

A multiple-instance scoring method to predict tissue-specific *cis*-regulatory motifs and regions

Jin Gu

MOE Key Laboratory of Bioinformatics and Bioinformatics Division

TNLIST / Department of Automation, Tsinghua University

Beijing 100084, China

Email: jgu@tsinghua.edu.cn

Abstract—Transcription is the central process of gene regulation. In higher eukaryotes, the transcription of a gene is usually regulated by multiple *cis*-regulatory regions (CRRs). In different tissues, different transcription factors bind to their *cis*-regulatory motifs in these CRRs to drive tissue-specific expression patterns of their target genes. By combining the genome-wide gene expression data with the genomic sequence data, we proposed multiple-instance scoring (MIS) method to predict the tissue-specific motifs and the corresponding CRRs. The method is mainly based on the assumption that only a subset of CRRs of the expressed gene should function in the studied tissue. By testing on the simulated datasets and the fly muscle dataset, MIS can identify true motifs when noise is high and shows higher specificity for predicting the tissue-specific functions of CRRs.

Keywords- *Multiple-instance learning, cis-regulatory region(s), cis-regulatory motif(s)*

I. INTRODUCTION

In higher eukaryotes, the transcription of a gene is usually regulated by multiple *cis*-regulatory regions (CRRs). Different transcription factors bind to the *cis*-regulatory motifs (hereinafter referred as motifs for short) in these CRRs and lead to specific expression patterns of the gene in different tissues. The lengths of motifs vary from several to more than twenty nucleotides and usually within six to twelve, while CRRs commonly cover several hundred nucleotides. Identifications of the tissue-specific motifs and corresponding CRRs are essential for understanding the complex gene transcriptional regulations. Now genome-wide experimental identifications of the

tissue-specific CRRs and motifs remain time-consuming and expensive. So many computational methods have been developed to solve this problem [1-5]. One popular approach is to find the enriched motifs in the candidate CRRs of the co-expressed genes or highly expressed genes. Given the following inputs: 1) in the studied tissue, the genes are labeled as positive or negative according to whether they are expressed (or highly expressed); 2) each gene has one or more candidate CRRs and the functional CRRs should be enriched for the positive genes but it is unknown which CRRs are functional; and 3) the enrichments of motifs have been calculated in each CRR, the proposed method should identify the enriched tissue-specific motifs and CRRs by analyzing the motifs' enrichments in the CRRs of positive genes. Most previous methods assume that the motifs are enriched in all candidate CRRs of the positive (highly expressed or co-expressed) genes [6-9]. But this assumption may not be correct [10]. For example, *eve*, an important gene for fly embryo development, has four CRRs, which function in different segments of the embryo to drive *eve*'s stripe expression pattern [1, 3]. It is a better assumption that only a subset of the candidate CRRs of the positive genes are functional and the tissue-specific motifs are enriched in these functional CRRs in the studied tissue.

In recent years, multiple-instance learning proposes a new machine learning framework to deal with the labeling incomplete data [11-13]. Above problem can be formulated in a multiple-instance learning framework: 1) each bag (gene) is labeled as positive or negative according to the genome-wide expression data; 2) each bag (gene) has multiple instances (CRRs) which are unlabeled, but positive instances (CRRs) are enriched in the positive bags (genes) than in the negative bags; 3) the features of each instance (CRR) are the motifs' enrichments (counts or scores) in that CRR. Zhang et al. proposed a multiple-instance learning method, named multiple-instance learning via embedded instance selection (MILES) method to identify motifs [14]. They regarded the motifs directly as instances but the method did not consider the problem of multiple CRRs for each gene. Actually, their method is largely similar to the maximum scoring approach proposed by Andrew et al., which used the maximum matched score to represent each motif's feature for each gene [15]. Here, we proposed a more proper multiple-instance learning description of above problem: 1) define the genes as bags, and labels of genes are given as supervised information in the studied tissue; 2) define the candidate CRRs as instances and each instance is assigned to a unique bag (gene);

and 3) define the feature space of instances as the vector consisting of the scores or enrichments of candidate motifs.

