Prediction of Binding Sites in the Mouse Genome Using Support Vector Machines

Yi Sun [1*], Mark Robinson [2*], Rod Adams [1*], Alistair Rust[3*], and Neil Davey [1*]

[1*] Science and technology research school, University of Hertfordshire,
United Kingdom, AL10 9AB
{comrys, r.g.adams, n.davey}@herts.ac.uk
[2*]Department of Biochemistry and Molecular Biology, Michigan State University,
East Lansing MI 48824, USA
blobby@msu.edu
[3*]Institute for Systems Biology,
1441 North 34th Street,
Seattle, WA 98103, USA
arust@systemsbiology.org

Abstract. Computational prediction of *cis*-regulatory binding sites is widely acknowledged as a difficult task. There are many different algorithms for searching for binding sites in current use. However, most of them produce a high rate of false positive predictions. Moreover, many algorithmic approaches are inherently constrained with respect to the range of binding sites that they can be expected to reliably predict. We propose to use SVMs to predict binding sites from multiple sources of evidence. We combine random selection under-sampling and the synthetic minority over-sampling technique to deal with the imbalanced nature of the data. In addition, we remove some of the final predicted binding sites on the basis of their biological plausibility. The results show that we can generate a new prediction that significantly improves on the performance of any one of the individual prediction algorithms.

1 Introduction

In this paper, we address the problem of predicting transcription factor (TF) binding sites (binding motifs) within sequences of regulatory DNA. Currently, experimental methods for characterising the binding sites found in regulatory sequences are both costly and time consuming. Computational predictions are therefore often used to guide experimental techniques. Computational prediction of *cis*-regulatory binding sites is widely acknowledged as a difficult task [12]. Binding sites are notoriously variable from instance to instance and in higher eukaryotes they can be located considerable distances, both upstream and downstream, from the gene being regulated.

There are many different algorithms for searching for binding sites in current use, such as those proposed in [1] and [2]. However, most of them produce a high rate of false positive predictions. The use of algorithmic predictions prone to high rates of false positives is particularly costly to experimental biologists using the predictions to guide experiments. Moreover, many algorithmic approaches are inherently constrained with respect to the range of binding sites that they can be expected to reliably predict. Given the differing aims of these algorithms it is reasonable to suppose that an efficient method for integrating predictions from these diverse strategies should increase the range of detectable binding sites. Furthermore, an efficient integration strategy may be able to use multiple sources of information to remove many false positive predictions, while also strengthening our confidence about many true positive predictions. In [6], five popular motif discovery algorithms are run multiple times with different parameters, then multiple results are collected and grouped by a score rank. The final predictions are obtained based on voting, smoothing and extracting methods. In [7], a software tool, *MultiFinder*, was developed. It performs automated motif searching using four different profile-based motif finders (algorithms), and results from each motif finder are ranked according to the user specified scoring function. The user can select any combination of motif prediction tools.

The nature of the problem allows domain specific heuristics to be applied to the classification problem. Instead of applying voting as discussed in [6], and merging multiple predictions according to the user specified scoring function mentioned in [7], we attempt to reduce these false positive predictions using classification techniques taken from the field of machine learning. In [10] and [11], we found that the integrated classifier, or *meta classifier*, when using a support vector machine (SVM) [9] outperformed each of the original prediction algorithms. In particular the integrated classifier has a better tradeoff between recall and precision.

In this work, we extend our work in [11] by making a major change to the way the training sets are constructed. Previously we have only used proximal annotated DNA sequences close to a gene as both positive and negative examples of binding sites. However a potential problem with this approach is that the nucleotides labelled as not being part of a binding site may be incorrectly labelled, due to unreliable biological evidence. Here we introduce a new *background* dataset which draws negative examples from sequences that are 5000-4500 base pair (bp) away from any gene. In this way we hope to ensure that our negative examples are much less likely to be regulatory.

