International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:7, No:10, 2013

An Integrated Predictor for Cis-Regulatory Modules

Darby Tien-Hao Chang, Guan-Yu Shiu, You-Jie Sun

Abstract**—**Various cis-regulatory module (CRM) predictors have been proposed in the last decade. Several well-established CRM predictors adopted different categories of prediction strategies, including window clustering, probabilistic modeling and phylogenetic footprinting. Appropriate integration of them has a potential to achieve high quality CRM prediction. This study analyzed four existing CRM predictors (ClusterBuster, MSCAN, CisModule and MultiModule) to seek a predictor combination that delivers a higher accuracy than individual CRM predictors. 465 CRMs across 140 Drosophila melanogaster genes from the RED fly database were used to evaluate the integrated CRM predictor proposed in this study. The results show that four predictor combinations achieved superior performance than the best individual CRM predictor.

Keywords**—**Cis-regulatory module, transcription factor binding site.

I. Introduction

A cis-regulatory module (CRM) is a stretch of DNA sequence of 10 to 1000 base pairs (bp) that contains three to five transcription factor binding sites (TFBSs) [1]. It is critical to the transcription of its downstream genes. Understanding CRMs helps to know gene regulation and the related biological mechanisms [1]-[4].

Various CRM predictors have been proposed in the last decade. The prediction strategies of these CRM predictors can be roughly split into three categories (Table I). The first category is window clustering, which identifies DNA regions with significantly high densities of binding sites [5]. The second category is probabilistic modeling, which describes binding site clusters as statistical models and extracts those have higher scores than a background model [6]-[8]. Most predictors in this category adopted the hidden Markov model (HMM) [9]. The third category is phylogenetic footprinting, which searches for conserved regions that contain binding site clusters [8]. A disadvantage of the predictors in this category is that they require sequence data of closely related genomes to compute conservation. In addition to related genomes, some predictors ask users to input the motifs of TFBSs. On the other hand, predictors that have a built-in pattern mining algorithm can generate the motifs of TFBSs in the runtime without depending on the input. The required input data of the abovementioned CRM predictors are summarized in Table II.

TABLE I
Search Strategies of CRM Predictors

Predictor	Window Clustering	Probabilistic Modeling	Phylogenetic Footprinting
CisModule		✓	
ClusterBuster		✓	
MSCAN	✓		
MultiModule		✓	✓

TABLE II
Input Data of CRM Predictors

Method	Genome Sequence	Sequences of Related Genome	Motifs of TFBSs
CisModule	✓		
ClusterBuster	✓		✓
MSCAN	✓		✓
MultiModule	✓	✓	

The above mentioned CRM predictors have different advantages. However, there is no study that integrates multiple predictors to improve CRM prediction. This study analyzed four existing CRM predictors (ClusterBuster, MSCAN, CisModule and MultiModule) to seek a predictor combination that delivers a higher accuracy than individual CRM predictors.

II. Method

A. Data Collection

This study collected CRMs from the Regulatory Element Database for Drosophila (REDfly) database, which is the most comprehensive database of *Drosophila melanogaster* CRMs [10], [11]. The REDfly version 3.0 contained 1,365 *D. melanogaster* CRMs and 3,446 TFBSs that collected from more than 200 articles.

Among the 1,365 CRMs, three redundant records and seven records that lacked the downstream genes were first removed. This study focused on promoter regions that (a) locate upstream a gene's start codon and (b) cover at least a CRM. 140 promoter regions of 2,000 bp were used as the positive set in this study. For the negative set, the 129 regions that (a) locate at 1 to 2,000 bp downstream the corresponding genes of the positive set and (b) cover no CRMs were used. As a result, the positive and negative sets were 140 and 129 DNA regions of 2,000 bp, respectively. The sequences in the corresponding regions were extracted from the FlyBase database [12], [13]. The FlyBase database integrates various genomic data, such as sequences, expressions and annotations, of *D. melanogaster*. The sequences of other *Drosophila* species were downloaded from the UCSC database [14], [15] as the closely related genomes for conservation computation.

The motifs of TFBSs in this study were obtained from the improved *Drosophila melanogaster* Major Position Matrix Motifs (iDMMPMM) [16] and JASPAR databases [17]. The

D. T.-H. Chang (corresponding author), G.-Y. Shiu, and Y.-J Sun are with the Department of Electrical Engineering, National Cheng Kung University, Tainan, 70101 Taiwan(fax:+886-6-2345482; e-mail: darby@mail.ncku.edu.tw, n26001571@mail.ncku.edu.tw, n26001644@mail.ncku.edu.tw).

iDMMPMM database version 2012, which is dedicated for *D. melanogaster*, contained37 TFBSs collected from DNase I footprinting, SELEX and/or ChIP-chip experiments. The JASPAR database is a comprehensive library that contains 1,316 TFBSs of 19 species with literature support. The JASPAR database version 4.0_ALPHA contained 108 *D. melanogaster* TFBSs. The union of the iDMMPMM and JASPAR included 122 TFBSs, which formed the TFBS collection in this study.

