Predicting Binding Sites in the Mouse Genome

Yi Sun, Mark Robinson, Rod Adams, Neil Davey
Science and Technology Research Institute, University of Hertfordshire,
United Kingdom
y.2.sun, m.robinson, r.g.adams, n.davey@herts.ac.uk

Alistair Rust
Institute for Systems Biology,
1441 North 34th Street,
Seattle, WA 98103, USA
arust@systemsbiology.org

Abstract

The identification of cis-regulatory binding sites in DNA in multicellular eukaryotes is a particularly difficult problem in computational biology. To obtain a full understanding of the complex machinery embodied in genetic regulatory networks it is necessary to know both the identity of the regulatory transcription factors together with the location of their binding sites in the genome. We show that using an SVM together with data sampling, to integrate the results of individual algorithms specialised for the prediction of binding site locations, can produce significant improvements upon the original algorithms applied to the mouse genome. These results make more tractable the expensive experimental procedure of actually verifying the predictions.

1. Introduction

In this paper, we present research demonstrating the utility of integrating multiple sources of binding site predictions and genomic annotation evidence using classification techniques employed in the machine learning field, to better identify transcription factor binding sites in regulatory regions from the mouse genome (*M.musculus*). There is a vast multitude of algorithms to search for binding sites in current use. However, most of them are severely limited in their accuracy and yield many false positive results. That imposes a serious problem for practicing biologists, as experimentally validating a prediction is costly.

In [7] we attempted to reduce these false positive predictions using classifications techniques employed in the field of machine learning on yeast (*S.cerevisiae*) regulatory regions.

In this paper we show how algorithmic predictions and genomic annotation evidence can be combined so that a Support Vector Machine (SVM) can perform a new prediction that significantly improves on the performance of any one of the individual algorithms. Moreover we show how the number of false positive predictions can be reduced by approximately 50.0%.

2. Problem domain

Gene regulatory networks (GRNs) encode developmental programs and underlie many important biological systems. They are composed of functional units, genes and their associated regulatory regions, along with any regulatory interactions arising between these functional units. The connectivity of gene regulatory networks is determined by the location and identity of *cis-regulatory* binding sites in gene regulatory sequences. These binding sites locate transcription factors to regions of the genome where they can exert a regulatory influence on the expression of a specific gene, or set of genes. Discovery and characterisation of cis-regulatory binding sites is currently a critical bottleneck in the analysis of GRNs. Computational binding site predictions offer the possibility of high-throughput genome wide analyses and as such are the focus of considerable research.

A wide range of algorithmic strategies have been developed to tackle the problem of computational binding site prediction. These strategies include scanning sequences for matches to known binding sites, statistical analysis of sequence features, looking for over-represented patterns in clusters of co-expressed gene promoters and phylogenetic analysis. Each of these approaches have different dependencies on data availability and as a result different applicability depending on the type of data available. A ma-

0-7695-3069-9/07 $25.00 © 2007 IEEE
DOI 10.1109/ICMLA.2007.28

jor problem with all binding site prediction algorithms is that they are prone to particularly high rates of false positive. One approach to reducing false positive predictions is to integrate predictions from multiple sources. Previous work has explored the use of a range of classification algorithms from the machine learning field for this task using yeast (*S.cerevisiae*) promoter sequences [7], establishing that this is a viable approach for efficiently reducing the rate of false positive predictions. Gene regulation in *S.cerevisiae* is however much simpler than that typically found in multicellular eukaryotes. Here we present research demonstrating the utility of integrating multiples sources of binding site predictions and genomic annotation evidence with machine learning algorithms in the mouse, *M.musculus*, promoter sequences.

3. Data and algorithm description

The dataset consists of a merger of annotated transcription factor sites in a set of promoters in the mouse genome (*M.musculus*), curated from the *ABS*[1] and *ORegAnno*[2] databases. 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.

Seven sources of evidence 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[3] 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[4]: *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.

The data is a sequence of 60851 nucleotides, each of which may be part of a binding site. For each nucleotide there is a prediction result from each of the seven sources of evidence, which may be either real valued or binary. Each nucleotide also has a label denoting whether it is part of a known binding site. The data therefore consists of 60851 7-ary real vectors, each with an associated label as shown in Figure 1.

The data set was divided into a training set that consisted of $2/3$ of the data, the remaining $1/3$ was used as the test set. 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. However, in the case of the test set, the full set of data is considered.

