Bioinformatics, 33(1), 2017, 1–7
doi: 10.1093/bioinformatics/btw552
Advance Access Publication Date: 7 September 2016
Original Paper

OXFORD

Genome analysis

A novel method for predicting activity of cis-regulatory modules, based on a diverse training set

Wei Yang and Saurabh Sinha*

Department of Computer Science, University of Illinois, Urbana-Champaign, Urbana, IL, USA

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on February 28, 2016; revised on July 26, 2016; accepted on August 17, 2016

Abstract

Motivation: With the rapid emergence of technologies for locating cis-regulatory modules (CRMs) genome-wide, the next pressing challenge is to assign precise functions to each CRM, i.e. to determine the spatiotemporal domains or cell-types where it drives expression. A popular approach to this task is to model the typical k-mer composition of a set of CRMs known to drive a common expression pattern, and assign that pattern to other CRMs exhibiting a similar k-mer composition. This approach does not rely on prior knowledge of transcription factors relevant to the CRM or their binding motifs, and is thus more widely applicable than motif-based methods for predicting CRM activity, but is also prone to false positive predictions.

Results: We present a novel strategy to improve the above-mentioned approach: to predict if a CRM drives a specific gene expression pattern, assess not only how similar the CRM is to other CRMs with similar activity but also to CRMs with distinct activities. We use a state-of-the-art statistical method to quantify a CRM's sequence similarity to many different training sets of CRMs, and employ a classification algorithm to integrate these similarity scores into a single prediction of the CRM's activity. This strategy is shown to significantly improve CRM activity prediction over current approaches.

Availability and Implementation: Our implementation of the new method, called IMMBoost, is freely available as source code, at https://github.com/weiyangedward/IMMBoost.

Contact: sinhas@illinois.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene expression is regulated by the interaction of transcription factors (TFs) and their associated cis-regulatory sequences in the genome. These sequences, called 'cis-regulatory modules' (CRMs), harbor a collection of transcription factor-binding sites (TFBSs) that implement the 'logic' underlying precise spatiotemporal patterns of gene expression such as those seen in developing embryos (Davidson, 2001). A major challenge in the study of CRMs is to accurately differentiate genomic sequences that drive a specific spatiotemporal expression pattern from sequences that are either not functional or drive a different expression pattern. This 'CRM classification' problem is the subject of our study.

Powered by rapidly advancing sequencing technologies, several experimental techniques have recently emerged for regulatory element identification. These techniques include DNase-seq (Boyle *et al.*, 2008), ChIP-Seq (Visel *et al.*, 2009), FAIRE-seq (Giresi *et al.*, 2007) and ATAC-seq (Buenrostro *et al.*, 2013) among others, and have expanded the catalog of putative CRMs (mostly in human cell lines and model organisms) by orders of magnitude. Despite their widespread success (Bernstein *et al.*, 2010), these

techniques, which rely upon DNA accessibility profiles and epigenomic marks to locate CRMs, do not offer a solution to the more challenging problem of identifying the expression pattern driven by a CRM.

Computational approaches to predict CRMs associated with a regulatory network have gained in popularity as an alternative to experimental genome-wide profiling (Aerts, 2012). A common approach is to start with a set of known TFs from the network, along with their binding site motifs, and to search genome-wide for the putative binding regions (Blatti *et al.*, 2015; Frith *et al.*, 2003; Philippakis *et al.*, 2005; Sun *et al.*, 2012). However, this approach is often limited by missing knowledge of TFs relevant to the regulatory system and/or their binding site motifs. An alternative computational approach, called 'supervised CRM prediction', instead relies on a set of CRMs that drive a specific expression pattern to train a model that is used to search for other CRMs with similar activity (Kantorovitz *et al.*, 2009). This 'motif-blind' strategy does not rely on prior knowledge of TFs relevant to the regulatory network, or their binding site motifs. It is therefore widely applicable to a number of different regulatory networks and expression patterns, although it is less amenable to discovery of CRMs with pre-defined motif arrangements. We have also shown that it accurately predicts CRMs in less annotated genomes that are highly diverged from the species that the training CRMs belong to (Kazemian *et al.*, 2014). As an example of this approach, Lee *et al.* (2011) showed that support vector machine (SVM) trained on 'k-mer' (short word) frequencies in a training set can accurately predict regulatory sequences bound by EP300/CREBBP. Narlikar *et al.* (2010) built a linear regression model from sequence features to predict human heart CRMs. Kazemian *et al.* (2011) used an 'interpolated Markov model' (IMM) to learn from short words of varying length and identified CRMs that are active in specific expression domains in *Drosophila* species. An IMM combines Markov chains of different orders (up to 5 in our case), to account for the fact that k-mer signatures of TF binding sites in a CRM may not be described by a fixed order Markov model. Motif-blind approaches to CRM classification have been explored by several other authors as well (e.g. Ghandi *et al.*, 2014).

However, even the state-of-the-art methods of this genre are prone to false positives and may report CRMs that either drive gene expression in spatiotemporal domains different from the predicted domain (incorrect CRM activity prediction) or fail to drive gene expression completely (incorrect CRM prediction) (Kazemian *et al.*, 2011; Narlikar *et al.*, 2010). The goal of our study is to derive a more robust and predictive model that harnesses the benefits of current statistical methods for supervised CRM prediction, but alleviates the issue of false positives while not requiring additional data beyond DNA sequence information.

