Identifying Binding Sites in Sequential Genomic Data

Mark Robinson, Cristina González Castellano, Rod Adams,
Neil Davey, Yi Sun

Science and Technology Research Institute
Univesrity of Hertfordshire
UK
{M.Robinson, R.G.Adams, Y.2.Sun, N.Davey}@herts.ac.uk
c.gonzalezcastellano@yahoo.es

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 integrate the results of individual algorithms specialised for the prediction of binding site locations, can produce significant improvements upon the original algorithms. These results make more tractable the expensive experimental procedure of actually verifying the predictions.

Keywords: Computational Biology, Support Vector Machine, Imbalanced data, Sampling, Transcription Factor Binding Sites

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 domain specific 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.

Computational prediction of *cis*-regulatory binding sites is widely acknowledged as a difficult task [1]. Binding sites are notoriously variable from instance to instance and they can be located considerable distances from the gene being regulated in higher eukaryotes. Many algorithmic approaches are inherently constrained with respect to the range of binding sites that they can be expected to reliably predict. For example, co-regulatory algorithms would only be expected to successfully find binding sites common to a set of co-expressed promoters, not any unique binding sites that might also be present. Scanning algorithms are likewise limited by the quality of the position weight matrices available for the organism being studied. 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. The use of algorithmic predictions prone to high rates of false positive is particularly costly to experimental biologists using the predictions to guide experiments. High rates of false positive predictions also limits the utility of prediction algorithms for their use in network reconstruction. Reduction of the false positive rates is therefore a high priority.

In this paper we show how the algorithmic predictions 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 around 80%.

2 Background

The use of a non-linear classification algorithm for the purposes of integrating the predictions from a set of *cis*-regulatory binding site prediction algorithms is explored in this paper. This is achieved by first running a set of established prediction algorithms, chosen to represent a range of different algorithmic strategies, on a set of annotated promoter sequences. Subsequently, an SVM is trained to classify individual sequence positions as a component of either a binding site or the background sequence. The set of predictions from the original algorithms, appropriately sampled to account for the imbalanced nature of the data set, and labeled with experimental annotations is used for the training inputs.

A wide range of binding site prediction algorithms were used in this study. They were selected to represent the full range of computational approaches to the binding site prediction problem. The algorithms chosen were either reported in the literature or were developed in-house or by our collaborators in the case of PARS, Dream and Sampler. Table 1 lists the algorithms used along with references. Where possible, parameter settings for the algorithms were taken from the literature, if not available, default settings were used.

Table 1. The 12 Prediction Algorithms used.

Strategy	Algorithm
Scanning algorithms	Fuzznuc MotifScanner [2] Ahab [3]
Statistical algorithms	PARS Dream (2 versions) [4] Verbumculus [5]
Co-regulatory algorithms	MEME [6] AlignACE [7] Sampler
Evolutionary algorithms	SeqComp [8] Footprinter [9]

3 Description of the Data

Experimentally annotated sequences were used in this study. The yeast, *S.cerevisiae* was selected for the model organism; the use of this particularly well studied model organism ensures that the annotations available are among the most complete available. 112 annotated promoter sequences were extracted from the *S.cerevisiae* promoter database [10] 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 [10]. 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 [11].

Predictions made by the original algorithms across the dataset were placed into a matrix consisting of 67,782 12-ary real valued vectors, each associated with a binary label indicating the presence or absence of an experimental annotation at that position, see Figure 1.

Each 12-ary vector represents the predictions from all 12 original algorithms for a particular position in the dataset. All predictions in the matrix were normalised as real values in the range [-1,1] with the value of 0 allocated to sequence positions where an algorithm was unable to be run. 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 112 annotated promoter sequences. We use a window size of 7, 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 and would otherwise pose a significant obstacle to the training of the classifiers.

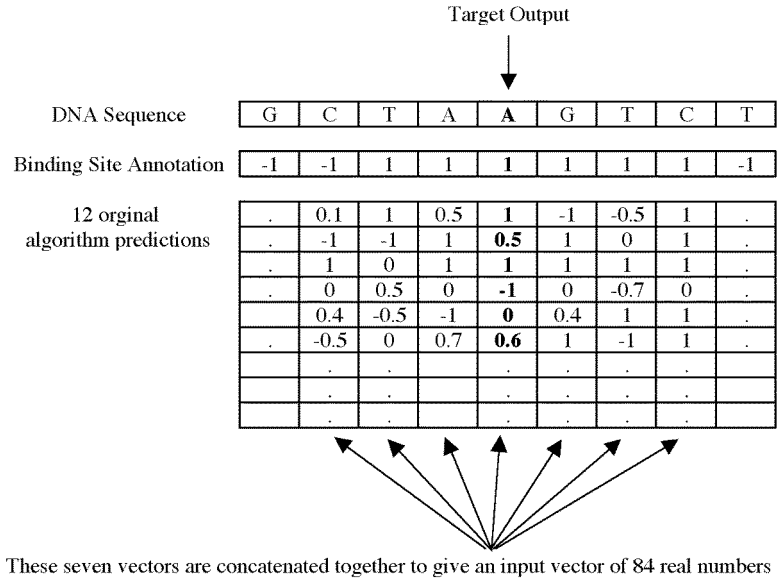

Fig. 1. The formation of the windowed data. The 12 predictions from the original algorithms for the target site are concatenated with the predictions from the 3 sites on either side. This gives an input vector of 12 by 7 real numbers. The corresponding label of this vector is the annotation of the central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

A number of statistics summarising the dataset are shown in Table 2.