In the dataset, there are fewer than 2.93% binding positions amongst all the vectors, so this is an imbalanced dataset [4]. An imbalanced dataset imposes a problem for supervised classification algorithms, as they are expected to over-predict the majority class, namely the non binding site category. One of the techniques to overcome this problem is to apply the data based method: under-sampling of the majority class and over sampling of the minority class. For under sampling, a subset of data points from the majority class is randomly selected. For over sampling, both the SMOTE algorithm [2] and Gaussian mixture models are used. The process of integrating, sampling and classifying the data, is illustrated in Figure 1.

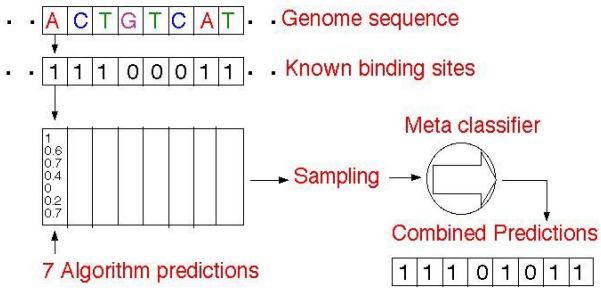

Figure 1. The integration, sampling and classification of the data. For each location in the sequence, the prediction results of the seven algorithms were integrated into one single vector. The data was under and over sampled, and then classified using a meta-classifier.

4. Data visualisation

Computational prediction of cis-regulatory binding sites is widely acknowledged as a difficult task [8]. Binding sites are notoriously variable from instance to instance and they can be located considerable distances from the gene being regulated in higher eukaryotes.

[1] http://genome.imim.es/datasets/meta2005/index.html
[2] http://www.oreganno.org/oregano/Index.jsp
[3] http://jaspar.genereg.net/
[4] http://genome.ucsc.edu/

Before attempting classification, we first look at the underlying data distribution by means of classical principal component analysis PCA [1], which linearly projects data into a two-dimensional space, where it can be visualised.

We visualise the non binding sites using PCA, then project the transcription factors binding sites into the same PCA projection space. The result is shown in Figure 2. It can be seen that the 2 classes are extremely difficult to separate.

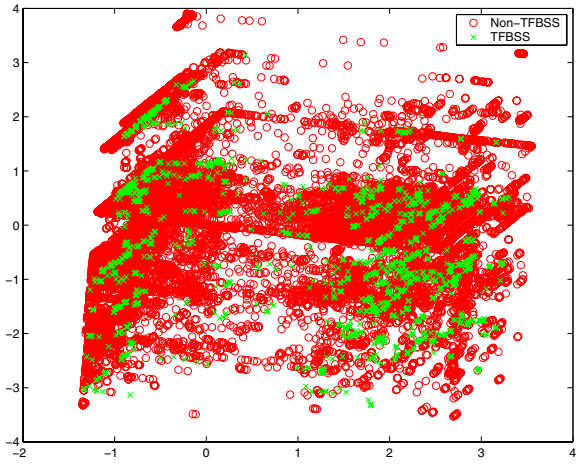

Figure 2. Projection of the non binding sites using PCA, where the binding sites are also projected into the non-binding sites' first two principal components space.

5. Sampling

In our dataset, there are less than 2.93% binding positions amongst all the vectors, so this is an extremely *imbalanced* dataset [5]. Since the dataset is imbalanced, the supervised classification algorithms will be expected to over predict the majority class, namely the non-binding site category. This is demonstrated in the results shown in Section 8.1 for the unsampled data. There are various methods of dealing with imbalanced data [9]. In this work, we concentrate on the data-based method [2]: 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. The actual ratio of minority to majority class is set to 1 in this work.

For under-sampling, we randomly selected a subset of data points from the majority class. The over-sampling case is more complex. In this work we apply two methods to tackle the problem.

The first is the synthetic minority over-sampling technique. In [5], 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 as proposed in [2]. 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. We take 9 nearest neighbours, and increase the number of items in the minority class by a factor of 7.

Another approach we applied in this work is to use a Gaussian mixture model [1].

We apply a Gaussian mixture model for the class-conditional probability density of binding sites. In a Gaussian mixture model, the probability density function of each class is independently modelled as a linear combination of Gaussian basis functions. The number of basis functions, their position and variance and their mixing coefficients are all parameters of the model. The *expectation-maximisation* (EM) algorithm [3] is used to estimate parameters of a mixture model for an optimal fit to the training data.

In our experiment, we first estimated parameters of the class-condition density from the minority class. Synthetic data from the GMM is generated in order to increase the number of data points in the minority class. A 3 source, spherical Gaussian mixture model, is used in this work.