We use a 6-ary real valued vector, each element of which is a prediction result from one of the algorithms, for a particular nucleotide position, as the input of the system. The data consists of a merger of promoters from the mouse genome (*M.musculus*), annotated with transcription factor (TF) binding sites taken from the *ABS*[1] and *ORegAnno*[2] databases. In total there are 47 promoter sequences (regulatory region containing transcriptional start site), including 142 TF binding sites. The data also includes 250 upstream, non-coding sequences from which negative examples may be taken (*background*). The background sequences were extracted using the UCSC genome website[3].

In this work, one challenging aspect is the imbalanced nature of the data and that has led us to explore some powerful techniques to address this issue. The data has two classes: either binding sites or non-binding sites, with about 97% being non-binding sites. We combine random selection under-sampling and SMOTE [3] over-sampling

[1] http://genome.imim.es/datasets/meta2005/index.html

[2] http://www.oreganno.org/oregano/Index.jsp

[3] http://genome.ucsc.edu/

techniques. In addition, we remove some of the final predicted binding sites on the basis of their biological plausibility. The proposed method can be seen in Figure 1.

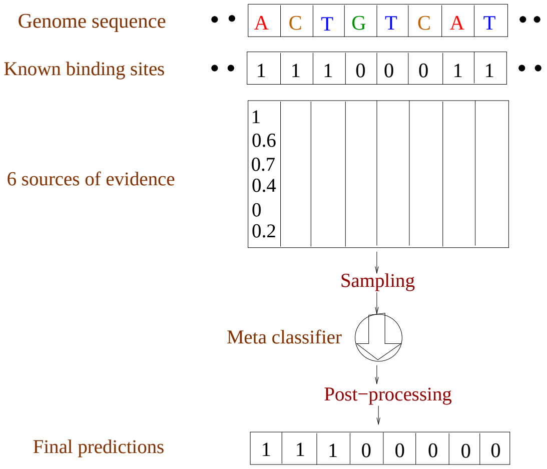

Fig. 1. The integration, sampling and classification of the data. The 6 algorithms give their own value for each sequence position and one such column is shown. The 6 results are combined into a 6-ary real valued vector. The data was under and over sampled, and then classified using a meta-classifier.

2 The description of the dataset

As mentioned in Section 1, the data consists of a merger of promoters annotated with transcription factor binding sites for mouse from the *ABS* and *ORegAnno* databases. This data is denoted as *ABS-And-OReg* data. The data also includes 250 upstream, non-coding sequences, denoted as *background* data.

– *ABS-And-OReg*
  There are 47 annotated promoter sequences in total. Sequences extracted from *ABS* are typically around 500 base pairs *(bp)* in length and those taken from *ORegAnno* are typically around 2000bp in length. Most of the promoters are upstream of their associated gene although a small number extend over the first exon and include intronic regions: where promoters were found to overlap they were merged. The total dataset is comprised of 60851 nucleotides, each of which may be part of a binding site.
– *Background*
  250 regions were randomly picked from across the mouse genome (forward strand genes only). The first 500bp from each sequence were selected i.e. the nucleotides that are 5000-4500 away from the gene with which they are associated. The idea is to extract non-coding sequences that are also probably non-regulatory.

> A check is also made that the selected region is indeed at least 4500 base pairs away from any neighbouring gene. It is common that a neighbouring gene can be close by and/or overlapping. The data is a sequence of 124467 nucleotides, and is believed to contain no TF binding sites.

For each nucleotide there is a real valued result from each of the six sources of evidence. Each nucleotide also has a label denoting whether it is part of a known binding site.

Six sources of evidence were generated from UCSC genome website, and were used as input in this study. Computational predictions of binding sites were generated using *MotifLocator* and *EvoSelex*. *MotifLocator* uses the PHYLOFACTS matrices from the JASPAR database[4] to scan for stringent matches in the sequences. *EvoSelex* uses motifs from [4] and the *Fuzznuc* algorithm to search for consensus sequences. A number of sources of genomic annotation evidence were extracted from the UCSC genome browser[5]: *Regulatory Potential* (RP) is used to compare frequencies of short alignment patterns between known regulatory elements and neutral DNA. The RP scores were calculated using alignments from the genomes of human, chimpanzee, macaque, rat, mouse, cow and dog. *PhastCons* is an algorithm that computes sequence conservation from multiple alignments using a phylo-HMM strategy. The algorithm was used with two levels of stringency. The *CpGIsland* algorithm finds ‘CG’ nucleotide sub-sequences in the regulatory region which are typically found near transcription start sites and are rare in vertebrate DNA.