Supervised CRM prediction methods learn a model (sequence features) from a set of CRMs that drive a specific spatiotemporal expression pattern ('training data'), and then search for other CRMs matching that model. Methods seeking to improve this approach have explored the integration of experimental data beyond sequence (Ahmad *et al.*, 2014; Erwin *et al.*, 2014; Kleftogiannis *et al.*, 2015). However, without acquiring additional types of data, opportunities for CRM prediction improvement are limited. Here, we ask whether or not it is possible to improve the ability to predict CRMs of a specific activity (spatiotemporal expression pattern) by incorporating the knowledge of CRMs that drive other related or distinct expression patterns. To answer this question, we started with a previously published state-of-the-art supervised CRM prediction method (IMM), whose accuracy was demonstrated on a comprehensive benchmark of many different expression domains in *Drosophila* development. Here, we developed a 'meta-technique' on top of IMM, where a separate IMM model is trained on CRMs associated with each of 38 different expression domains, and the activity of a candidate CRM is predicted based on its match to each of those models. We call this meta-technique 'IMMBoost', since it is used here to boost the accuracy of the IMM method. (The term 'boost' is used with its common English meaning, not the specific technique of 'boosting' used in machine learning.) We note however that the same strategy can in principle be applied to potentially improve the performance of any other model of sequence features in a set of co-functional CRMs. Upon assessing how well the candidate CRM matches each of the training sets of co-functional CRMs ('CRM sets'), IMMBoost uses a classifier [SVM or Random Forest (RF), or both] to make its final decision about that CRM's activity.

We evaluated the performance of IMMBoost on two important tasks: (i) to differentiate CRMs active in a spatiotemporal expression domain from random non-coding regions; (ii) to discriminate CRMs that are active in a specific expression domain from CRMs that drive expression in other domains. Cross validation was used to show that IMMBoost improves predictive power over IMM, for both tasks. We also built an ensemble model that combines IMMBoost and IMM to further increase prediction accuracy. A head-to-head comparison between this ensemble model and another state-of-the-art k-mer based method (Lee *et al.*, 2011) further demonstrated the advantage of learning from CRMs that are active in other expression domains. Closer examination of the parameters of trained IMMBoost models revealed that in several cases the knowledge and use of CRMs with expression patterns different from that of the candidate CRM play a critical role in accurate activity assignment for that CRM. This observation provides further evidence for the basic premise of our new technique: that to predict if a CRM drives a specific expression pattern, it helps to assess how similar the CRM is to other CRMs with that pattern and also to CRMs with distinct patterns.

2 Materials and methods

2.1 Data sets

We defined 38 CRM sets, each comprising *Drosophila melanogaster* CRMs that drive expression in a common domain, largely by borrowing from (Kazemian *et al.*, 2014), with minor updates based on the REDfly database (Supplementary Materials S1). A CRM set is named by its expression domain, with prefix 'mapping1' or 'mapping2' added to refer to different levels of specificity in an expression domain ('mapping2' represents more broadly defined domains than 'mapping1'). Each data set consisted of a CRM set ('positive' examples) and 100 sequences ('negative' examples) randomly sampled from DNA-accessible, inter-genic regions with lengths and G/C content similar to the CRMs. DNA-accessibility information was taken from BDTNP Chromatin Accessibility (DNase) peaks at stage 5 of *D. melanogaster* development from UCSC Tracks (http://genome.ucsc.edu/) and repeats in these regions were masked using Tandem Repeats Finder (Benson, 1999). Additionally, we ensured the sampled regions did not overlap with any of the existing training CRMs or annotated genes for *D. melanogaster*. Note that the above positive and negative examples were used both for training classifiers and testing them, in a cross-validation scheme. We chose DNA-accessible regions for negative data since they include CRM-like features such as TF-binding motifs, that target the ability of an algorithm to distinguish sequences with specific motif architectures as opposed to any motif architecture (Arvey *et al.*, 2012). To generate negative examples of CRMs from

non-target domains, we randomly picked 100 unique *D. melanogaster* CRMs from domains different from the target domain. We further obtained orthologous CRM sequences from 10 other *Drosophila* species using LiftOver program from the UCSC Genome Browser web site with minMatch = 0.25 and other parameters set to default values. During classifier training, sequence features of any example (positive or negative) were obtained from the sequence of that CRM or random segment, along with its orthologous sequences from other *Drosophila* species. During testing, sequence features were taken from *D. melanogaster* segments only. We ensured that the same CRM does not appear both in the training and test examples, as might happen in some cases when a CRM belongs to two expression domains. All data sets are made available in the Supplementary Materials S1 and S2 at https://github.com/weiyangedward/IMMBoost.