Due to the imbalances of the numbers of positive and negative genes and the high noises for screening the conserved candidate regulatory regions, classical two-class multiple-instance learning algorithm cannot achieve a bearable performance (mi-SVM & MI-SVM, provided in MILL package <http://www.cs.cmu.edu/~juny/MILL/>, were tested). In this study, we proposed a new multiple-instance learning method, named multiple-instance scoring (MIS), to alleviate the imbalances. By testing on the simulated data and the real data in fly, MIS shows higher power for identifying motifs and can achieve higher specificity for predicting the tissue-specific CRRs.

II. MATERIALS AND METHODS

A. *Prepare the labeled bags (genes)*

The expression data were extracted from the BDGP *in situ* database (<http://www.fruitfly.org/cgi-bin/ex/insitu.pl>). The list of detected genes in fly embryonic muscle (with ontology “embryonic/larval muscle system”) was downloaded from the database. To reduce the bias caused by the incomplete experiment and house-keeping genes, the genes with less than 10 images and the genes detected in more than 20 tissues were filtered. In the studied tissue, the detected genes were labeled as positive bags and the genes not detected were labeled as negative bags.

B. *Assign instances (CRRs) to bags (genes)*

In this study, all the coordination, gene annotations and genomic sequences were downloaded from FlyBase (v4.2.1) (<http://www.flybase.org>). The PhastCons conservation scores were downloaded from UCSC Genome Browser ftp site (<http://genome.ucsc.edu>). According to the gene annotations, the noncoding regions, including intergenic regions, introns, and un-translated regions (UTRs), were extracted. A 100bp sliding window with 50bp step was used to scan across the noncoding regions and average PhastCons score was computed in each window. According to the average scores in the sliding windows, the genomic regions with average scores > 0.6 and length 200bp~2,000bp were extracted.

These extracted conserved regions were assigned to the nearby gene if they are located in the -3k~+1k flanking regions around the transcription start site (TSS) or transcriptional terminal site (TTS) of the gene (another dataset with -5k~+1k flanking

regions was also constructed. Due to the limitation of space, the results were not shown in the article). The conservation in different genomic regions varies significantly. To lower the bias, the regions were sorted according to the products of their lengths and the average scores, and then only the top four regions were kept as instances (CRRs) for each bag (gene). The core promoter regions (-300bp~+100bp around TSS) were also added as CRRs. This procedure is similar to [7].

C. Calculate the features (motifs' enrichments) for instances (CRRs) in positive bags (genes)

Sixty-nine PWMs related to fly were extracted from TRANSFAC (v11.2) [16]. The nucleotide contents in each position were normalized as percentage in the matrices.

In the studied tissue, the CRRs of the negative bags (genes) were combined as the background sequences. Then the features (motifs' enrichment, one by one) for each instance (CRR) of the positive bags (genes) were computed against the background sequence using CLOVER program [17]. The times of random sampling (parameter $-r$) were set to 100,000 to estimate the p-value. Then the p-value was log-transformed as enrichment score D for k -th feature (motif) in the j -th instance (CRR) of the i -th bag (gene):

$$D_{ijk} = -\log(p_{ijk})$$

Larger D means that the instance (CRR) is more relevant to the positive bag and more distant to the negative under the k -th feature. The detail for CLOVER algorithm can be referred to [17]. For ranking the instances according to a set of m motifs, simple linear sum was used to transform the feature vector to a single score:

$$D_{ij} = \sum_{k=1}^m D_{ijk}$$

D. Multiple-instance scoring (MIS) method

After calculating the motifs' enrichment scores for all instances (CRRs) in positive bags (genes) against the instances in negative bags, the multiple-instance scoring (MIS) method takes the instance with the maximum enrichment score $\max_j(D_{ij})$ to represent i -th positive bag:

1) All the instances (CRRs) of positive bags (genes) were decreasingly sorted according to their features (enrichment scores) D_{ij} for all i and j . The ranked instances are re-denoted as single subscript l ;