B. Predictor Integration

To combine multiple predictors for an integrated score, the first step is to recover the raw score of individual predictors. We traced the source code and inserted appropriate code to output raw scores. The resultant raw scores were not documented in the original papers of the adopted CRM predictors. The next step is to weight different raw scores. This study adopted the following two equations:

$$S(s_1, s_2) = \alpha \cdot s_1 + (1-\alpha) \cdot s_2 \text{ and} \tag{1}$$

$$S_z(s_1, s_2) = \alpha_z \cdot \frac{s_1-\mu_1}{\sigma_1} + (1-\alpha_z) \cdot \frac{s_2-\mu_2}{\sigma_2}, \tag{2}$$

where s_1 and s_2 are respectively the raw scores of two CRM predictors, μ_1 and μ_2 are respectively the means of the raw scores of the two CRM predictors, while σ_1 and σ_2 are respectively the standard deviations of the raw scores of the two CRM predictors. The two equations support only binary combinations. Tertiary combinations are more complex but did not help the prediction performance (data not shown). S(•) in (1) is simply a weighted sum of two raw scores. α is a weighting parameter, which was introduced to alleviate the scale difference between raw scores from different CRM predictors. S_z(•) in (2) is a weighted sum of Fisher's z-transformed scores (widely called z-scores), where α_z is a weighting parameter. Z-scores were introduced to standardize/normalize the raw score distribution, i.e.to make the distribution zero mean and unity standard deviation. In this study, the mean and standard deviation of a CRM predictor were the mean and standard deviation of the raw scores of the CRM predictor on the 269 sequences in the positive and negative sets.

II. Result

A. Evaluation Setting

In a conventional evaluation setting, every sample falls into one and only one of the following four outcomes: true positive (TP, positive sample correctly predicted as positive), false negative (FN, positive sample incorrectly predicted as negative), true negative (TN, negative sample correctly predicted as negative) and false positive (FP, negative sample incorrectly predicted as positive). This study contains 269 samples thus 269 predictions are expected. A sequence with no CRMs reported is considered as a negative prediction. However, in practice, CRM predictors might predict multiple regions in a sequence. Predicted CRMs in positive samples that do not overlap with any actual CRMs in REDfly made evaluation complex. They were incorrect positive predictions but they were in the positive samples. In this study, a fifth category, FP_{pos}, was introduced to represents such false positives in the positive samples. If a positive sample has *n* predicted CRMs, each predicted CRM contributes 1/*n* to TP or FP_{pos}, depending on whether the predicted CRM overlaps with any actual CRMs in REDfly. The remaining three outcomes (FN, TN and FP) are the same as their conventional definitions. This evaluation setting ensures equal contribution of every sample. Finally, in this study a predicted CRM is defined as overlapping with an actual CRM as follows:

$$\text{overlap}(p, a, o) = \begin{cases} true & \text{if } o > \min\left(\frac{p}{2}, \frac{a}{2}, 200\right) \\ false & \text{otherwise,} \end{cases} \tag{3}$$

where *p*, *a* and *o* are the lengths of the predicted CRM, the actual CRM and their overlap. Equation (3) ensures that the length of a valid overlap of two CRMs in this study exceeds half of the smaller one or 200 bp.

B. Evaluation Index

After defining TP, FP_{pos}, FN, TN and FP, various indices such as true positive rate (TPR) and false positive rate (FPR) can be used to evaluate the CRM predictors. This study adopted area under curve (AUC) as the evaluation index because AUC reveals an overall performance so that the trade-off between TPR and FPR can be neglected [18]. However, we observed a problem when applying conventional AUC directly in evaluating CRM predictors (Fig. 1). In Fig. 1, the solid and dotted lines represent two predictors. The AUC of the solid line (area of II+III) is smaller than that of the dotted line (III+V). However, the solid line is obvious better than the dotted one in the range that the solid line makes predictions (i.e. the range of FPR ≤ 0.3). The large AUC of the dotted line only comes from that the predictor of the dotted line prefers to report more predictions.