Table 2. Summary of the data used.

Total number of sequences	112
Total sequence length	67782 bp
Average sequence length	605 bp
Average number of TFBS sites per sequences	3.6
Average TFBS width	13.2 bp
Total number of TFBS sites	400
Number of unique TFBS sites	69
TFBS density in total dataset	7.8%

4 Performance Metrics

As approximately 8% of the dataset (see Table 2) is annotated as being a part of a binding site, this dataset is imbalanced. If the algorithms are to be evaluated in a useful manner simple error rates are inappropriate, it is therefore 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 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 Techniques for Learning Imbalanced Datasets

Without addressing the imbalance of the two classes in the data, classifiers will produce negligible true positive predictions. This is due to the fact that predicting that every base-pair is not part of a binding site will give high accuracy, being correct 92% of the time (with no false positives). However such a predictor is obviously worthless.

In this paper we address the problem of our imbalanced data in two ways: firstly by using data based sampling techniques [12, 13] and secondly by using different SVM error costs for the two classes [14].

5.1 Sampling Techniques

One way to address imbalance is simply to change the relative frequencies of the two classes by under sampling the majority class and over sampling the minority class. Under sampling the majority class can be done by just randomly selecting a subset of the class. Over sampling the minority class is not so simple and here we use the Synthetic Minority Oversampling Technique (*SMOTE*) [12]. 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. In the first experiment the number of items in the minority class was first doubled and the

number of randomly selected majority class members was then set to ensure that the final ratio of minority to majority class was 0.5. This value was selected using 5-fold cross validation experiments.

5.2 Different SVM error costs

In the standard SVM the primal Lagrangian that is minimized is:

$$L_p = \frac{\left\|\mathbf{w}^2\right\|}{2} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\mathbf{w}.\mathbf{x}_i + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n} r_i\xi_i$$

$$\text{subject to: } 0 \le \alpha_i \le C \text{ and } \sum_{i=1}^{n}\alpha_i y_i = 0$$

Here ***C*** represents the trade-off between the empirical error, $\boldsymbol{\xi}$, and the margin. The problem here is that both the majority and minority classes use the same value for *C*, which as pointed out by Akbani et al [15] will probably leave the decision boundary too near the minority class. Veropoulos et al [14] suggest that having a different *C* value for the two classes may be useful. They suggest that the primal Lagrangian is modified to:

$$L_p = \frac{\left\|\mathbf{w}^2\right\|}{2} + \quad C^+\sum_{\{i|y_i=+1\}}\xi_i + \; C^-\sum_{\{i|y_i=-1\}}\xi_i - \sum_{i=1}\alpha_i\left[y_i\left(\mathbf{w}.\mathbf{x}_i + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n} r_i\xi_i$$

$$\text{subject to: } 0 \le \alpha_i \le C^+ \text{ if } y_i = +1, \quad 0 \le \alpha_i \le C^- \text{ if } y_i = -1 \quad \text{and } \sum_{i=1}^{n}\alpha_i y_i = 0$$

Here the trade-off coefficient *C* is split into C^+ and C^- for the two classes, allowing the decision boundary to be influenced by different trade-offs for each class. Thus the decision boundary can be moved away from the minority class by lowering C^+ with respect to C^-.

Akbani et al [15] argue that using this technique should improve the position of the decision boundary but will not address the fact that it may be misshapen due to the relative lack of information about the distribution of the minority class. So they suggest that the minority class should also be over-sampled, using SMOTE, to produce a method they call SMOTE with Different Costs (SDC). This is one of the techniques we evaluate here.

6 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.

7 Results

Before presenting the main results we should point out that predicting binding sites accurately is extremely difficult. The performance of the best individual original algorithm (Fuzznuc) is:

	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 the key aim of our combined classifier is to significantly reduce the number of false positives given by the original algorithms.

7.1 Results Using Sampling

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. This gives a training set of 32,615 84-ary vectors and a test set of 16,739 vectors.

In the results here 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. 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 ($C = 1000$, $\gamma = 0.001$), the test set was presented and the results are as follows:

	Recall	Precision	F-Score	FP-Rate
Best Original Algorithm	0.400	0.222	0.285	0.106
Combined Classifier - Sampling	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. However this is still too large a figure to make the classifier useful to biologists. So we turn to our second Combined Classifier using *SDC*.

7.2 Results Using *SDC*

First the minority class was over-sampled using SMOTE. The size of the minority class was tripled to 7731 vectors so that the ratio of majority to the minority class was now about 4 : 1. Once again 5-fold cross validation was used to find appropriate values for the three free parameters of the SVM with different costs, namely C^+, C^- and γ. The best values found were: $C^+ = 680$, $C^- = 1320$ and $\gamma = 0.0001$.