2) Calculate MIS , a running statistic, by going through the sorted instance (CRR) list:

$$MIS(k) = \left(\frac{1}{k} \sum_{l=1}^k D_l \right) \left(\frac{1 - q^{-\lambda C}}{1 + q^{-\lambda C}} \right) \quad q = 2, \lambda = 10 \text{ for this study}$$

C is defined as the percentage of positive bags (genes) whose instances (CRRs) are ranked before k . C is a key variable, which links the instances (CRRs) with bags (genes): single CRR is assumed to be strong enough to drive the expression of the corresponding gene;

3) When going through the decreasingly sorted instance (CRR) list, the first part of MIS $\left(\frac{1}{k} \sum_{l=1}^k D_l \right)$ is gradually decreased and the second part $\left(\frac{1 - q^{-\lambda C}}{1 + q^{-\lambda C}} \right)$ is gradually increasing. So MIS will reach its maximum MIS^* ($MIS^* = \max_k (MIS(k))$) at k^* ($k^* = \arg \left(\max_k (MIS(k)) \right)$) when running through the sorted instance list. MIS^* was used to evaluate the motifs' relevance to the studied bags (genes) and the instances (CRRs) with rank before k^* are classified as positive for this motif or motif combination.

Given a motif or a motif combination, we can calculate its MIS^* and k^* according to above procedure in the studied tissue. But the raw scores cannot be directly used to evaluate the statistical significance. False discovery rates (FDRs) were calculated as the method in GSEA: first, the labels of genes were randomly shuffling; second, compute the global FDR by comparing the distribution of the MIS^* in the original dataset and the distribution of the MIS^* in the shuffled datasets (see detail in [18]).

E. Evaluation of MIS method's performance

Experimentally identified tissue-specific CRRs. REDfly (v2.1) database have collected 665 experimentally identified CRRs (<http://redfly.ccr.buffalo.edu/>) [19]. To validate the genome-wide predictions, a testing dataset was constructed based REDfly. The 665 collected CRRs were clustered by genomic location and filtered by length

(200bp~1,500bp) to 251 non-overlapped reference CRRs. Then the tissue-specific functions of these CRRs were manually processed according to the database annotations. No further filters, such as evolutionary conservation was used to process the CRRs.

Comparison with other methods. 1) Traditional scoring (TS) method: label all CRRs (instances) of the positive genes (bags) as positive, and construct the new positive bags each containing single positive instance. Then the same procedure was run as MIS method. 2) The tissue-specific motifs reported by CLOVER are also compared. Because the previous studies do not provide stand-alone software and their used testing datasets are largely different, the comparisons with the Bayesian network method [6] and Enhancer Index [9] were not included in this study.

III. RESULTS AND DISCUSSIONS

A. Results on simulated datasets

Firstly, simulated data were used to test the performances of the multiple-instance scoring (MIS) method. The positive dataset contained 100 bags (genes) and the negative dataset contained 400 bags (genes). Each bag was assigned four instances (*cis*-regulatory regions, CRRs) with equal length 500bp. For positive bags (genes), 1~4 positive instances (CRRs) were assigned according to a pre-defined probability. The positive instances (CRRs) were generated by implanting known transcription factor binding sites (five motifs' PWMs were used: BCD, MEF2, STAT, TWI and UBX) into a random sequence.

When only one motif's PWM is used to calculate the enrichment score D_{ij} for the i -th gene (bag) and j -th CRR (instance), the resulting scores computed by MIS can be used to rank the enriched motifs in the CRRs of the positive genes. The top five motifs reported by CLOVER, MIS and TS are listed in (Table I). The three methods show competitive performances when the noise is lower than 60%, but MIS is much more stable when the noise is up to 80%.

To identify the positive CRRs, the top five motifs were used to compute the combined enrichment score D . The results show that the sensitivity of MIS is lower than TS, but the specificity is much higher: if the noise is no more than 60%, the classifier can achieve ~90% specificity for identifying the positive regions of the

positive genes (Table II). For genome-wide predictions, specificity is relatively more important than sensitivity due to large background sequences. MIS would provide more stable predictions under high noises.

