Combining Experts In Order to Identify Binding Sites in Genomic Data

Faisal Rezwan, Yi Sun
Rod Adams, Neil Davey
Science and Technology
Research Institute
Univesrity of Hertfordshire
F.Rezwan @herts.ac.uk

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

Mark Robinson
Department of Biochemistry
and Molecular Biology,
Michigan State University,
East Lansing MI 48824, USA
blobby@msu.edu

Abstract

The identification of cis-regulatory binding sites in DNA is a 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 classify the combination of the results of individual algorithms specialised for the prediction of binding site locations, can produce significant improvements upon the original algorithms. The resulting classifier produces fewer false positive predictions and so reduces the expensive experimental procedure of verifying the predictions.

1 Introduction

Binding site prediction is both biologically important and computationally interesting. One aspect that is challenging is the imbalanced nature of the data and that has allowed us to explore some powerful techniques to address this issue. In addition the nature of the problem allows biological heuristics to be applied to the classification problem. Specifically we can remove some of the final predicted binding sites as not being biologically plausible.

Computational predictions are invaluable for deciphering the regulatory control of individual genes and by extension aiding in the automated construction of the genetic regulatory networks to which these genes contribute. Improving the quality of computational methods for predicting the location of transcription factor binding sites (TFBS) is therefore an important research goal. 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. Larger scale studies, reconstructing the regulatory networks for entire systems or genomes, are therefore particularly reliant on computational predictions, there being few alternatives available.

DNA molecules are composed of a long chain of linked monomers, known as nucleotide bases, which come in four different types. The sequence of bases in a DNA sequence can be used to encode information necessary for the proper function of many biological systems. Two important examples include the gene sequences which encode an organism's complement of proteins and the regulatory sequences which by binding transcription factors help determine the coordinated expression of the proteins in space and time. Functional annotation of DNA sequences has taken an increasingly important role in the post-genomic era. Many regions of considerable functional importance, such as binding sites for transcription factors, consist of subtle signals encoded in the DNA sequence. Detection of these regions in genomic sequences is a critical step in our evolving understanding of gene regulation and gene regulatory networks. Transcription factor binding sites are notoriously variable from instance to instance and they can be located considerable distances from the gene being regulated in higher eukaryotes. Computational prediction of cis-regulatory binding sites is widely acknowledged as a difficult task (Tompa, Li et al. 2005).

In this paper we show how algorithmic predictions can be combined so that a Support Vector Machine (SVM) can subsequently 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 around 80%. We use two different data sets: for our major study we use a set of annotated yeast promoters take from the SCPD (Zhu and Zhang 1999), and then in order to validate the method with a complex multi-cellular species, the mouse, we used a set of 47 experimentally annotated

promoters extracted from the ABS (Blanco, Farre et al. 2006) and ORegAnno databases (Montgomery, Griffith et al. 2006) .

2 Background

The use of a non-linear classification algorithm for the purposes of integrating difference sources of evidence relating to cis-regulatory binding site locations, such as the predictions generated from a set of cis-regulatory binding site prediction algorithms, is explored in this paper. This is achieved by first generating a number of algorithmic predictions (a real number between 0 and 1 representing the probability that a nucleotide is part of a binding site, see Section 3) for a set of annotated (labelled) promoter sequences. These predictions are concatenated into vectors and an SVM is trained to classify them as either being part of a binding site or part of the background sequence.

A wide range of binding site prediction algorithms were used in this study. Those used for the analysis on yeast were selected to represent the full range of computational approaches to the binding site prediction problem. The algorithms chosen were typically taken from literature although some were developed in-house or by our collaborators in the case of PARS, Dream and Sampler. Table 1 lists the algorithms used with the yeast dataset, details can be found in (Robinson, Sun et al. 2006). Where possible, parameter settings for the algorithms were taken from the literature, if not available, default settings were used. A different set of algorithms were used when dealing with the mouse dataset to take advantage of the tracks available from the UCSC genome browser (Karolchik, Baertsch et al. 2003); once again they represent a range of different algorithmic approaches along with some additional sources of relevant evidence.

Table 1. The 12 Prediction Algorithms used with the yeast dataset. Note Dream was run using two different modes of operation.

Strategy	Algorithm
Scanning algorithms	Fuzznuc MotifScanner Ahab
Statistical algorithms	PARS Dream (2 versions) Verbumculus
Co-regulatory algorithms	MEME AlignACE Sampler
Evolutionary algorithms	SeqComp Footprinter

Table 2 lists the sources of evidence used with the mouse dataset. Each of these sources was extracted from the UCSC genome browser (Karolchik, Baertsch et al. 2003) for the promoter regions of interest.

