DNA motif screening for genes expressed in the mushroom body brain structure of *Drosophila melanogaster* using Fisher's exact test and the χ^2 test for homogeneity

Raymond A. 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

Abstract

Several existing bioinformatics software for identifying DNA motifs are combined with a novel tool to develop a new approach for identifying potential enhancer elements for a set of genes related to learning and memory in *Drosophila*. Using six different DNA motif discovery software tools, we identified 160 conserved motifs among the DNA sequences driving mushroom body expression. We took two different approaches to identify DNA motifs among 28 lines known to display γ-lobe expression: the first based on the length of individual line sequences, and the second based upon stochastic line selection. 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. Eight candidate DNA motifs were identified. The frequency of each motif's occurrence was noted among the 28 lines displaying γ-lobe expression against 28 randomly selected lines known to not display γ-lobe expression. The Fisher's exact test and χ^2 test of homogeneity were performed for each motif and the p-values were compared to a level of significance established with the use of a Bonferroni correction. This information can be used to identify potential enhancer elements for the genes expressed in the mushroom body γ neurons. With these results, we can further identify additional genes expressed in the γ neurons responsible for dopamine and octopamine signals. This study would ultimately help understand the underlying pathological mechanisms of ADHD, autism, schizophrenia, Parkinson's disease, and drug abuse/addiction.

Background

D. melanogaster is a powerful model organism due to its well characterized genetics, fully sequenced genome, and its sophisticated nervous system comparable to humans; moreover, the study of insect mushroom bodies in the insect brain is important to understand learning, memory, and other brain functions. The mushroom bodies are a widely studied and well defined structure. The mushroom bodies of *D. melanogaster* are composed of three distinct structures: α, β, and γ lobes.

Figure 1: *D. melanogaster* also known as the fruit fly. Source: http://caltech.edu

Our research targets γ neurons due to their important function of mediating dopamine and octopamine signals responsible for learning and memory. The understanding of these mechanisms can lead to uncovering important pathways associated with abnormal dopamine function. Abnormal dopamine function is responsible for various neuropsychiatric diseases: ADHD, autism, schizophrenia, and drug abuse/addiction.

Objectives

1. Find DNA motifs in the γ neurons critical to the study of neural transmissions which have a potential effect in neuropsychiatric disorders.
2. Screen the discovered DNA motifs to reduce the number of potential sequence-specific binding sites.
3. Help in the study of the pathogenesis mechanisms underlying learning and memory impairments by identification of enhancer elements.
4. Develop open source software applications that will allow researchers to access data related to *D. melanogaster* more easily.

Materials and Methods

The expression data videos for all lines having a DNA sequence length less than 1000 base pairs were surveyed to determine whether or not a line displayed γ-lobe expression. The DNA sequence data for all lines displaying γ-lobe expression was then used to discover DNA motifs with the aide of the following DNA motif discovery tools: YMF, MEME, CONSENSUS, MDscan, Weeder, and Gibbs Motif Sampler. A novel tool, FlyTrap, was developed to assist with the creation of the FASTA input files for these DNA motif tools. Five DNA motif discovery runs were performed, with the lines in the first run being selected based on sequence length, and the lines for the other four runs being randomly selected.

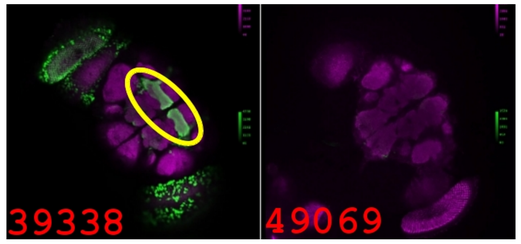

Figure 2: Annotated screenshots of the video expression data for lines 39338 and 49069. The γ-lobe expression for line 39338 is demarcated by the yellow ellipse. Adapted from http://flyweb.janelia.org

A sample of 28 lines not displaying γ-lobe expression and having less than 997 base pairs in their individual DNA sequences was drawn randomly. The number of these lines having selected motifs in their DNA sequence data was recorded and compared to the number of lines displaying γ-lobe expression in the form of a 2×2 contingency table for each motif. These tables were then used to perform two statistical significance tests: Fisher's exact test and the χ^2 test for homogeneity. The result of a test was only deemed to be significant if the computed p-value was less than an adjusted level of significance (Bonferroni correction).

Fisher's exact test and the χ^2 test for homogeneity Data are summarized in the form of a 2×2 contingency table.

	1	0
γ^+	x_1	$n_1 - x_1$
γ^-	x_2	$n_2 - x_2$

Table 1: An example of a 2×2 contingency table.