2.2 Training models

For each expression domain d, two fifth-order IMMs were trained as in (Kazemian *et al.*, 2011): one on the positive examples (IMM^{d+}) and one on negative examples (IMM^{d-}). We used these two trained IMMs to compute the 'IMM score' of any sequence S, representing how well it matches the expression domain d:

$$\mathrm{IMM}(S,d) = \log \frac{P(S \mid \mathrm{IMM}^{d+})}{P(S \mid \mathrm{IMM}^{d-})} \quad (1)$$

In the second phase of IMMBoost, each positive or negative example S was represented by a 38-dimensional vector (since there are 38 CRM sets), where the dth dimension is the score IMM*(S,d)*. When performing cross-validation on a target domain t, we took care to exclude the test sequences of domain t from the training sets used to learn models IMM^{t+} and IMM^{t-}. To learn the IMMBoost-SVM model, we trained an SVM classifier via the LIBLINEAR package (Fan *et al.*, 2008) using a linear kernel with L2-regularization and L2-loss function. The best parameter for cost ('-c') was selected based on a grid search over training data. IMMBoost-RF was trained using the R package 'randomForest' (Liaw and Wiener, 2002). We trained 1000 decision trees to perform classification, where each tree had $\sqrt{38} \approx 6$ nodes and each leaf node had a minimum support of 1. To compare IMMBoost variants to a baseline other than the IMM, we implemented a 'kmer-SVM' classifier, using the frequencies of all k-mers ($k = 6$) in the sequence as feature vector (Lee *et al.*, 2011). We used the LIBSVM package (Chang and Lin, 2011) for this implementation, and chose a Gaussian kernel. The best parameters for gamma '-g' and cost '-c' were selected according to a grid search on training data. To build the IMMBoost-Ensemble model, we first performed a min-max normalization of the prediction score from IMM, IMMBoost-RF and IMMBoost-SVM, and then summed over the normalized scores of a test sequence to obtain the final prediction score.

2.3 Identify strong spatiotemporal domain predictors

To determine which expression domains were the main contributors in predicting CRM activity in a specific target domain, we quantified the contribution of each predictor using feature importance in IMMBoost-RF results. The feature importance is a z-score of the average decrease of prediction accuracy among all trees due to permuting one feature at a time, and was reported by the R package used for this task. We averaged feature importance over 50 data sets (10 repeats of 5-fold cross validation) for each predictor and normalized all predictors to a scale of 0–1. Expression domains with a normalized feature importance larger than 0.5 were considered to be strong predictors.

2.4 Clustering of spatiotemporal expression domains by CRMs

After counting the number of CRMs that were shared between two expression domains, we normalized this count by the size of the smaller set, obtaining the 'overlap coefficient' (Supplementary Table S3). The formula is the following, where A and B are the CRM sets representing two expression domains:

$$\mathrm{overlap}(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \quad (2)$$

Altogether, each expression domain was defined by a 38-dimensional vector of overlap coefficients. We performed a hierarchical clustering of all domains represented by these overlap vectors. In the clustering results, the distance between two expression domains indicates the similarity of their CRM sets.

3 Results

3.1 IMMBoost: a supervised CRM prediction method trained on heterogeneous training data

Given 'training sets' of CRMs from different expression domains, the goal is to learn to predict activity in each expression domain, i.e. whether any given sequence drives expression in that domain. Normally, supervised CRM prediction methods (Kazemian *et al.*, 2011) perform this task for each domain separately, by learning sequence features of CRMs of that domain. IMMBoost also learns to predict each domain separately, but does so by examining sequence features of CRMs of that domain as well as any other expression domains for which CRMs are available.

An outline of IMMBoost is shown in Figure 1 (For a more detailed description, see 'Materials and methods' section.). The algorithm first constructs separate IMMs from each expression domain's training set; the positive data for any individual IMM model corresponds to the CRMs that pattern genes in the relevant expression domain, while the negative data corresponds to randomly selected, accessible non-coding genomic sequences or to CRMs from all other spatiotemporal domains (Fig. 1A). The expression domain-specific IMM is trained based on frequencies of k-mers in the positive and negative data (sequences) for that domain, and can be used to assign a score (log likelihood ratio) to any given sequence. After this initial step of training IMMs for each expression domain, a given sequence is mapped to a 'feature vector', where each entry of the vector is the score of the sequence under the IMM for a specific expression domain (each column in Fig. 1B). In other words, the sequence is represented by its similarity to CRMs of each expression domain. Finally, a SVM or RF classifier is trained for an expression domain using the above-mentioned feature vector for each training example of that domain (Fig. 1C) (This classifier-training step is performed separately for each expression domain.).

We denote the SVM and RF versions of IMMBoost as 'IMMBoost-SVM' and 'IMMBoost-RF' respectively. By incorporating the IMM scores for all expression domains, the classifiers integrate information from multiple expression domains in deciding if a sequence is capable of driving gene expression in the target expression domain. Whether we wish to predict membership in one expression domain or another, the vector representation of the test sequence does not change, only the weights learned by the RF or SVM change to reflect the changing classification task.