3 Methods

3.1 Sampling

In our dataset, there are less than 2.93% binding positions amongst all the vectors, so this is an extremely *imbalanced* dataset [8]. Since the dataset is imbalanced, the supervised classification algorithms will be expected to over predict the majority class, namely the non-binding site category. There are various methods of dealing with imbalanced data [13]. In this work, we concentrate on the data-based method [3]: using under-sampling of the majority class (non-binding sites) and over-sampling of the minority class (binding site examples). We combine both over-sampling and under-sampling methods in our experiments.

For under-sampling, we randomly selected a subset of data points from the majority class. In [8], the author addresses an important issue that the class imbalance problem is only a problem when the minority class contains very small subclusters. This indicates that simply over sampling with replacements may not significantly improve minority class recognition. To overcome this problem, we apply a synthetic minority over-sampling technique (SMOTE) as proposed in [3]. For each member of the minority class its nearest neighbours in the same class are identified and new instances are created, placed randomly between the instance and its neighbours.

[4] http://jaspar.genereg.net/

[5] http://genome.ucsc.edu/

3.2 Biologically Constrained Post-Processing

We propose a two-step post-processing over the SVM predictions. First, since TF binding sites are almost never found within an *exon*, an exon prediction can be considered to be negative evidence for a TF binding site at a given position. Although exon predictions are still not perfect, they are much more robust than TF binding site predictions by several orders of magnitude. There is much less noise in the signals that delimit them in the sequence. Therefore, predicted components of a TF binding site will be removed if they are within a predicted exon position.

One important concern when applying classifier algorithms to the output of many binding site prediction algorithms is that the classifier decisions could result in biologically unfeasible results. The original algorithms only predict reasonable, contiguous sets of base pairs as constituting complete binding sites. However when combined in our meta-classifier each base pair is predicted independently of the neighbouring base pairs, and it is therefore possible to get lots of short predicted binding sites of length one or two base pairs. In this and a previous study, it was observed that many of the predictions made by the classifiers were highly fragmented and too small to correspond to biological binding sites. It was not clear whether these fragmented predictions were merely artifacts or whether they were accurate but overly conservative.

Since the limits of biologically observed binding site lengths are typically in the range 5-30 bp, we simply remove any predicted TF binding site with a length smaller than 5bp. It was found that removal of the fragmented predictions had a considerable positive effect on the performance measures, most notably for *Precision*.

3.3 Classifier Performance

In cases such as the imbalanced data simple error rates are inappropriate - an error rate of 2.93% can be obtained by simply predicting the majority class. Therefore it is necessary to use other metrics. Several common performance metrics, such as Recall (also known as Sensitivity), Precision, False Positive rate (FP-Rate) and F-Score, can be defined using the confusion matrix (see Table 1) computed from the test results:

Table 1. A confusion matrix

	Predicted Negatives	**Predicted Positives**
Actual Negatives	True Negatives (TN)	False Positives (FP)
Actual Positives	False Negatives (FN)	True Positives (TP)

$$\text{Recall} = \frac{\text{TP}}{(\text{TP + FN})}, \quad \text{Precision} = \frac{\text{TP}}{(\text{TP + FP})},$$

$$\text{F-Score} = \frac{2\cdot\text{Recall}\cdot\text{Precision}}{\text{Recall+Precision}}, \quad \text{FP-Rate} = \frac{\text{FP}}{\text{FP+TN}}.$$

Furthermore the Correlation Coefficient (CC) [12], is given below:

$$\text{CC} = \frac{\text{TP}\cdot\text{TN} - \text{FN}\cdot\text{FP}}{\sqrt{(\text{TP +FN})(\text{TN+FP})(\text{TP+FP})(\text{TN+FN})}},$$

Note that for all the measures except FP-Rate a high value is desirable. Precision is the proportion of the positively categorised samples that are actually part of a binding site. Increasing the Precision of the prediction is one of the main goals of our meta-classifier. However increasing Precision is normally accompanied by a decrease in the Recall, so the F-Score, which takes into account both Recall and Precision, is a useful measure of overall performance. The Correlation Coefficient (at nucleotide level) measures the correlation of the prediction with the target. The FP-Rate is the proportion of all the negative samples that are incorrectly predicted. The original algorithms generally have a high FP-Rate and reducing this is another major goal of our classifier.