TABLE I. THE TOP FIVE MOTIFS REPORTED BY CLOVER, MIS AND TS

Noise	CLOVER	MIS	TS
0%	STAT, MEF2, TWI, BCD, UBX	STAT, MEF2, BCD, TWI, UBX	STAT, MEF2, BCD, TWI, UBX
20%	STAT, MEF2, TWI, BCD, UBX	STAT, MEF2, BCD, UBX, TWI	MEF2, STAT, UBX, BCD, TWI
40%	STAT, TWI, BCD, UBX, SN	STAT, TWI, BCD, UBX, ABDB	STAT, TWI, UBX, BCD, DL
60%	TWI, TATA, STAT, MEF2, BYN	STAT, TATA, TWI, MEF2, BYN	STAT, TATA, TWI, MEF2, BCD
80%	AP1, DEAF1, CROC, ADF1, SRYBETA	STAT, TWI, BCD, ADF1, CROC	AP1 ADF1, FTZ, SRYBETA, CROC
100%	EVE, TCF, BRK, AP1, SD	TCF, HSF, EVE, SD, AP1	TCF, EVE, HSF, SD, BRK

The motifs denoted by bold font mean these motifs are in the five motifs which are used to generate the positive regions.

TABLE II. SUMMARY OF THE RESULTS FOR CLASSIFYING THE INSTANCES (CRRs) ON THE SIMULATED DATASET

Noise	MIS			TS		
	#PR	#NR	Spec.	#PR	#NR	Spec.
0%	101	1	99.02%	119	9	92.97%
20%	64	0	100.00%	132	19	87.42%
40%	52	6	89.66%	106	40	72.60%
60%	41	5	89.13%	108	27	80.00%
80%	41	13	75.93%	17	38	30.01%
100%	12	31	27.91%	17	28	37.78%

#PR: the number of positive CRRs of the positive genes which have been predicted as positive; #NR: the number of negative CRRs of the positive genes which have been predicted as positive.

B. Identifications of muscle-specific motifs and CRRs

In fly, the transcriptional regulations of muscle are relatively well-studied. So the MIS method was further tested on the muscle-specific dataset. Because not all the motifs are functional in the studied tissues, the motifs which are enriched in the

candidate CRRs of muscle-specific genes were first identified. Three different methods: CLOVER, MIS and TS were used to rank the motifs. The motifs with FDR < 30% computed by MIS/TS and with p-value < 0.05 computed by CLOVER were listed in Table III. For CLOVER only two of the six significant motifs were related to muscle system: *twi* (I\$TWI_Q6) and *sna* (I\$SN_02). For TS, three of the four selected motifs were related to muscle: *twi* (I\$TWI_Q6), *sna* (I\$SN_02) and *mef2* (V\$MEF2_02). For MIS, five of the seven selected motifs were related to muscle: *twi* (I\$TWI_Q6), *sna* (I\$SN_02), *mef2* (V\$MEF2_02), *ap* (I\$AP_Q6) and *retn* (I\$DRI_01) (According to the annotations of the transcription factors in muscle in the Interactive Fly <http://www.sdbonline.org/fly/aimorph/mesoderm.htm>). These results indicate that MIS can achieve higher power to identify the tissue-specific motifs in muscle.

TABLE III. THE TOP MOTIFS REPORTED BY CLOVER, MIS, TS AND THE CORRESPONDING P-VALUE AND FDR

Tissue	CLOVER (p-value)	MIS (FDR)	TS (FDR)
Muscle	I\$DREF_Q3: 0.0027	<u>I\$AP_Q6</u> : 13.64%	<u>V\$MEF2_02</u> : 09.09%
	I\$CF1_02: 0.0093	I\$DREF_Q3: 18.18%	I\$DREF_Q3: 27.27%
	<u>I\$TWI_Q6</u> : 0.0099	<u>V\$MEF2_02</u> : 24.24%	<u>I\$TWI_Q6</u> : 29.09%
	<u>I\$SN_02</u> : 0.0137	<u>I\$DRI_01</u> : 24.24%	<u>I\$SN_02</u> : 29.55%
	I\$CF1_01: 0.0147	I\$STAT_01: 27.27%	
	I\$ZEN_Q6: 0.0191	<u>I\$TWI_Q6</u> : 28.57%	
		<u>I\$SN_02</u> : 29.55%	