Special attention was devoted to the type of negative data used to train each domain-specific IMM, depending on the classification task. When trying to distinguish CRMs in a particular expression domain

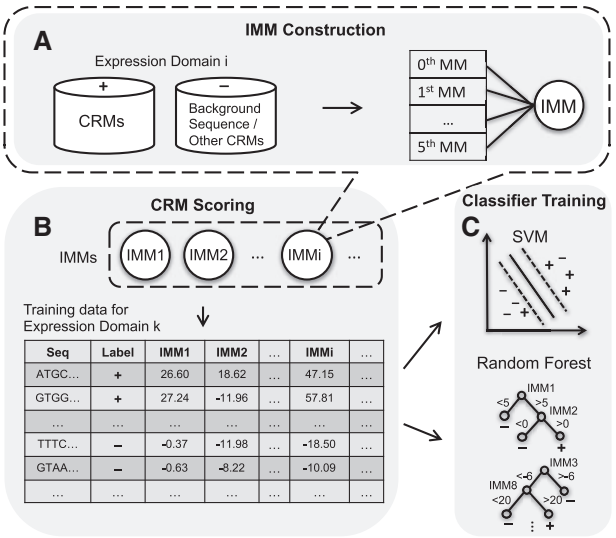

Fig. 1 Overview of IMMBoost. **(A)** We first train fifth order IMMs for an expression domain using CRMs that drive gene expression in the domain as positive data, and negative data as either random, accessible, inter-genic sequences or CRMs from other expression domains. **(B)** We apply all trained IMMs to score a sequence for similarity to each domain, and construct a feature vector out of these IMM-based similarity scores. '+' labels indicate a positive example for the expression domain k while '−' indicates a negative example for the same domain. **(C)** We train a SVM or a RF classifier on the training sequences of the expression domain k using IMM-based feature vectors

from non -functional sequences, IMMBoost uses randomly selected, length-matched, DNA-accessible inter-genic sequences as negative data. However, when learning to discriminate whether a CRM drives expression in one expression domain versus other domains, it uses CRMs from other spatiotemporal domains as negative data, while excluding domains similar to the target domain. The two tasks superficially appear similar, but distinguishing CRMs from non-CRM sequences is generally easier than distinguishing CRMs between domains, as the CRMs might have similar regulators and binding sites at the sequence level.

3.2 IMMBoost improves classification accuracy significantly

We evaluated IMMBoost-SVM and IMMBoost-RF for their ability to predict CRM activity in each of 38 different spatiotemporal expression domains in *Drosophila*, as catalogued by Kazemian *et al.* (2014) (Each data set corresponds to an expression domain and comprises between 6 and 109 CRMs, and appropriate negative data.) Although the CRMs for each expression domain were derived from *D. melanogaster*, the classifier training phase examined these CRMs as well their orthologs from 10 closely related *Drosophila* species, to better learn sequence features. (For more detail, see 'Materials and methods' section.) Ten random trials of five-fold cross validation were conducted for each expression domain, and performance was measured by the area under the curve (AUC) of the receiver operating characteristic (ROC). The mean AUC across all 10 random trials was used as the final evaluation metric for any given expression domain and prediction method. As a baseline, we used the IMM model trained on the expression domain being tested: in previous work we have shown the IMM to be as good as or better than several alternative methods for this task, including motif-based methods and several 'motif-blind' statistical methods (Kazemian *et al.*, 2011). Moreover, using IMM as the baseline allowed us to evaluate the contribution of the 'meta' strategy of IMMBoost, while keeping the underlying scoring method the same.

First, we compared IMMBoost with the baseline IMM method for the task of discriminating CRMs of a specific expression domain from random inter-genic sequences of similar length and G/C content as the CRMs, located in accessible regions of DNA. Both IMMBoost-SVM and IMMBoost-RF outperform IMM in this regard, with AUCs of 0.711, 0.730 and 0.662 respectively on average across the 38 expression domains (Fig. 2A, panels 1 and 2). The IMMBoost variants showed significant improvement over IMM for several expression domains. The most dramatic improvement was observed on a data set of 'malpighian tubule' CRMs (Supplementary Table S1), with IMMBoost-SVM and IMMBoost-RF AUCs of 0.712 and 0.655, respectively, with the baseline IMM failing completely on this data set (AUC 0.412).

Next, we compared IMMBoost variants with IMM on the more challenging task of discriminating CRMs of a specific expression domain from CRMs of other domains. Again, IMMBoost-SVM and IMMBoost-RF outperformed IMM, with AUCs (averaged across the 38 different classification tasks) of 0.667, 0.680 and 0.654, respectively (Fig. 2A, panels 3 and 4). The margin of improvement though was smaller than the results noted in the previous paragraph, and the baseline IMM model outperformed IMMBoost-SVM and IMMBoost-RF for certain domains, such as 'eye', 'glia', 'dorsal ectoderm' and 'endoderm' (Supplementary Table S2). Nevertheless, for the majority of expression domains, the IMMBoost methods improved classification accuracy.

In order to further improve predictive performance, we constructed an ensemble classifier that integrates the predictions of IMMBoost-SVM, IMMBoost-RF and IMM, by reporting the sum of normalized prediction scores from each model (see 'Materials and methods' section). This new method is called IMMBoost-Ensemble. For the task of discriminating CRMs of a given domain from random non-coding sequences, IMMBoost-Ensemble (average AUC 0.731) showed significant improvements over the baseline IMM method (average AUC 0.662) (Fig. 2B, panel 1). Also, seven additional expression domains were tractable when using IMMBoost-Ensemble (domains have AUC $\geq$ 0.6 under IMMBoost-Ensemble and AUC $<$ 0.6 under IMM): malpighian tubules, female gonad, mesectoderm, eye, glia ('mapping1.glia' and 'mapping2.glia'), and adult mesoderm (Supplementary Table S1).