Table 2. The 7 Prediction Algorithms used with the mouse dataset.

Strategy	Algorithm
Scanning algorithms	MotifLocator
	EvoSelex
Evolutionary algorithms	Regulatory Potential
	PhastCons (Conserved)
	PhastCons (Most conserved)
Indirect evidence	CpGIsland
Negative evidence	Exon

3 Description of the Data

High quality experimentally annotated datasets were used in this study. In all cases it is important to be aware that such annotations are limited to positive observations and as such cannot guarantee completeness. It is possible that additional binding sites exist in the sequences used and will here be classified as background. Any additional binding sites which are present but which are not included in the annotations will necessarily affect our evaluation of prediction accuracy in this study.

The yeast, S.cerevisiae was selected for the model organism for the first experiment; the use of this particularly well studied model organism ensures that the annotations available are among the most complete. 112 annotated promoter sequences were extracted from the S.cerevisiae promoter database (SCPD) (Zhu and Zhang 1999) for training and testing the algorithms. For each promoter, 500 base-pairs (bp) of sequence taken immediately upstream from the transcriptional start site was considered sufficient to typically allow full regulatory characterisation in yeast. In cases where annotated binding sites lay outside of this range, then the range was expanded accordingly. Likewise, where a 500 bp upstream region would overlap a coding region then it was truncated accordingly. Further details about how the data was obtained can be found in (Robinson, Sun et al. 2006).

The dataset for the second experiment uses annotated transcription factor sites for the mouse, M.musculus, taken from the ABS and ORegAnno databases. There are 47 annotated promoter sequences in total. Sequences extracted from ABS are typically around 500 base pairs in length and those taken from ORegAnno are typically around 2000 bp 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. MotifLocator uses the PHYLOFACTS matrices from the JASPAR database (Wasserman and Sandelin 2004) to scan for good matches in the sequences. EvoSelex uses motifs from (Rajewsky, Vergassola et al. 2002) and the Fuzznuc algorithm to search for consensus sequences. A number of sources of evidence were extracted from the UCSC genome browser: 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 mouse, rat, human, chimpanzee, macaque, dog, and cow. 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, conserved and most conserved, which are included as separate sources of evidence. The CpGIsland algorithm finds CG sequences in the regulatory region which are typically found near transcription start sites and are rare in vertebrate DNA. Finally, Exon predictions are included for those sequences where the sequence extends over the first exon and into the next intronic region and should be considered a type of negative evidence.

For both experiments, each source of evidence is placed into a matrix consisting of a vector of inputs for each sequence position, each associated with a binary label indicating the presence or absence of an experimental annotation at that position, see Figure 1, which illustrates the input vectors for the yeast dataset.

The algorithms either output a binary value designating the prediction of being within a binding site or not, or a probability. All predictions in the matrix were then normalised as real values in the range [-1,1] with the value of 0 allocated to sequence positions where an input source was unavailable. In other words each feature is scaled individually as a number between -1 and +1. Additionally, we contextualize the training and test datasets to ensure that the classification algorithms have data on contiguous binding site predictions. This is achieved by windowing the vectors within each of the annotated promoter sequences. We use a window size of 7 (found after testing of various window sizes, see (Sun, Robinson et al. 2005)) providing contextual information for 3 bp either side of the position of interest.

Additionally this procedure carries the considerable benefit of eliminating a large number of repeated or inconsistent vectors which are found to be present in the data. (Sun, Robinson et al. 2005) These arise when for instance the 12 yeast algorithms produce the same set of predictions for different nucleotides; if the annotations are all the same then these are repeats and if different these are inconsistent data items. The inclusion of such items could otherwise pose a significant obstacle to the training of the classifiers. With windowing the input vector is increased from 12 to 84 and the chance of all 84 values repeating is therefore much reduced. A similar process was undertaken for the mouse dataset.