4 Experiments: binding sites prediction

4.1 Simulation setup

First the *ABS-And-OReg* data was divided into a training set that consisted of $2/3$ of the data, the remaining $1/3$ including 20 promoter sequences was used as the test set. We consider the following cases: 1) all non-binding site examples are selected from the *ABS-And-OReg* data; 2) all non-binding site examples are selected from the *background* data; 3) we repeat case 2) using only 4 features, that is without the two prediction algorithms *MotifLocator* and *EvoSelex* as inputs. In the last two cases, the training sets are actually a combination of *ABS-And-OReg* and *background* data, since all training examples of components of TF binding sites are from *ABS-And-OReg* and all non-binding site examples are from *background*.

Amongst the data, there are repeated vectors, some with the same label (repeated items), and some with contradictory labels (inconsistent items). These items are unhelpful in the training set and were therefore removed. The training datasets are then consistent. However, in the case of the test set, the full set of data is considered.

In the *ABS-And-OReg* data, there are fewer than 2.93% binding positions amongst all the vectors, so this is imbalanced data. To cope with this problem we used sampling. For under sampling, a subset of data points from the majority class is randomly selected. In this work, we apply SMOTE for over sampling, where we take 9 nearest neighbours, and increase the number of items in the minority class by a factor of 7. The final ratio of majority to minority class is set to 1 in all the following experiments. Note that we normalise the consistent training set before sampling so that each feature has zero mean and unit standard deviation.

After sampling, there are 3 different training sets based on each case mentioned above.

Case 1: original data from *ABS-And-OReg* denoted *orig*.

Case 2: postive examples from *ABS-And-OReg* and negative examples from *background* denoted *orig+bg*.

Case 3: As case 2 but using only four features, denoted by *orig+bg_4f*.

Table 2 gives the size of these datasets.

In the following experiments, we apply an SVM for classification. The SVM software is publicly available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm/. The *radial basis kernel* is employed. Therefore the SVM has two free parameters: the *cost* C and γ related to the radial basis kernel function. The range for C is set to $[20,\ 250,\ 500,\ 1000,$

$2000,\ 5000]$ and for γ is $[0.001,\ 0.01,\ 0.1,\ 1,\ 10]$. In all the following experiments, the best values of C and γ are $C = 5000$ and $\gamma = 10$, selected by standard 5-fold cross validation.

Table 2. Description of datasets used in this work (bp denotes base pair).

Type	**Dataset**	**Negative** (bp)	**Positive** (bp)	**Size** (bp)
Original	*ABS-And-OReg*	59070	1782	60851
Original	*Background*	124467	0	124467
Training	*orig*	5446	5446	10892
	orig+bg	5757	5757	11514
	orig+bg_4f	5264	5264	10528
Test	test (from *ABS-And-OReg*)	18124	784	18908

4.2 Experimental results

Before presenting the main results we should point out that predicting binding sites accurately is extremely difficult. The best individual original algorithm (*EvoSelex*) produces over 11 times as many false positives as true positives on the test set. This makes the predictions almost useless to a biologist as most of the suggested binding sites will need expensive experimental validation and most will not be useful. Therefore the key aim of our combined classifier is to reduce the number of false positives while increasing the number of true positives given by the original algorithms.

Table 3 shows experimental results without post-processing. For comparison, we also give results of the two original prediction algorithms: *MotifLocator* and *EvoSelex*, over the test set.