Then the motifs with FDR < 30% were used to construct the classifier to identify the CRRs functional in muscle. The selected motifs may have redundant information for classifying, so forward selection process was used to find the optimal motif combination. Also, because the false negatives in the in situ data (such as the important transcription factor *twi*, which is not annotated by the ontology “embryonic/larval muscle system”) and the inaccuracies for preparing the candidate regions by comparative genomic methods, an independent experimentally identified dataset derived from REDfly (Table IV) was used to estimate the classifier’s performance.

For MIS, 107 regions (87 genes) were classified as positive (three motifs were selected by the forward selection: V\$MEF2_02, I\$AP_Q6 and I\$DRI_01). On the

REDfly dataset, MIS achieved 57.89% (11/19) sensitivity, 25.00% specificity (11/44), F-value 0.3492. For TS, 205 regions (138 genes) were classified as positive (only V\$MEF2_02 was selected). On the REDfly dataset, TS achieved 63.16% (12/19) sensitivity, 15.38% (12/78) specificity, F-value 0.2474. These results indicate that MIS can achieve much higher specificity but not significantly reduce sensitivity (Figure 1).

TABLE IV. THE SUMMARY ON THE FLY MUSCLE TESTING DATASET

#PG	#PR	#PR/#PG	#NG	#NR	#PRED	#NRED
215	638	2.97	1631	5450	19	232

#PG: the number of the positive genes; #PR: the number of the CRRs of the positive genes; #NG: the number of the negative genes; #NR: the number of the CRRs of the negative genes; #PRED: the number of muscle CRRs annotated in REDfly; #NRED: the number of the other CRRs.

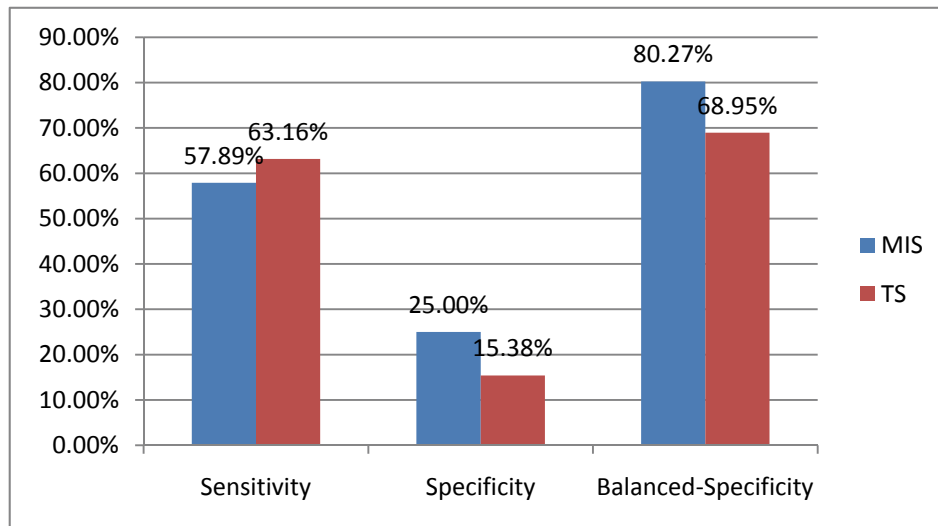


Figure 1. The performances for identifying the muscle CRRs.

The genome sequences and global gene expression data are quickly accumulating, but the complex relations between the genes and the CRRs make the computational predictions of the tissue-specific motifs and CRRs still difficult. The multiple-instance scoring (MIS) method models the relation between CRRs and the genes as that only a subset of CRRs of the expressed gene should function in the studied tissue. By testing on the simulated and the experimental datasets, the MIS method can achieve higher performance.