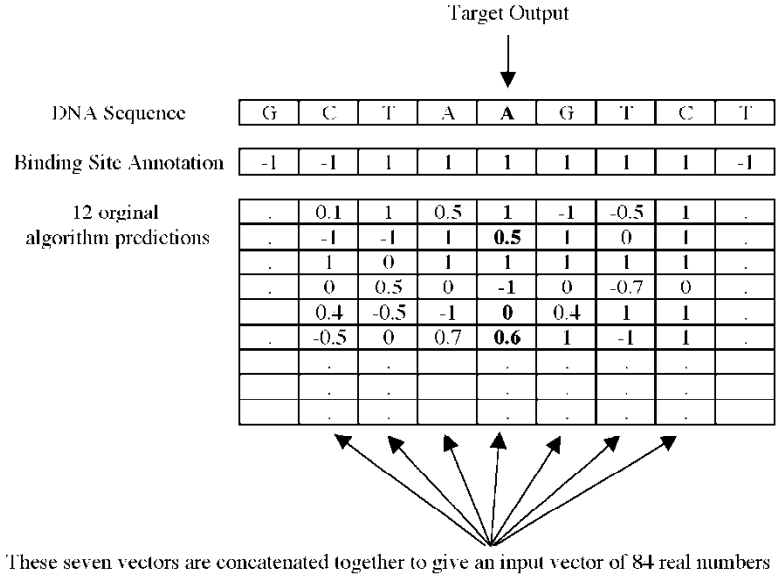

Fig. 1. The formation of the windowed data for the yeast dataset. The 12 predictions from the original algorithms for the target site are concatenated with the predictions from the 3 sites on either side. This gives an input vector of 12 by 7 real numbers. The corresponding label of this vector is the annotation of the central nucleotide.

4 Performance Metrics

As only approximately 8% of the yeast dataset is annotated as being a part of a binding site, this dataset is imbalanced (as is the mouse dataset). If the algorithms are to be evaluated in a useful manner simple error rates are inappropriate, it is therefore necessary to use other metricsSeveral common performance metrics, such as *Recall* (also known as *Sensitivity*), *Precision (also known as Specificity)*, *False Positive rate (FP-Rate)* and *F-Score*, can be defined using a confusion matrix (see Table 3) of the classification results. *Precision* describes the proportion of predictions that are accurate; *Recall* describes the proportion of binding site positions that are accurately predicted; *FP-Rate* describes the proportion of the actual negatives that are falsely predicted as positive; and the *F-Score* is the weighted harmonic mean of *Precision* and *Recall*. There is typically a trade off between *Precision* and *Recall*, making the *F-Score* particularly useful as it incorporates both measures. In this study, the weighting factor, β, was set to 1 giving equal weighting to both *Precision* and *Recall*. It is worth noting that for all these metrics a higher value represents improved performance with the solitary exception of *FP-rate* for which a lower value is preferable.

Table 3. The definition of performance measures

	Predicted Negatives	**Predicted Positives**
Actual Negatives	True Negatives - *TN*	False Positives - *FP*
Actual Positives	False Negatives - *FN*	True Positives - *TP*

$$Recall = \frac{TP}{TP + FN} \quad Precision = \frac{TP}{TP + FP}$$

$$FP\_Rate = \frac{FP}{FP + TN} \quad F\_Score = \frac{\left(1 + \beta^2\right) Recall \times Precision}{\beta^2 Recall + Precision}$$

5 Results

5.1 Results for the yeast genome

Before presenting the main results we should point out that predicting binding sites accurately is extremely difficult. For the yeast dataset the performance of the best individual original algorithm (Fuzznuc) is shown in Table 4.

Table 4: Confusion Matrix for the yeast data for the best original algorithm, Fuzznuc

	Predicted Negatives	**Predicted Positives**
Actual Negatives	*TN= 83%*	*FP = 10%*
Actual Positives	*FN = 4%*	*TP = 3%*

Here we can see over three times as many false positives as true positives. 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 a key aim of our combined classifier is to significantly reduce the number of false positives given by the original algorithms.

As described above the imbalanced nature of the data must be addressed. First the data is divided into a training set and test set, in the ratio 2 to 1. For the yeast dataset this gives a training set of 32,615 84-ary vectors and a test set of 16,739 vectors. It is not necessary to use cross validation of the training / test set division as the test set is so large (Henery 1994).
In the results here for the yeast dataset the majority class in the training set is reduced, by random sampling, from 30,038 vectors to 9,222 and the minority class was increased from 2,577 vectors to 4,611 vectors using the SMOTE algorithm. Therefore the ratio of the majority class to the minority class is reduced from approximately 12 : 1 to 2 : 1. Other ratios were tried but this appears to give good results (Sun, Robinson et al. 2005). The test set was not altered at all.

As described earlier an SVM with Gaussian kernel was used as the trainable classifier, and to find good settings for the two free parameters of the model, C and γ standard 5-fold cross validation was used. After good values for the parameters were found (C = 1000, γ = 0.001), the test set was presented and the results are given in Table 5:

Table 5: Results for the yeast dataset

	Recall	**Precision**	**F-Score**	**FP-Rate**
Best Original Algorithm	0.400	0.222	0.285	0.106
Meta Classifier -	0.305	0.371	0.334	0.044

The first notable feature of this result is that the combined classifier has produced a weaker Recall than the best original algorithm. This is because it is giving fewer positive predictions, but it has a much

higher precision. Of particular significance is that the FP-Rate is relatively low at 0.04, so that only 4% of the actual non-binding sites are predicted incorrectly.