	Recall	Precision	F-Score	FP-RATE
Best Original Algorithm	0.400	0.222	0.285	0.106
Combined Classifier - Sampling	0.305	0.371	0.334	0.044
SDC	0.283	0.375	0.324	0.036

This method has produced a good classifier, but it is not much better than the classifier using a straightforward SVM and sampling. However the FP-Rate has been further reduced to 0.036.

7.3 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 SDC algorithm and remove all those that do not meet this criterion.

	Recall	Precision	F-Score	FP-Rate
Best Original Algorithm	0.400	0.222	0.285	0.106
Combined Classifier - Sampling	0.305	0.371	0.334	0.044
SDC	0.283	0.375	0.324	0.036
SDC + Post-Processing	0.264	0.517	0.350	0.021

This produces our best result by some way. The Precision of the prediction has been increased to 0.517 and the FP-Rate is now down to just 2%.

To see how this has come about Figure 2 shows a fragment of the genome with the original algorithmic predictions, the SVM predictions, the result of post-processing the SVM predictions and the actual annotation. It can be seen that for this fragment the removal of the implausible predictions eliminates almost all the false positive predictions.

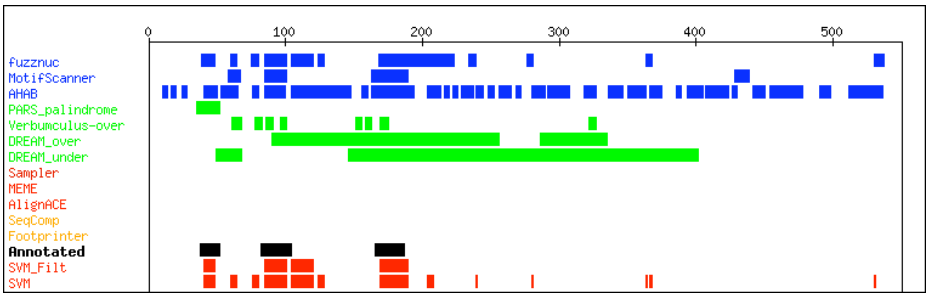

Fig. 2. A fragment of the genome with the 12 original predictions, the actual annotations in black. The last row shows the predictions of the SVM and above it the effect of removing unrealistically short predictions.

8 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 particular domain.

On the technical issue of dealing with the highly imbalanced data we found that both sampling of the two classes and using the SDC algorithm gave similar results, with both methods dealing well with our data.

References

[1] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, 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, vol. 23, pp. 137-44, Jan 2005.

[2] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau, "A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling," Bioinformatics, vol. 17, pp. 1113-22, Dec 2001.

[3] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia, "Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo," BMC Bioinformatics, vol. 3, p. 30, Oct 24 2002.

[4] I. Abnizova, A. G. Rust, M. Robinson, R. Te Boekhorst, and W. R. Gilks, "Transcription binding site prediction using Markov models," J Bioinform Comput Biol, vol. 4, pp. 425-41, Apr 2006.

[5] A. Apostolico, M. E. Bock, S. Lonardi, and X. Xu, "Efficient detection of unusual words," J Comput Biol, vol. 7, pp. 71-94, Feb-Apr 2000.

[6] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," Proc Int Conf Intell Syst Mol Biol, vol. 2, pp. 28-36, 1994.

[7] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend, "Functional discovery via a compendium of expression profiles," Cell, vol. 102, pp. 109-26, Jul 7 2000.

[8] C. T. Brown, A. G. Rust, P. J. Clarke, Z. Pan, M. J. Schilstra, T. De Buysscher, G. Griffin, B. J. Wold, R. A. Cameron, E. H. Davidson, and H. Bolouri, "New computational approaches for analysis of cis-regulatory networks," Dev Biol, vol. 246, pp. 86-102, Jun 1 2002.

[9] M. Blanchette and M. Tompa, "FootPrinter: A program designed for phylogenetic footprinting," Nucleic Acids Res, vol. 31, pp. 3840-2, Jul 1 2003.

[10] J. Zhu and M. Q. Zhang, "SCPD: a promoter database of the yeast Saccharomyces cerevisiae," Bioinformatics, vol. 15, pp. 607-11, Jul-Aug 1999.

[11] M. Robinson, Y. Sun, R. T. Boekhorst, P. Kaye, R. Adams, N. Davey, and A. G. Rust, "Improving computational predictions of cis-regulatory binding sites," Pac Symp Biocomput, pp. 391-402, 2006.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[13] P. Radivojac, N. V. Chawla, A. K. Dunker, and Z. Obradovic, "Classification and knowledge discovery in protein databases," J Biomed Inform, vol. 37, pp. 224-39, Aug 2004.

[14] K. Veropoulos, N. Cristianini, and C. Campbell, "Controlling the Sensitivity of Support Vector Machines," in Proceedings of the International Joint Conference on Artificial Intelligence, (IJCAI99), Stockholm, 1999.

[15] R. Akbani, Stephen Kwek, and Nathalie Japkowicz, "Applying Support Vector Machines to Imbalanced Datase," in 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004, Pisa, 2004, pp. 39-50.