The test statistic for each test is given by:

$$\text{Test Statistic} := \begin{cases} x_1 & \text{Fisher's exact test} \\ \sum_{i=1}^{2} \frac{(x_i - n_i\hat{\theta})^2}{n_i\hat{\theta}(1-\hat{\theta})} & \chi^2 \text{ test for homogeneity} \end{cases}$$

Bonferroni correction Let X denote the number of significant results obtained—due to chance alone—when testing a set of k hypotheses simultaneously, with a given significance level of α. Then

$$\begin{aligned} \mathbb{P}(X \geq 1 \mid \alpha) &= 1 - \mathbb{P}(X = 0) \\ &= 1 - (1 - \alpha)^k. \end{aligned}$$

As k increases, so too does $\mathbb{P}(X \geq 1)$. One way to address this issue is to only declare results of a test to be significant if the p-value is less than α/k.

Results

- From this study, we identified 28 lines showing γ-lobe expression out of 1043 revised files.
- From 160 discovered motifs, only eight were found to be of interest and from these eight motifs, only one was found to be statistically significant.

48001	46896	48588	47792	38832	47708	40019
45998	49218	49656	46974	45256	45258	45051
47840	46565	46676	39338	45228	46672	46537
46669	39813	48921	45092	47617	39900	38737

Table 2: The 28 lines identified to display γ-lobe expression.

39370	50474	49942	49591	45669	45180	46000
46902	49035	48435	45100	50284	47410	47319
49936	39092	46175	45752	45076	45301	45755
39170	38831	45045	49374	45294	46111	46666

Table 3: The 28 randomly selected lines not displaying γ-lobe expression.

ATCGCTTCAA

	1	0	
γ^+	3	25	28
γ^-	0	28	28
	3	53	56

TSCMTSCYCCYC

	1	0	
γ^+	4	24	28
γ^-	0	28	28
	4	52	

GGCACCGG

	1	0	
γ^+	3	25	28
γ^-	0	28	28
	3	53	56

YSCMKSCYMCTS

	1	0	
γ^+	7	21	28
γ^-	0	28	28
	7	49	56

WGTGTGTGTATG

	1	0	
γ^+	3	25	28
γ^-	0	28	28
	3	53	56

AAGGTCKT

	1	0	
γ^+	4	24	28
γ^-	0	28	28
	4	52	56

CTCGTATC

	1	0	
γ^+	3	25	28
γ^-	0	28	28
	3	53	56

YCSTTTYTSYCT

	1	0	
γ^+	3	25	28
γ^-	0	28	28
	3	53	56

Table 4: Contingency tables for each of the eight candidate motifs.

	Test			
	χ^2		Fisher's exact	
Motif	Test statistic	p-value	Test statistic	p-value
ATCGCTTCAA	3.1682	0.0751	3.0000	0.1182
TSCMTSCYCCYC	4.3077	0.0379	4.0000	0.0558
GGCACCGG	3.1682	0.0751	3.0000	0.1182
YSCMKSCYMCTS	8.0000	0.0047	7.0000	0.0051
WGTGTGTGTATG	3.1682	0.0751	3.0000	0.1182
AAGGTCKT	4.3077	0.0379	4.0000	0.0558
CTCGTATC	3.1682	0.0751	3.0000	0.1182
YCSTTTYTSYCT	3.1682	0.0751	3.0000	0.1182

Table 5: A comparison of the χ^2 and Fisher's exact test for each of the candidate motifs.

Conclusions

- With the known motifs we can look for specific binding sites for proteins such as transcription activators and repressors, splicing machinery, and chromatin remodeling components responsible for tissue-specific expression.
- The FlyTrap GUI application allowed us to access data related to *D. melanogaster* more easily.

Future Work

- Examine the expression patterns of the additional lines containing the motif under study to substantiate it as a key enhancer element for γ-lobe expression.
- Search for the significant motif in other *Drosophila* species to explore whether it is conserved across species.
- Make the source code of FlyTrap, along with all the other data files that it needs to run, publicly available under a GPLv3 license.

References

1. 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.
2. Bloomington Drosophila Stock Center. (2014). Janelia GAL4 stocks.
3. Conover, WJ. (1999). *Practical nonparametric statistics*. Hoboken, New Jersey: John Wiley and Sons Inc.
4. 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.
5. Pfeiffer BD, Jenett A, Hammonds AS, Ngo TT, Misra S. (2014). GAL4 driver collection of Rubin Laboratory at Janelia Farm.
6. 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.
7. 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 Educacion, Deporte y Cultura), Chih. Chih. Mexico 2014.