5.2 Results for the mouse genome

In order to examine if our approach is also applicable to the much more complicated case of multi-cellular eukaryotes, we now give results for the mouse, *M.musculus* genome. Prediction in this case is significantly more difficult. The mouse genome contains significantly more non-coding DNA sequence than the yeast genome, thereby increasing the search space. Furthermore, complex, multi-cellular organisms, such as the mouse, exhibit more complex organisation of the gene regulatory regions. Genes are often regulated by a number of spatially distinct regulatory modules, each containing a number of transcription factor binding sites. These modules can be located not just in the regions proximal to the promoter but also many thousands of base pairs away, both upstream and downstream as well as inside intronic regions. Furthermore, there are a number of other biological features found in non-coding sequence which are not necessarily related to transcription factor binding or gene regulation at all. All these factors tend to increase the difficulty of making accurate computational predictions of binding sites.

Firstly we give the confusion matrix for the best individual matrix (*MotifLocator*).

Table 6: Confusion matrix of the best individual algorithm for the mouse data

	Predicted Negatives	**Predicted Positives**
Actual Negatives	*TN= 75%*	*FP = 22%*
Actual Positives	*FN = 1.5%*	*TP = 1.5%*

It is clear from the low true positives and the high false positives that this problem is indeed harder than the equivalent problem in yeast, as would be expected.
The results are shown in the following table:

Table 7: Results for the mouse dataset

	Recall	**Precision**	**F-Score**	**FP-Rate**
Best Original Algorithm	0.495	0.063	0.111	0.224
Meta Classifier -	0.300	0.159	0.208	0.069

Once again the Precison has been improved, at some cost to the Recall. However, once more, the FP-Rate is greatly reduced.

6 Discussion

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 we can make a significant improvement from their 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 problem area (Yellaboina, Seshadri et al. 2004).

We have investigated the contribution that each algorithm makes to the final prediction (Sun, Robinson et al. 2005) and find that there is a wide difference between the various methods. More work needs to be done to unravel both the detailed nature of the predictions and the biological significance of the results.

When we tested the method on the much more difficult case of the mouse genome we also found that the number of false positive predictions could be significantly reduced. The reduction of false positives by a factor of 6 relative to the reduction of the true positives by a factor of 2 illustrates that the processes is preferentially filtering noise from the predictions. One limitation of these results is the large reduction in Recall. Further work will extend the range of sources used as evidence, it is hoped that by incorporating a larger pool of evidence that less genuine predictions will be missed. The approach will also be applied to other available organism datasets to test the generality of these results. One particular goal is to apply the approach to systems where experimental validation of the predictions can

be made, circumventing the uncertainty surrounding the completeness of the promoter annotations currently available.

References

Blanco, E., D. Farre, et al. (2006). "ABS: a database of Annotated regulatory Binding Sites from orthologous promoters." Nucl. Acids Res. **34**(suppl_1): D63-67.

Henery, R. J. (1994). Methods for comparison. Machine learning, neural and statistical classification, Ellis Horwood**:** 107-124.

Karolchik, D., R. Baertsch, et al. (2003). "The UCSC Genome Browser Database." Nucl. Acids Res. **31**(1): 51-54.

Montgomery, S. B., O. L. Griffith, et al. (2006). "ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation." Bioinformatics **22**(5): 637-640.

Rajewsky, N., M. Vergassola, et al. (2002). "Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo." BMC Bioinformatics **3**(1): 30.

Robinson, M., Y. Sun, et al. (2006). "Improving computational predictions of cis-regulatory binding sites." Pac Symp Biocomput: 391-402.

Sun, Y., M. Robinson, et al. (2005). Integrating binding site predictions using non-linear classification methods. Machine Learning Workshop, Sheffield, LNAI.

Sun, Y., M. Robinson, et al. (2005). Using Real-Valued Meta-classifiers To Integrate Binding Site Predictions. IJCNN, Montreal, CA.

Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol **23**(1): 137-44.

Wasserman, W. W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements." Nat Rev Genet **5**(4): 276-87.

Yellaboina, S., J. Seshadri, et al. (2004). "PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes." Nucleic Acids Res **32**(Web Server issue): W318-20.

Zhu, J. and M. Q. Zhang (1999). "SCPD: a promoter database of the yeast Saccharomyces cerevisiae." Bioinformatics **15**(7-8): 607-11.