Table 3. Classification results without post-processing (in percentage %)

	Recall	*Precision*	*F-Score*	*FP-Rate*	*CC*
MotifLocator	42.5	7.1	12.1	24.2	8.4
EvoSelex	34.8	8.0	13.0	17.2	9.1
orig	43.1	12.5	19.4	13.0	17.2
orig+bg	66.1	13.3	22.1	18.7	23.4
orig+bg_4f	60.5	16.3	25.7	13.4	26.0

The first notable feature of these results is that the meta classifiers have produced stronger Recalls and Precisions than those of the two original algorithms. Therefore, the F-Score, which can be viewed as an average of the Recall and Precision, has also been increased. The nucleotide level correlation coefficient has been significantly improved. As for the FP-Rate, the meta classifiers trained on *orig+bg_4f* and *orig*, have reduced the FP-Rate by 22.1% and 24.4%, respectively, compared with *EvoSelex*, while the meta classifier trained on *orig+bg* has increased the FP-Rate by 8.7%.

The second notable feature of these results is that the meta classifier trained on *orig+bg_4f*, which used only 4 features, produced a better performance than the one that used all 6 features when looking at the F-Score and CC values, which assess the overall performance of a classifier.

One more notable feature of these results is that one can obtain a better overall performance when using non-binding site examples from the *background* set rather than from the *ABS-and-OReg* dataset.

Finally we investigate how the results can be further improved by removing those predictions of base-pairs being part of a binding site that are not biologically plausible. As described earlier we find that removing predictions that are either within exons or not part of a contiguous predicted binding site of at least five nucleotides gives a better result. So here we take the predictions of our experimental results and remove all those that do not meet the criteria. The results can be seen in Table 4.

Table 4. Classification results with and without post-processing (in percentage %)

	Recall	*Precision*	*F-Score*	*FP-Rate*	*CC*
EvoSelex	34.8	8.0	13.0	17.2	9.1
orig+bg_4f	60.5	16.3	25.7	13.4	26.0
orig+post_processing	40.6	13.7	20.4	11.1	17.9
orig+bg+post_processing	61.0	14.8	23.8	15.2	24.2
orig+bg_4f+post_processing	58.0	17.5	26.9	11.8	26.8

It shows that all FP-Rates are reduced when compared with the best original algorithm *EvoSelex*. In addition, comparing *orig+bg_4f+post_processing* with *orig+bg_4f*, one can see that the FP-Rate has been further reduced to 11.8%. Looking at two overall performance values, F-Score and CC, it shows that the accuracy of predictions is further improved after post-processing. Interestingly, *orig+bg+post_processing* has a larger number of true positives (Recall) than *orig+bg_4f+post_processing*. However, *orig+bg_4f+post_processing* has a lower FP-Rate and better overall performance on F-Score and CC. Specifically, *orig+bg_4f+post_processing* has increased the Recall by 66.7%, the Precision by 118.8%, the F-Score by 106.9% and CC by 194.5%, while reduced the FP-Rate by 31.4% when compared with the original prediction algorithm *EvoSelex*.

To further analyse our method, we investigate in more detail the predictions on each test promoter. Figure 2 shows the nucleotide level correlation coefficient within each promoter between the known nucleotide positions and the predicted nucleotide positions for each prediction algorithm.

It can be seen that there are more higher correlation (bright patterns) between the known nucleotide positions and the predicted nucleotide positions based on each promoter given by the 3 meta classifiers. It indicates that the two original prediction algorithms can only successfully find few parts of binding sites, while the meta classifiers can detect more parts of binding sites by integrating several diverse sources. In addition, although both meta classifiers *orig* and *orig+bg* include *MotifLocator* and *EvoSelex* as part of the input, these two prediction algorithms do not contribute significantly in the final decision to the meta classifiers. For example, there is a relatively high CC value in both *MotifLocator* and *EvoSelex* predictions within test promoter 6, but all 3 SVM meta classifiers produce a lower CC value. One more example is test promoter 5. Both *MotifLocator* and *EvoSelex* predictions have low correlation with the known nucleotide positions, but the 3 meta classifiers give a very high CC value. It suggests that those 4 suggestive evidences rather than the two original prediction algorithms are much more important for the classification.