Although the MIS method shows better specificity, it remains further improvement. Except the false positives of the comparative genomic methods for

preparing the candidate CRRs and the noises in the gene expression data, the uncertainties in the interactions between multiple motifs worsen the situation. Here a simple forward selection was used to select the optimal motif combination by considering their “add” effect. Beyond this simple method, more sophisticated models can be used to compute the enrichment of multiple motifs, such as Hidden-Markov Model [20] and TFBS alignment model [3]. These models will be tested systematically in the future version of MIS.

ACKNOWLEDGMENT

The author thanks Yanda Li for extensive discussions. The author also thanks Xuegong Zhang, Xiaowo Wang and Tao Peng for useful discussions.

REFERENCES

- [1] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen, "Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome," *Proc Natl Acad Sci U S A*, vol. 99, pp. 757-62, Jan 22 2002.
- [2] B. P. Berman, B. D. Pfeiffer, T. R. Laverty, S. L. Salzberg, G. M. Rubin, M. B. Eisen, and S. E. Celniker, "Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*," *Genome Biol*, vol. 5, p. R61, 2004.
- [3] O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen, and J. Taipale, "Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity," *Cell*, vol. 124, pp. 47-59, Jan 13 2006.
- [4] M. Markstein, P. Markstein, V. Markstein, and M. S. Levine, "Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo," *Proc Natl Acad Sci U S A*, vol. 99, pp. 763-8, Jan 22 2002.
- [5] X. Yu, J. Lin, D. J. Zack, and J. Qian, "Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors," *BMC Bioinformatics*, vol. 8, p. 437, Nov 9 2007.
- [6] X. Chen and M. Blanchette, "Prediction of tissue-specific cis-regulatory modules using Bayesian networks and regression trees," *BMC Bioinformatics*, vol. 8 Suppl 10, p. S2, 2007.
- [7] L. A. Pennacchio, G. G. Loots, M. A. Nobrega, and I. Ovcharenko, "Predicting tissue-specific enhancers in the human genome," *Genome Res*, vol. 17, pp. 201-11, Feb 2007.
- [8] L. A. Pennacchio, N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B. L. Black, O. Couronne, M. B. Eisen, A. Visel, and E. M. Rubin, "In vivo enhancer

- analysis of human conserved non-coding sequences," *Nature*, vol. 444, pp. 499-502, Nov 23 2006.
- [9] A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio, "VISTA Enhancer Browser--a database of tissue-specific human enhancers," *Nucleic Acids Res*, vol. 35, pp. D88-92, Jan 2007.
- [10] M. S. Halfon, "(Re)modeling the transcriptional enhancer," *Nat Genet*, vol. 38, pp. 1102-3, Oct 2006.
- [11] T. G. Dietterich, R. H. Lathrop, and T. LozanoPerez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31-71, Jan 1997.
- [12] S. Andrews, T. Hofmann, and I. Tsochantaridis, "Multiple instance learning with generalized support vector machines," in *In Proceedings of the AAAI National Conference on Artificial Intelligence*, 2002.
- [13] Z. H. Zhou, "Multi-instance learning from supervised view," *Journal of Computer Science and Technology*, vol. 21, pp. 800-809, Sep 2006.
- [14] Y. Zhang, Y. Chen, and X. Ji, "Motif discovery as a multiple-instance problem," *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, pp. 805-809, 2006.
- [15] A. D. Smith, P. Sumazin, and M. Q. Zhang, "Tissue-specific regulatory elements in mammalian promoters," *Mol Syst Biol*, vol. 3, p. 73, 2007.
- [16] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res*, vol. 31, pp. 374-8, Jan 1 2003.
- [17] M. C. Frith, Y. Fu, L. Yu, J. F. Chen, U. Hansen, and Z. Weng, "Detection of functional DNA motifs via statistical over-representation," *Nucleic Acids Res*, vol. 32, pp. 1372-81, 2004.
- [18] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A*, vol. 102, pp. 15545-50, Oct 25 2005.
- [19] M. S. Halfon, S. M. Gallo, and C. M. Bergman, "REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in *Drosophila*," *Nucleic Acids Res*, Nov 26 2007.
- [20] Q. Zhou and W. H. Wong, "CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling," *Proc Natl Acad Sci U S A*, vol. 101, pp. 12114-9, Aug 17 2004.