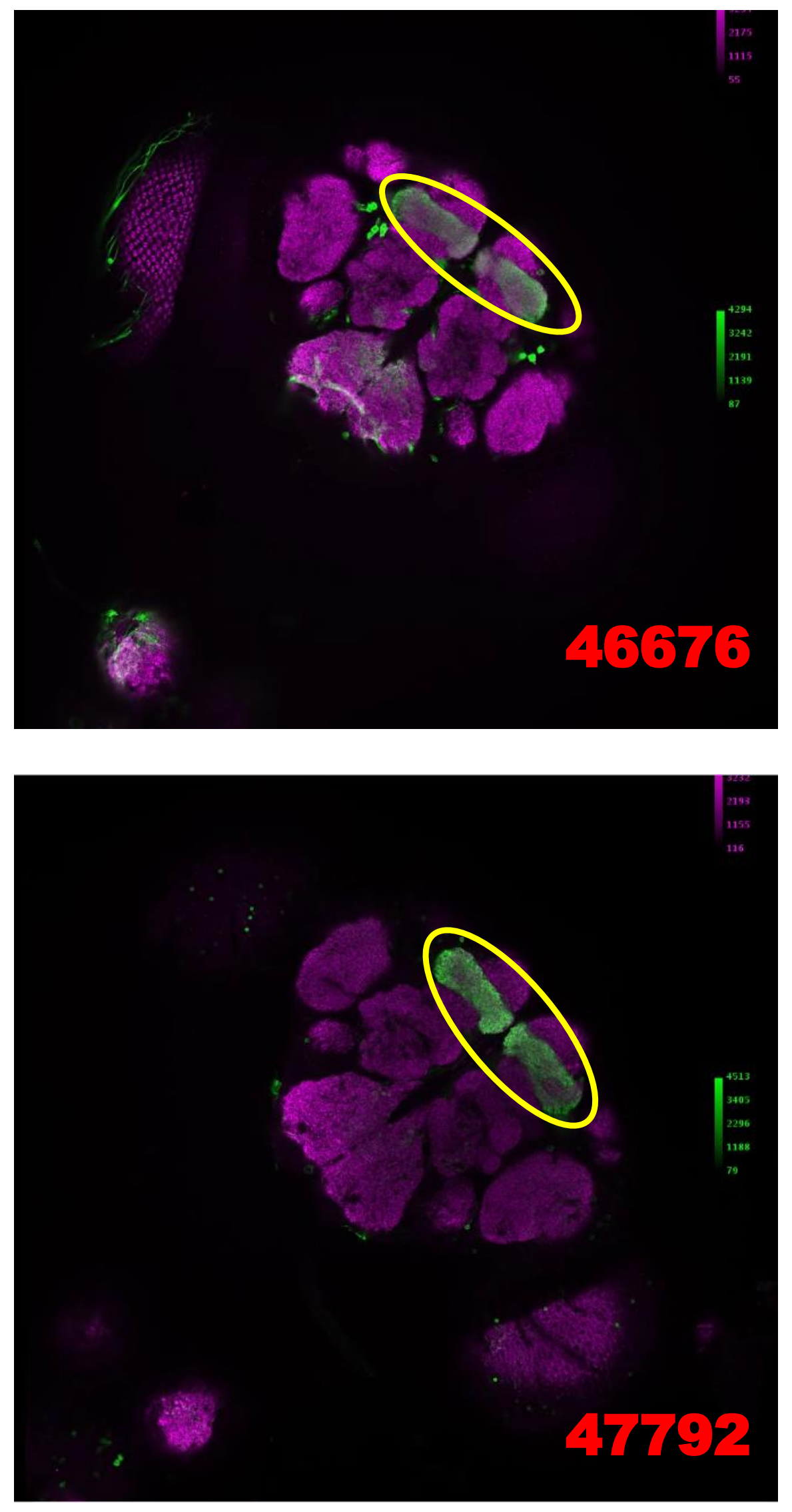
- The study of mushroom bodies in the insect brain is important to understand learning, memory, and other brain functions. Also, the mushroom bodies are a widely studied and well-defined structure.