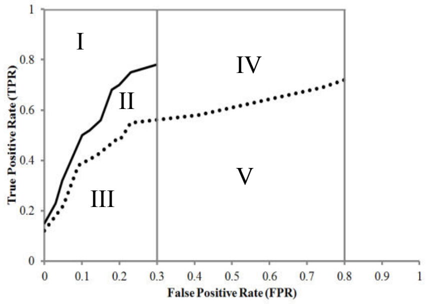

Fig. 1 Illustration of the proposed AUC%. The solid and dotted lines represent two predictors. The AUC of the solid and dotted lines are the area of II+III and III+V, respectively. The proposed AUC% is the AUC normalized by the theoretic maximum at the range that the predictor made predictions. The AUC% of the solid and dotted lines are (II+III)/(I+II+III) and (III+V)/(I+II+III+IV+V), respectively

In this study an AUC% was proposed to solve this problem. The AUC% of a predictor was the original AUC normalized by the theoretic maximum at the range that the predictor made predictions. Namely, the AUC% is the ratio of the predictor's AUC to a perfect predictor. In Fig. 1, the AUC% of the solid and dotted predictors are (II+III)/(I+II+III) and (III+V)/(I+II+III+IV+V), respectively. The AUC% also equals to the original AUC divided by the maximum FPR. With this adjustment, the solid line in Fig. 1 is superior to the dotted one, which is consistent with intuition.

C. Comparison with Existing Predictors

The performances of the four individual CRM predictors and their integrations are shown in Table III. To prevent overfitting due to parameter tuning, the results in Table III was based on (1) with $\alpha = 0.5$ (*i.e.*, the most trivial integration). The effects of weights and z-scores are discussed in the next subsection.

In Table III, the first four rows are individual CRM predictors. The best individual predictor was MSCAN, which achieved an AUC% of 16.02. The remaining six rows are integrated CRM predictors. Four integrated predictors delivered superior AUC% than MSCAN. The best integrated predictors was the integration of ClusterBuster and MultiModule, which achieved an AUC% of 17.31 (1.29 higher than MSCAN).In the four integrated predictors with better AUC% than MSCAN, three of them included MSCAN. This is reasonable because including a good predictor provides a good start point to improve. On the other hand, combining any of the three predictors with MSCAN improved MSCAN. This suggest that the strategy of integrating multiple CRM predictors is a promising direction and worthy of more efforts in future CRM studies.

An interesting observation is that the best integrated predictor did not include MSCAN. Actually, it was the integration of the worst two individual predictors. This is because that the two predictors captured distinct CRMs with different characteristics. Combining them largely increased the TPR with only a slightly FPR sacrifice. This complementarity is hard to predict. For example, replacing MultiModule with CisModule (they had comparable individual performances) in the best integration yielded a bad AUC% of 14.61. A further study to analyze the predictor complementarity is important so that effective integrations can be detected without exhausting evaluations.

TABLE III
COMPARISON WITH EXISTING CRM PREDICTORS

Predictor	AUC	AUC%
MSCAN	0.117	16.02
CisModule	0.140	13.97
MultiModule	0.134	13.52
ClusterBuster	0.061	8.80
MultiModule-ClusterBuster	0.171	17.31
MSCAN-CisModule	0.172	17.22
MSCAN-MultiModule	0.166	16.60
MSCAN-ClusterBuster	0.143	16.47
CisModule-ClusterBuster	0.146	14.61
CisModule-MultiModule	0.132	13.18

D.Effect of Different Integration Equations

This subsection focuses on the effects of using (1) and (2) and changing the weights in the equations. Fig. 2 shows the results, where only the four integrated predictors better than MSCAN were considered. In Fig. 2, S(•) (weighted raw scores, the solid lines) was superior to Sz(•) (weighted z-scores, the dotted lines)in three of the four integrations. In the integration of MultiModule-ClusterBuster (Fig. 2 (a)), Sz(•) was better than S(•) in some weights, but Sz(•) with the best αz was still worse than S(•) with the best α. This shows that the weighted sum of raw scores performed better than that of z-scores. Because the effect of Fisher's z-transformation is to alleviate the different means and standard deviations among predictors, this suggests that differences of means and standard deviations are useful information when integrating CRM predictors and should not be alleviated.

On the other hand, a linear weighting is required because the best weight was not close to 0.5 in most lines, solid or dotted. The exact weights are shown in Table IV. The extreme weights in MSCAN-CisModule and MSCAN-MultiModule show that the raw score of MSCAN dominated the two integrated predictors. But the improvement of 1.16 AUC% shown in Table IV and the rugged lines in Fig. 2 (c) indicate that MultiModule in the integration had a large effect, even with its small weight. Furthermore, the improvements of MultiModule-ClusterBuster and MSCAN-MultiModule (1.53 and 1.16 AUC%, respectively) in comparison with the improvements of integration shown in the previous subsection demonstrate the power of weighting. As a result, in integrating CRM predictors, the distributions of their raw scores should be preserved and the scale differences among them should be balanced with linear weighting.