More importantly, IMMBoost-Ensemble was clearly the best of the evaluated approaches for the task of discriminating CRMs of a given domain from CRMs of other domains. It yielded an AUC measure of 0.700 on average over the 38 data sets, marking an improvement over all three other methods, IMMBoost-RF (AUC 0.680), IMMBoost-SVM (AUC 0.667) and the baseline IMM (AUC 0.654). Again, seven additional expression domains proved amenable to supervised CRM prediction when using IMMBoost-Ensemble: malpighian tubules, ventral ectoderm, mesectoderm, larva, tracheal system, male gonad and cns (Supplementary Table S2). A head-to-head comparison between the ensemble method and baseline IMM showed the former to be as or more accurate than the latter on nearly every data set (Fig. 2B, panel 3), with AUC improvements $\geq$ 0.05 on 15 of 38 data sets. For the task of distinguishing CRMs of a given domain from CRMs of other domains, some of the greatest improvements were observed on the expression domains 'malpighian tubules', 'ventral ectoderm', 'mesectoderm' and 'dv neurogenicectoderm' (Fig. 2C).

Our comparisons above used the single-class IMM as baseline because we had observed this method to be the most effective in our previous work Kazemian *et al.* (2011). To compare IMMBoost-Ensemble to a second baseline, we trained and evaluated another state-of-the-art