6. Performance metrics

Since the dataset is imbalanced, simple error rates are inappropriate to evaluate the algorithms. 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 Positive (FP)
Actual Positives	False Negatives (FN)	True Positives (TP)

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}, \tag{2}$$

$$\text{F-Score} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall+Precision}}, \tag{3}$$

$$\text{FP-Rate} = \frac{\text{FP}}{\text{FP+TN}}. \quad (4)$$

Furthermore the Correlation Coefficient (*CC*), is given below:

$$CC = \frac{\text{TP} \cdot \text{TN} - \text{FN} \cdot \text{FP}}{\sqrt{(\text{TP} + \text{FN})(\text{TN}+\text{FP})(\text{TP}+\text{FP})(\text{TN}+\text{FN})}}, \quad (5)$$

Note that for all the measures except *FP-Rate* a high value is desirable. Most computational prediction algorithms have a high *Recall* by simply over predicting the binding site class (predicting every item to be positive gives a *Recall* of 1), and this is problematic. On the other hand *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 *FP-Rate* is the proportion of all the negative samples that are incorrectly predicted. The base algorithms generally have a high *FP-Rate* and reducing this is another major goal of our classifier.

7. Biologically Constrained Post-Processing

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. Therefore, predictions with a length smaller than a threshold value were removed and the effect on the performance measures observed. It was found that removal of the fragmented predictions had a considerable positive effect on the performance measures, most notably for *Precision* and that an optimal value for the threshold is 6 bp. Interestingly, this value corresponds roughly to the lower limit of biologically observed binding site lengths which are typically in the range 5-30 bp in length.

8. Results

We use 3 classifiers: Fisher's linear discrimination (FLD) [1], a single layer network (SLN) [1], and a support vector machine (SVM) using Gaussian kernel [6]. Optimal parameters for the SVM were found using 5-fold cross-validation.

8.1. SLN results for the original imbalanced data with no sampling

The confusion matrix of the SLN trained with imbalanced data is shown in Table 2 (the SVM has similar performance). The trained classifer simply predicted the majority class in all cases, proving that sampling is necessary in this work.

Table 2. The confusion matrix of the SLN trained with imbalanced data.

$TN = 18223$	$FP = 0$
$FN = 784$	$TP = 0$

8.2. Results using sampling

The performance of trained classifiers is shown in Table 3, together with two of the base algorithms.

Compared with the two base algorithms, all classifiers, except FLD decrease the *FP-Rate* and increase the *Precision* and the *CC* values. It can be seen that the SVM algorithm with SMOTE sampling gives the best *Precision* and *F-Score*. It improves the *Precision* by 90.0%, the *F-Score* by 71.5%, and decreases the *FP-Rate* by 41.9% when compared with the best base algorithm (that is *Evoselex*, which has the highest *F-Score*, *Precision* and lowest *FP-Rate* between the two base algorithms).

It also shows that the SVM algorithm with GMM sampling gives the lowest *FP-Rate*. However this is at a cost: in comparison to the best base algorithm the *Recall* has been decreased. The classifier has become more conservative, predicting binding sites less often but with greater accuracy.

The SLN algorithm works well on all common performance metrics when compared with the two base algorithms, and it gives the best *CC* improving the *CC* by 125.27% when compared with the *EvoSelex* algorithm. However, it has higher *FP-Rate* when compared with SVM.

Interestingly, one can see that FLD without sampling gives better values on *Recall*, *Precision*, and *CC* when compared with the two base algorithms, but not on *FP-Rate*, where it has a bigger value than *EvoSelex*. FLD with SMOTE sampling performs worse than the two base algorithms on *FP-Rate*. It suggests that FLD works better without sampling on this problem.

In summary, the SVM algorithm with the SMOTE sampling performs well and outperforms all the other classifiers

Table 3. Classification results.

Samplings	Classifier	*Recall*	*Precision*	*F-Score*	*FP-Rate*	*CC*
No sampling	**MotifLocator**	0.425	0.071	0.121	0.241	0.085
	EvoSelex	0.348	0.080	0.130	0.172	0.091
	FLD	0.647	0.112	0.190	0.221	0.198
SMOTE	**FLD**	0.749	0.075	0.136	0.398	0.142
	SLN	0.531	0.133	0.213	0.148	0.205
	SVM	0.417	**0.152**	**0.223**	0.100	0.199
GMM	**SVM**	0.226	0.139	0.172	**0.060**	0.132

on the biologically important values of *Precision*, *FP-Rate* and *F-Score*.