- 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 numbers 46676 and 47792. Source: http://flyweb.janelia.org



HHMI janelia farm research campus

1) The expression data videos were surveyed for all 1043 fly lines having a DNA sequence length less than 1000 base pairs to determine if gamma lobe expression (demarcated by the yellow ovals 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 HHMI Janelia Farm Research Campus.

3) Using the FASTA files from (2) and information obtained in (1), we constructed a plain text file containing the sequences of the Bloomington Drosophila Stock Center (BDSC) stock numbers showing gamma lobe expression.

4) The file from (3) was then submitted to YMF, MEME, and Melina-II (using the Gibbs MotifSampler, Weeder, Consensus, and MDscan discovery tools). The top five scoring motifs from each tool were considered for further analysis.

5) The FASTA files from (2) were then computationally searched to determine which motifs were present in a particular fly line's DNA sequence data.

RESULTS

Table 1: The 14 discovered fly lines showing gamma expression.

Stock #	Fragment	Gene	Location	Sequence Length (bp)
48001	R93E08	unc-5	2R: 11236857-11237190	333
46896	R89G05	Dscam3	3R: 13309632-13310008	376
48588	R13H03	Rab3	2R: 6611847-6612338	491
47792	R84A09	Deaf1	3L: 19811695-19812273	578
38832	R52E05	Glu-RIB	3L: 9251270-9251949	679
47708	R65G03	14-3-3zeta	2R: 5992501-5993251	750
40019	R78H10	sba	3R: 19721188-19721959	771
45998	R50F10	unc-13	4: 914152-914962	810
49218	R27C11	Fmr1	3R: 5934716-5935578	862
49656	R31A04	sol	X: 21221828-21222705	877
46974	R76H09	stj	2R: 9694356-9695282	926
45256	R39A09	Hr39	2L: 21237362-21238338	976
45258	R39A12	CG12071	3R: 26705583-26706568	985
45051	R14A04	RyR	2R: 4753537-4754533	996

Table 2: The 18 discovered DNA motifs organized by discovery tool.

Tool	Motif	Identifier
YMF	ACACACAC	Motif$_{1,1}$
	GAGRGRGA	Motif$_{1,2}$
	CCTCCTCC	Motif$_{1,3}$
	GGGGTGGY	Motif$_{1,4}$
	GGAAAASS	Motif$_{1,5}$
MEME	CWCACACACACA	Motif$_{1,6}$
	RCRGCMGCRAM	Motif$_{1,7}$
	RGARAGRKASA	Motif$_{1,8}$
	SVTTTTCC	Motif$_{1,9}$
	AGKGGGTGGMR	Motif$_{1,10}$
Gibbs	ACACACACACAS	Motif$_{1,11}$
CONSENSUS	CAYACACACACA	Motif$_{1,12}$
Weeder	CMCMCACA	Motif$_{1,13}$
	TGTGKGTRWG	Motif$_{1,14}$
	TCCGTTTTYCTC	Motif$_{1,15}$
	CCCCAC	Motif$_{1,16}$
	CRTTTTCCTC	Motif$_{1,17}$
MDscan	TGTGTGTGTGTG	Motif$_{1,18}$