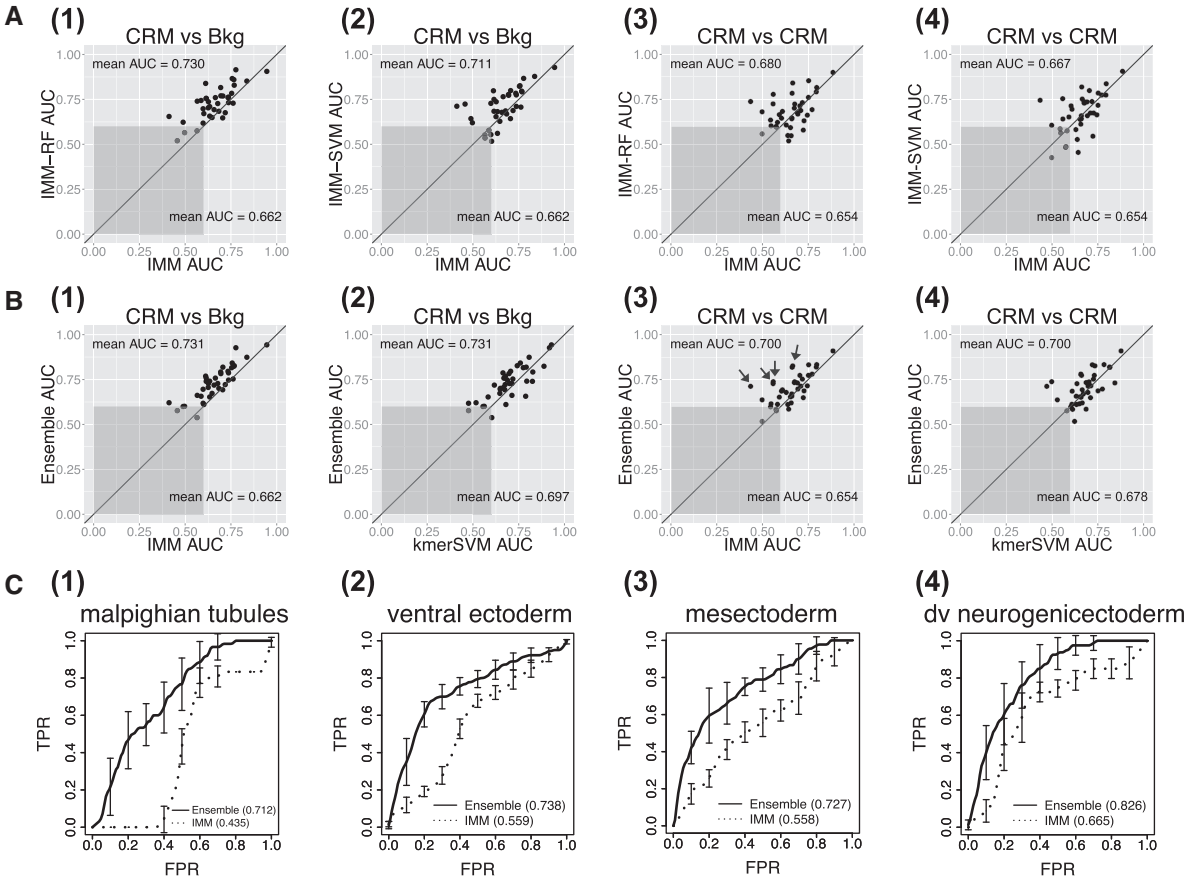

Fig. 2 IMMBoost improves supervised CRM prediction. **(A)** IMMBoost outperforms IMM in predicting CRMs. In each scatter plot (panels 1–4), each dot represents an expression domain and the AUC (average over ten 5-fold cross validation) of IMMBoost-RF ('IMM-RF') or IMMBoost-SVM ('IMM-SVM') is compared with AUC of the baseline IMM method. Comparisons are shown for the task of classifying CRMs of one expression domain versus random, accessible, inter-genic sequences (panels 1 and 2: 'CRM versus Bkg') and for task of discriminating CRMs of one domain from CRMs that are active in other domains (panels 3,4: 'CRM versus CRM'). The darker grey area indicates that both compared classifiers have AUCs < 0.6. Mean AUC across all 38 domains are shown for both methods (top left and bottom right corner). **(B)** IMMBoost-Ensemble outperforms baseline IMM and k-mer SVM in predicting CRMs. Panels have similar semantics to those in (A), except that here the AUC of IMMBoost-Ensemble (y axis) is compared with either the baseline IMM (panels 1 and 3) or the k-mer SVM (panels 2 and 4). Four expression domains with the most prominent improvement in IMMBoost-Ensemble over IMM are pointed out (arrows) in panel 3. **(C)** Accuracy on the aforementioned four expression domains is shown as ROC Curves. Each curve is an average of 10 5-fold cross validation results, where error bars denote standard deviation among the ten tests. The solid and dotted curves correspond to IMMBoost-Ensemble and (baseline) IMM models, respectively. Numbers within parentheses are AUC scores

algorithm for predicting CRMs, proposed by Lee *et al.* (2011). It considers k-mer frequencies in training CRMs, similar to the IMM method, but unlike the latter it employs an SVM that uses feature vectors comprising k-mer frequencies. This k-mer SVM approach trains the classifier solely based on the class of CRMs it hopes to predict, in a manner similar to our baseline IMM method. We implemented a k-mer SVM classifier with a Gaussian kernel we used the same training and evaluation scheme as that used for the IMM-based methods (The results in Lee *et al.*, 2011) were based on an SVM with spectrum kernel, but the authors had noted that the Gaussian kernel exhibits similar accuracy.). We found IMMBoost-Ensemble to be significantly more accurate than our implementation of k-mer SVM for both tasks discussed above (Fig. 2B, panels 2 and 4), once again demonstrating the advantage of utilizing CRMs from multiple expression domains in training predictors of a specific class of CRMs.

3.3 Characteristics of strong predictors

Our tests above demonstrate that when predicting CRM activity in a specific expression domain (the 'target' domain), it helps to consider if the test sequence is similar to CRMs of other expression domains as well. We posited that there must be some concordance between the CRM sequences of the target domain and those of any domains that were successfully exploited by the classifier ('predictor' domains). One possibility is that the predictor domains are biologically similar to the target domain, making them suitable for learning sequence features of the target domain. It is also possible that some predictor domains are highly distinct from the target domain and the classifier learns to recognize a CRM's activity in the target domain based on its dissimilarity from CRMs of the predictor domains.

To investigate the above possibilities, we first objectively recorded the overall similarity among expression domains. This was performed by hierarchical clustering of expression domains based on the number of CRMs shared between two domains (Supplementary Table S3, see 'Materials and methods' section, 'overlap coefficient'). As expected, the clustering results were often consistent with known biological relationships among expression domains (Supplementary Fig. S1), e.g. expression domains 'cardiac mesoderm', 'visceral mesoderm' and 'somatic muscle', related to muscle development, were

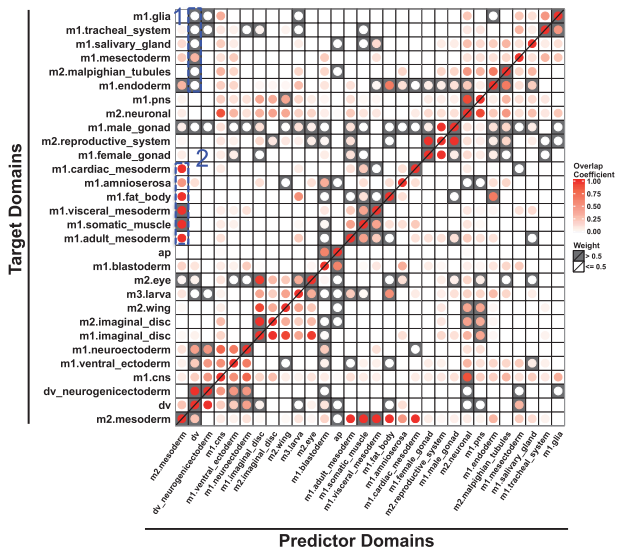

Fig. 3 IMMBoost learns from diverse expression domains to predict expression in a target domain. Each row is a target domain (CRMs are being predicted for this domain) and each column represents a predictor domain (sequence similarity to these domains is used by the classifier). We only show target domains with mean IMMBoost-RF AUC > 0.6, in order to limit our study to accurately predicted domains. The prefixes 'm1' and 'm2' are short for 'mapping1' and 'mapping2'. Rows and columns are ordered according to the hierarchical clustering of domains, based on their mutual overlaps (Supplementary Fig. S1), so that biologically similar expression domains are shown proximally. Similarity of two domains (normalized overlap coefficient, ranging from 0 to 1, see 'Materials and methods' section) is indicated by intensity of red circle in the cell. If a predictor domain is a strong predictor of a target domain (normalized feature importance > 0.5), the cell has a black border. Box 1 is an example showing how distantly related expression domains can be strong predictors (see text). Box 2 is an example showing that a predictor domain ('m2.mesoderm') may not strongly contribute to the prediction of a target domain despite having a large overlap coefficient with that target domain (four of the six cells in this box have a red circle but no black border)

