Improved enhancer discovery in *Drosophila* and other insects

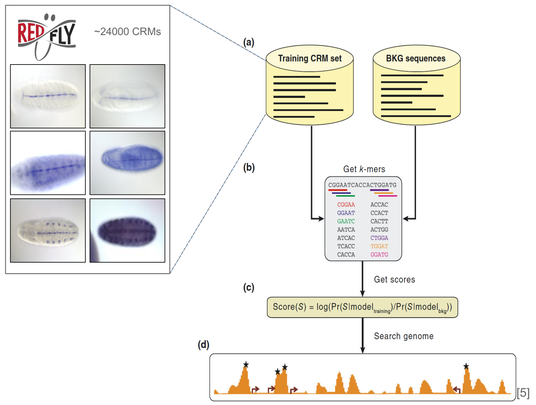
Hasiba Asma[1], Chad M. Jaenke[2], Michael L. Weinstein[2], Thomas M. Williams[2] and Marc S. Halfon[1]

[1]University at Buffalo-State University of New York, Buffalo, NY
[2]University of Dayton, Dayton, OH

Introduction

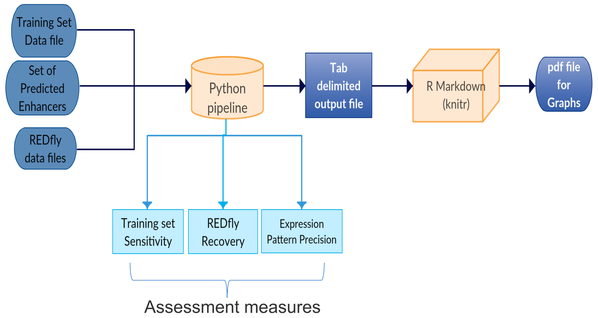
Transcriptional enhancers, or more broadly, cis-regulatory modules (CRMs), are essential building blocks of gene regulatory networks. We previously developed the SCRMshaw method for computational CRM discovery[1,2,3]. SCRMshaw uses the wealth of known *D. melanogaster* CRMs as training data to facilitate CRM discovery in not just *Drosophila* but in diverse holometabolous insects including mosquitoes, beetles, and bees. **Here we present three approaches for increasing SCRMshaw's effectiveness.**

SCRMshaw: predicting CRMs



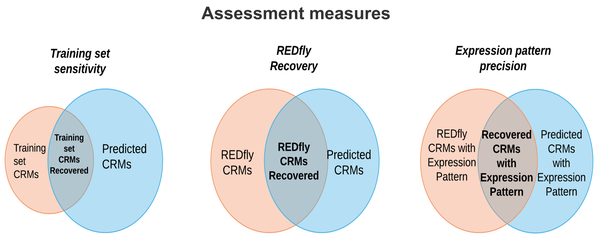
(a) SCRMshaw takes a training set of similar CRMs and a set of similarly-sized non-CRMs as a background (BKG) set, **(b)** builds up kmer profiles of the sequence sets, **(c)** generates scores using statistical models , **and (d)** searches the genome for high scoring windows (predicted CRMs).

1. pCRM_eval: A comprehensive pipeline for in silico evaluation of CRM prediction approaches[4]

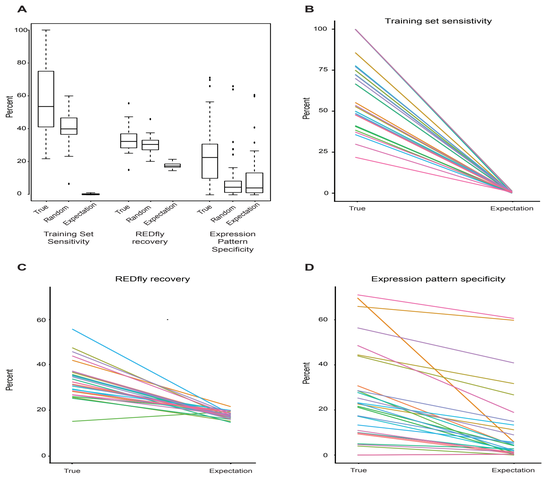


Three measures of Assessment

pCRMeval compares prediction results with the existing extensive corpus of validated *Drosophila* CRMs to calculate recovery of true CRMs and estimate the specificity of a given method. ***pCRMeval* can also assess the performance of a specific training set in terms of both sensitivity and specificity.**



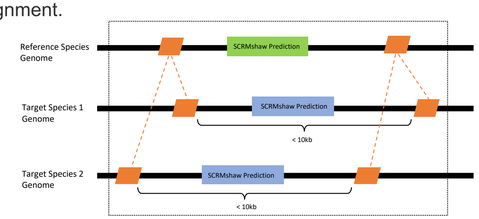
pCRM_eval demonstrates that SCRMshaw performs better than random expectations



(A) Aggregate performance for 29 true training sets, 62 random training sets, and random expectation. Comparison of *training set sensitivity* **(B)**, *REDfly recovery* **(C)**, and *expression pattern specificity* **(D)** for true versus random expectations for each of the 29 training sets.

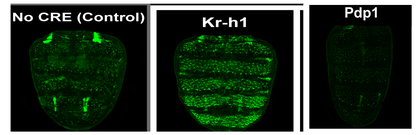
2. Toward individual prediction confidence scores

A true CRM might be predicted in multiple related species. We are developing a "weighted comparative confidence" score to identify such conserved CRMs, even in the absence of sequence alignment.



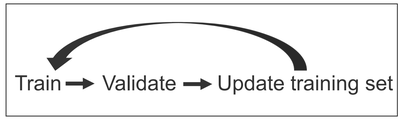
3. Iterative searching can serve to augment weak training sets to improve true-positive : false-positive ratios

We used SCRMshaw on a small training set of 7 CRMs to identify CRMs within a gene regulatory network for *Drosophila* abdominal pigmentation.



Empirical testing of 18 SCRMshaw predictions revealed 10 true (e.g. Kr-h1) and 8 false positive (e.g. Pdp1) prediction results.

These 10 new validated CRMs were combined with the original 7 and the 2.5-fold expanded training set used for a new round of SCRMshaw prediction.



Following is the summary of the results after getting predictions from updated training set.

In vivo validation	Total No. of CRMs	No. of CRMs predicted by original Training set	No. of CRMs predicted by updated Training set
True Positives	10	10/10	10/10*
False Positives	8	8/8	1/8*

*Notably prediction results from updated training set contain all the previous true positives and only one false positive demonstrating a **marked improvement in prediction specificity using the updated training data.**

Funding

This work is supported by United States Department of Agriculture (USDA)

1. Kantorovitz, Miriam R., et al.. Developmental cell 17.4 (2009): 568-79.
2. Kazemian, M., et al. Nucleic Acids Res 39.22 (2011): 9463-72.
3. Kazemian M, Halfon MS: CRM Discovery Beyond Model Insects.: Insect Genomics: Methods and Protocols; 2019: 117
4. Asma H, Halfon MS:. BMC Bioinformatics 2019, 20(1):174.
5. Suryamohan K, Halfon MS:. Wiley interdisciplinary reviews Developmental biology 2015, 4(2):59-84

University at Buffalo The State University of New York