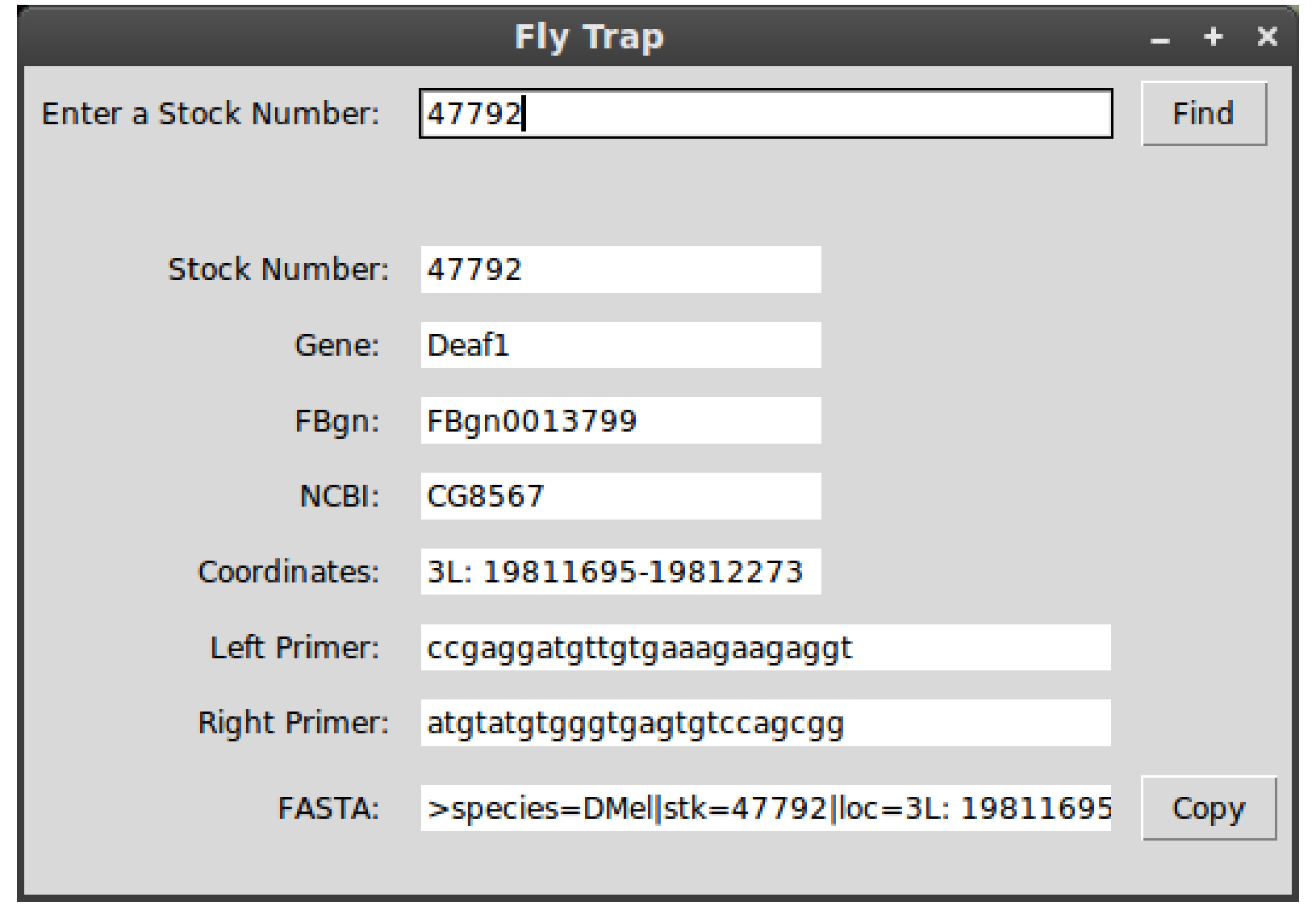
1) From the 1,043 JF lines surveyed, we identified 14 fly lines showing gamma lobe expression (Table 1).

2) We created 6,931 total FASTA files, one for each fly line.

3) Motif discovery on DNA sequences of the selected 14 fly lines returned 18 unique motifs (Table 2).

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 2).

Figure 2: Screenshot of FlyTrap being used to find information related to BDSC stock number 47792.



Enter a Stock Number: 47792

Stock Number: 47792

Gene: Deaf1

FBgn: FBgn0013799

NCBI: CG8567

Coordinates: 3L: 19811695-19812273

Left Primer: ccgaggatgttgtgaaagaagaggt

Right Primer: atgtatgtgggtgagtgtccagcgg

FASTA: >species=DMel|stk=47792|loc=3L: 19811695

CONCLUSIONS

- From this study, we found 14 fly lines showing gamma lobe expression out of 1,043 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

- Process the results of the computational motif search to identify the best motif candidates.

- 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.

Okumura T, *et al.* (2007). Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions. *Nucleic Acids Research* 35:W227-31

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, Ngo TT, Misra S. GAL4 driver collection of Rubin Laboratory at Janelia Farm.

Sinha S, & Tompa M. (2003). YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research* 31(13):3586-8.

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.

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.