grouped together by our clustering procedure. We then marked information about strong predictor domains for each target domain on top of the clustering results (Fig. 3, Supplementary Table S4 and S5). To evaluate the importance of a predictor domain, we extracted the feature importance from the RF classifier, computed by permuting the values of the feature and noting the decrease in predictive accuracy (see 'Materials and methods' section).

Focusing on the 30 domains where the classifier (IMMBoost-RF) shows AUC > 0.6, we first noted, as expected, that the target domain is a good predictor of itself in the majority of cases (black squares on diagonal, Fig. 3). Surprisingly, we found that for 11 of these 30 domains, the target domain is not a strong predictor of itself (white squares on diagonal, Fig. 3), i.e. in these cases the classifier for a specific target domain chose not to be guided by sequence features of CRMs of that domain. In some cases, the strong predictor domains are close to the target domain as per our hierarchical clustering. For example, CRMs of expression domain 'dv' (dorsal–ventral patterning) were strong predictors of activity in the domain 'dv_neurogenicectodern', with our clustering results suggesting that these two expression domains share many CRMs and are biologically similar. This implies that similar composition can enable a domain to be a good predictor for biologically related domains. On the other hand, a number of strong predictor domains are distantly related to their target domains (black squares off diagonal in Fig. 3), which suggests that domains with dissimilar CRM compositions can be as informative than ones that are similar to the target domain. For example, even though the expression domain 'dv' does not share many CRMs in common with the domains 'm1.glia', 'm1.tracheal_system', 'm1.salivary_gland', 'm2.malpighian_tubules' and 'm1.endoderm', activity in these domains is strongly predicted by comparison of sequence features to 'dv' CRMs (Fig. 3, Box 1).

Investigating the above matter further, we considered whether the strong predictive power of a predictor domain results simply from a large number of training CRMs known for that domain, possibly leading to more reliable characterization of sequence features representing that domain. We found that the domain 'dv' is a strong predictor for 16 expression patterns though it only has 19 CRMs in its training set. On the other hand, the largest CRM training set, 'm2.mesoderm', has 109 CRM, but only strongly predicts six target domains; furthermore, similarity to 'm2.mesoderm' CRMs does not strongly predict activity in domains 'm1.cardiac_mesoderm', 'm1.amnioserosa', 'm1.adult_mesoderm' or 'm1.fat_body' despite the fact that the 'm2.mesoderm' CRM set overlaps strongly with the CRM sets of these target domains (Fig. 3, Box 2). Overall, we did not find persuasive evidence that larger size of its CRM set makes an expression domain strongly predictive of more target domains. However, we did observe an effect of CRM set size on the extent of improvement afforded by IMMBoost-Ensemble relative to the baseline IMM method, as explained next.

We carefully examined the improvement of IMMBoost over baseline IMM across expression domains. First we noted that the improvement due to IMMBoost-Ensemble is highly correlated with the improvement due to IMMBoost-SVM as well as that due to IMMBoost-RF (Supplementary Fig. S5A and B), so that we could focus our attention on these methods. We then computed the difference between training and test data accuracy estimates (during cross validation) as a rough measure of 'over-fitting', for each data set and found a significant negative correlation between the extent of over-fitting and the gain in performance over IMM, both for IMM-SVM (Supplementary Fig. S5C) and for IMM-RF (Supplementary Fig. S5F). This supports the view that IMMBoost fails to improve on the baseline IMM on data sets where there is stronger evidence of over-fitting. An obvious possible cause of over-fitting is a smaller training set. Indeed, we observed correlations between the extent of over-fitting and CRM set size (Supplementary Fig. S5D and E). Overall, we conclude that with larger size CRM sets, IMMBoost-SVM and IMMBoost-RF, and as a result IMMBoost-Ensemble, tend to provide greater improvement over the baseline IMM method.

4 Discussion

We have presented IMMBoost, a new supervised learning method to predict CRMs active in a specific expression domain. The problem of supervised CRM prediction has been addressed by several previous models, such as IMM and kmer-SVM (Kazemian *et al.*, 2011; Lee *et al.*, 2011). In contrast to these previous CRM identification strategies, which learn from the known CRMs of one spatiotemporal expression domain, IMMBoost has the ability to integrate knowledge about active CRMs in other available domains as well, while automatically determining which other domains are informative for predicting activity in the target domain. More specifically, IMMBoost is a two-stage learning method: the first stage uses the previously demonstrated power of the IMM score for characterizing the similarity of a sequence to a class of CRMs, while the second stage uses classifiers to integrate IMM-based similarities of the test sequence to a number of different CRM classes. This second stage

and its underlying principle are the main novel contribution of our work. Our evaluation results suggest that IMMBoost leads to solid improvement over the previous best methods in predicting CRMs that are active in a specific gene regulatory network.