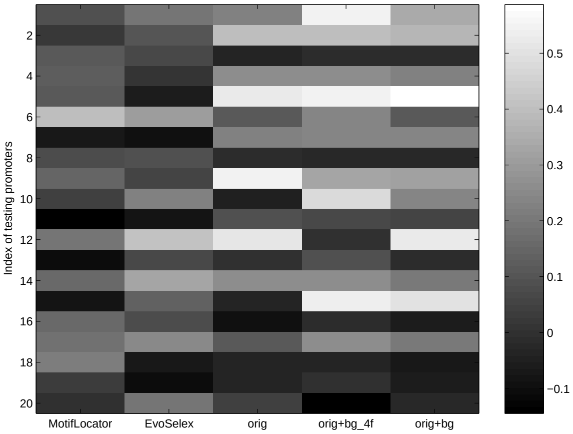

Fig. 2. Correlation coefficients between predicted positions and the known positions in each test promoter. Note that predictions from 3 meta classifiers are post-processed. High correlations are associated with brighter cells. Poor correlations are associated with darker cells.

5 Discussion

The identification of regions in a sequence of DNA that are regulatory binding sites is a very difficult problem. Here we have confirmed our earlier results showing that a meta classifier using multiple sources of evidence can do better than any of the original algorithms individually. In particular it was possible to reduce the number of false positive predictions.

Importantly we have also shown that using negative data that is very probably correctly labelled leads to a better prediction results. This is perhaps unsurprising, but it does suggest that some of the original data in the promoter sequences may be incorrectly labelled. This suggests that more binding sites exist on the promoter sequences than have been found by the expensive experimental techniques currently needed to produce such predictions.

Finally results that the meta classifier trained on only 4 features can produce a better performance than the one used all 6 features demonstrate the importance of feature selection. One needs to choose sources of complementary evidences which are in fact the most useful to consider. In the future, we intend to cope with this by applying the SVM classification based on *Recursive Feature Elimination* [5].

Much further work is needed to extend our current methods. The technique needs to be evaluated on other species and the biological significance of the predictions needs close examination. However it seems likely that the use of background data, as demonstrated here, will facilitate generally improved predictions.

References

1. Bailey, T.L., Elkan, C.: Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, 28-36, AAAI Press. (1994)
2. Blanchette, M., Tompa, M.: FootPrinter: A Program Designed for Phylogenetic Footprinting. Nucleic Acids Research. Vol. 31, No. 13, 3840-3842 (2003)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer,W.P.: Smote: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research. 16, 321–357 (2002)
4. Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J., Birney, E.: The Discovery, Positioning and Verification of a Set of Transcription-associated Motifs in Vertebrate. Genome Biol. 6(12), (2005)
5. Guyon, I., Weston, J., Barnhill, S., Vapnik,V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 46, 389-422 (2002)
6. Hu, J.J., Yang, Y.F.D., Kihara, D.: EMD: an Ensemble Algorithm for Discovering Regulatory Motifs in DNA Sequsences. BMC Bioinformatics. (2006)
7. Huber, B.R., Bulyk, M.L.: Meta-analysis Discovery of Tissue-specific DNA Sequence Motifs from Mammalian Gene Expressin Data. BMC Bioinformatics. (2006)
8. Japkowicz, N.: Class Imbalances: Are We Focusing on the Right Issure? In: Workshop on learning from imbalanced datasets, II, ICML. (2003)
9. Scholköpf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, (2002)
10. Sun, Y., Robinson, M., Adams, R., Kaye, P., Rust, A.G., Davey, N.: Using Real-valued Meta Classifiers to Integrate Binding Site Predictions. In: Proceedings of International Joint Conference on Neural Network. (2005)
11. Sun, Y., Robinson, M., Adams, R., Davey, N., Rust, A.: Predicting Binding Sites in the Mouse Genome. In: Proceedings The Sixth International Conference on Machine Learning and Applications (ICMLA'07). (2007)
12. Tompa, M., *et al.*: Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites. Nature Biotechnology. 23(1), (2005)
13. Wu, G., Chang, E.: Class-boundary Alignment for Imbalanced Dataset Learning. In: Workshop on learning from imbalanced datasets, II, ICML. (2003)