Figure 3 shows a fragment of the genome with the two original algorithmic predictions, the SVM predictions with both SMOTE and GMM samplings, and the actual annotation. This figure should not be considered representative for the performance on all genes, which is typically highly variable from one promoter sequence to another, however it illustrates a simple example. The "known" binding sites are shown in black and represent the best available information for the location of *in vivo* functional cis-regulatory elements.

It can be seen from the visualisations that in all examples the SVM is the most conservative predictor when compared with the two base algorithms. Both the trained classifiers are far more conservative than the original algorithms. The SVM with GMM algorithm is the best performing algorithm for *Vim-merged*, which has two annotated binding sites.

Note that it is dangerous to rely too heavily on the completemess of the annotated data, there being no way, currently, to assure that this is the case. These kinds of questions can only be answered conclusively by experimental validation of algorithm predictions.

8.3. Results after post-processing

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 not part of a contiguous predicted binding site of at least six nucleotides gives an optimal result. So here we take the predictions of the SMOTE+SVM and GMM+SVM and remove all those that do not meet this criterion. Table 4 shows results.

The SVM with SMOTE sampling algorithm produces our best result by some way. The *Precision* of the prediction has been increased to 0.166 and the *FP-Rate* is now down to just 0.088%. However, the SVM with GMM sampling algorithm has become more conservative, where the *Recall* has been further decreased when compared with the one without the post-processing. It gives the lowest *FP-Rate* over all our experiments, and the *Precision* has improved a little when compared with the one without the post-processing.

9. Conclusions

The identification of regions in a sequence of DNA that are regulatory binding sites is a very difficult problem. Individually the original prediction algorithms are inaccurate and consequently produce many false positive predictions. Our results show that by combining the predictions of the original algorithms and other sources of evidence we can make a significant improvement from the individual results. This suggests that the predictions that they produce are complementary, perhaps giving information about different parts of the genome. The only problem of our approach is that the combined predictor can indicate implausibly short binding sites. However we have shown that by simply rejecting these binding sites, using a length threshold, gives a very low rate of false positive predictions. This is exactly the result that we wanted: false positives are very undesirable in this particular domain. On the technical issue of dealing with the highly imbalanced data we found that the SMOTE sampling outperforms the GMM sampling on all performance metrics but the *CC*. However, more experiments with different structures of GMM for sampling need to be done.

References

[1] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc*, 39:1–38, 1977.

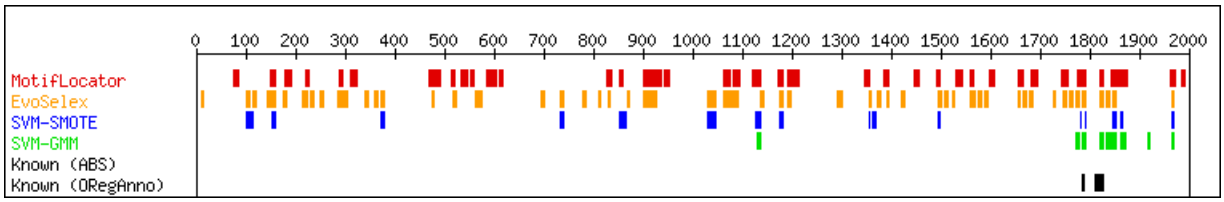

Figure 3. A fragment of the genome *Vim-merged* with the 2 original predictions, the actual annotations in black.

Table 4. Classification results with and without post-processing

	Recall	*Precision*	*F-Score*	*FP-Rate*	*CC*
EvoSelex	0.348	0.080	0.130	0.172	0.091
SMOTE+SVM	0.417	0.152	0.223	0.100	0.199
SMOTE+SVM+post-processing	0.404	**0.166**	**0.235**	0.088	0.210
GMM+SVM+post-processing	0.203	0.142	0.167	**0.053**	0.127

[4] L. Ettwiller, B. Paten, M. Souren, F. Loosli, J. Wittbrodt, and E. Birney. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrate. *Genome Biol.*, 6(12), 2005.

[5] N. Japkowicz. Class imbalances: Are we focusing on the right issure? *Workshop on learning from imbalanced datasets, II, ICML*, 2003.

[6] B. Scholkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2002.

[7] Y. Sun, M. Robinson, R. Adams, P. Kaye, A. Rust, and N. Davey. Using real-valued meta classifiers to integrate binding site predictions. In *Proceedings of International Joint Conference on Neural Network*, 2005.

[8] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, W. J. K. Y. Fu, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van. Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23:137–44, January 2005.

[9] G. Wu and E. Chang. Class-boundary alignment for imbalanced dataset learning. *Workshop on learning from imbalanced datasets, II, ICML*, 2003.