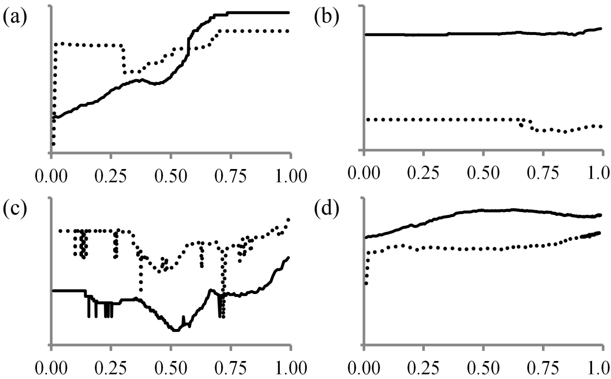

Fig. 2 Effect of different integration equations. Solid and dotted lines indicate S(•) in (1) and Sz(•) in (2), respectively. The y-axis is the AUC% while the x-axis is the weight (α for solid lines and αz for dotted lines).This figure includes four integrations: (a) MultiModule-ClusterBuster, (b) MSCAN-CisModule, (c) MSCAN-MultiModule and (d) MSCAN-ClusterBuster. The x-axis is the weight of the former predictor of an integration while the weight of the later predictor is 1 - x

TABLE IV
IMPROVEMENTS OF WEIGHTING

		Best Weighting		
Predictor	α= 0.5	α/α_z	AUC%	Improvement
MultiModule-ClusterBuster	17.31	0.73	18.84	1.53
MSCAN-CisModule	17.22	0.99	17.37	0.15
MSCAN-MultiModule	16.60	0.99	17.76	1.16
MSCAN-ClusterBuster	16.47	0.62	16.55	0.08

III. CONCLUSION

CRM plays a critical role in transcriptional regulation. It is important to predicting CRMs via computational methods to save time and cost of biological experiments. This study proposed an integrated predictor to improve CRM prediction and an evaluation setting to deal with the characteristics of CRM prediction. The experimental results show that the proposed predictor achieved superior performance to individual predictors.

REFERENCES

[1] Su, J., S. A. Teichmann, and T.A. Down, Assessing computational methods of cis-regulatory module prediction. PLoS Computational Biology, 2010. 6(12): p. e1001020.

[2] Davidson, E. H., The regulatory genome: gene regulatory networks in development and evolution. 2010: Academic Press.

[3] Kazemian, M., M. H. Brodsky, and S. Sinha, Genome Surveyor 2.0: cis-regulatory analysis in Drosophila. Nucleic Acids Res, 2011. 39(Web Server issue): p. W79-85.

[4] Levine, M. and E. H. Davidson, Gene regulatory networks for development. Proc Natl Acad Sci U S A, 2005. 102(14): p. 4936-42.

[5] Johansson, O., et al., Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. Bioinformatics, 2003. 19(Suppl 1): p. i169-i176.

[6] Zhou, Q. and W. H. Wong, CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci U S A, 2004. 101(33): p. 12114-9.

[7] Frith, M.C., Cluster-Buster: finding dense clusters of motifs in DNA sequences. Nucleic Acids Research, 2003. 31(13): p. 3666-3668.

[8] Zhou, Q. and W. H. Wong, Coupling hidden Markov models for the discovery of Cis -regulatory modules in multiple species. The Annals of Applied Statistics, 2007. 1(1): p. 36-65.

[9] Baum, L. E. and T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains. The annals of mathematical statistics, 1966. 37(6): p. 1554-1563.

[10] Gallo, S. M., et al., REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic Acids Res, 2011. 39(Database issue): p. D118-23.

[11] Gallo, S. M., et al., REDfly: a Regulatory Element Database for Drosophila. Bioinformatics, 2006. 22(3): p. 381-3.

[12] Drysdale, R. A. and M. A. Crosby, FlyBase: genes and gene models. Nucleic Acids Res, 2005. 33(Database issue): p. D390-5.

[13] Marygold, S. J., et al., FlyBase: improvements to the bibliography. Nucleic Acids Res, 2013. 41(Database issue): p. D751-7.

[14] Karolchik, D., The UCSC Genome Browser Database. Nucleic Acids Research, 2003. 31(1): p. 51-54.

[15] Fujita, P. A., et al., The UCSC Genome Browser database: update 2011. Nucleic Acids Res, 2011. 39(Database issue): p. D876-82.

[16] Kulakovskiy, I. V. and V.J. Makeev, Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. Biophysics, 2010. 54(6): p. 667-674.

[17] Portales-Casamar, E., et al., JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res, 2010. 38(Database issue): p. D105-10.

[18] Witten, I. H. and E. Frank, Data mining : practical machine learning tools and techniques. 2nd ed. Morgan Kaufmann series in data management systems. 2005, Amsterdam ; Boston, MA: Morgan Kaufman. xxxi, 525.