Our new method significantly increases the accuracy of classifying CRMs versus random non-coding sequences (average AUC of 0.731 in IMMBoost-Ensemble, compared with 0.662 in IMM), and makes seven more expression domains amenable to supervised CRM prediction compared with the state-of-the-art. Moreover, it significantly improves performance on the harder task of discriminating CRMs specific to one regulatory network from CRMs of other networks (average AUC of 0.700 in IMMBoost-Ensemble, compared with 0.654 in IMM), and it again makes seven additional domains amenable to supervised CRM prediction. (The AUC numbers summarized in this paragraph are averaged over 38 expression domains, and over 10 trials of 5-fold cross validation for each domain.)

We note that the benchmark used here (obtained from Kazemian *et al*., 2014) is one of the most comprehensive benchmark for evaluating CRM activity prediction methods: the CRMs and their associated regulatory functions have been directly tested through in vivo reporter assays rather than being based on circumstantial evidence such as epigenomic marks, accessibility and expression data on nearby genes. Importantly, the benchmark includes dozens of different expression domains, each under control of its own regulatory network. Our demonstration of the added value of the new meta-strategy across this diverse benchmark therefore speaks to the generalizability of the approach.

By investigating the relative importance of features used by the IMMBoost-RF classifier, we found, as expected, that to predict CRM activity in a specific expression domain it helps to learn from other CRMs active in that domain or in biologically similar domains. On the other hand, we also found that distinct expression domains can be informative in predicting activity in the target domain, vindicating the basic premise underlying the new strategy we pursued in this work. We expect that as the knowledge base of CRMs and their diverse activities grows in the future, the new strategy proposed here will prove ever more useful in predicting novel CRMs, especially in genomes where experimental data suggesting their locations are less abundant.

Funding

This work was supported in part by USDA (grant 2012-67013-19361, PI: M.S. Halfon, Subaward R775499 to S.S.), the NIH (grant R01GM114341 to S.S.) and by the Simons Foundation (Grant 'Molecular Roots of Social Behavior', PI: L. Stubbs and G.E. Robinson). We thank Marc S. Halfon and Kushal Suryamohan for sharing data about CRM sets and their biological relationships.

Conflict of Interest: none declared.

References

Aerts,S. (2012) Chapter five - computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Current Topics in Developmental Biology, Transcriptional Switches during Development*. Academic Press, **98**, 121–145.

Ahmad,S.M. *et al*. (2014) Machine learning classification of cell-specific cardiac enhancers uncovers developmental subnetworks regulating progenitor cell division and cell fate specification. *Development*, **141**, 878–888.

Arvey,A. *et al*. (2012) Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome Res*., **22**, 1723–1734.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*., **27**, 573–580.

Bernstein,B.E. *et al*. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol*., **28**, 1045–1048.

Blatti,C. *et al*. (2015) Integrating motif, DNA accessibility and gene expression data to build regulatory maps in an organism. *Nucleic Acids Res*., **43**, 3998–4012

Boyle,A.P. *et al*. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

Buenrostro,J.D. *et al*. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

Chang,C.C. and Lin,C.J. (2011) LIBSVM. A library for support vector machines. *ACM Trans. Intell. Syst. Technol*., **2**, 27:1–27:27.

Davidson,E.H. (2001) *Genomic Regulatory Systems: In Development and Evolution* Academic Press, San Diego.

Erwin,G.D. *et al*. (2014) Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput. Biol*, **10**, e1003677.

Fan,R.E. *et al*. (2008) LIBLINEAR: a library for large linear classification. *J Mach. Learn. Res*., **9**, 1871–1874.

Frith,M.C. *et al*. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*., **31**, 3666–3668.

Ghandi,M. *et al*. (2014) Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol*., **10**, e1003711.

Giresi,P.G. *et al*. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res*., **17**, 877–885.

Kantorovitz,M.R. *et al*. (2009) Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. *Dev. Cell*, **17**, 568–579.

Kazemian,M. *et al*. (2014) Evidence for deep regulatory similarities in early developmental programs across highly diverged insects. *Genome Biol. Evol*., **6**, 2301–2320.

Kazemian,M. *et al*. (2011) Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. *Nucleic Acids Res*., **39**, 9463–9472.

Kleftogiannis,D. *et al*. (2015) DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res*., **43**, e6–e6.

Lee,D. *et al*. (2011) Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res*., **21**, 2167–2180.,

Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R. News*, **2**, 18–22.

Narlikar,L. *et al*. (2010) Genome-wide discovery of human heart enhancers. *Genome Res*., **20**, 381–392.

Philippakis,A.A. *et al*. (2005) Modulefinder: a tool for computational discovery of cis regulatory modules. *Pac. Symp. Biocomput*., 519–530.

Sun,H. *et al*. (2012) Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection. *Nucleic Acids Res*., **40**, e90–e90.

Visel,A. *et al*. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.