Neural Networks 21 (2008) 856–861

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2008 Special Issue

Combining experts in order to identify binding sites in yeast and mouse genomic data

Mark Robinson [a,b], Cristina González Castellano [a], Faisal Rezwan [a], Rod Adams [a], Neil Davey [a,*], Alastair Rust [c], Yi Sun [a]

[a] Science and Technology Research Institute, University of Hertfordshire, UK
[b] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, USA
[c] Institute for Systems Biology, Seattle, USA

ARTICLE INFO

Article history:
Received 18 January 2008
Received in revised form
18 July 2008
Accepted 29 July 2008

Keywords:
Computational biology
Support vector machine
Imbalanced data
Sampling
Transcription factor binding sites

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 and 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 at 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 et al., 2005).

Computational analysis of DNA sequences typically relies on a string based representation where four characters represent the sequence of nucleotides defining a DNA sequence. The use of string

* Corresponding author.
E-mail addresses: M.Robinson@herts.ac.uk (M. Robinson), c.gonzalezcastellano@yahoo.es (C. González Castellano), F.Rezwan@herts.ac.uk (F. Rezwan), R.G.Adams@herts.ac.uk (R. Adams), N.Davey@herts.ac.uk (N. Davey), arust@systemsbiology.org (A. Rust), Y.2.Sun@herts.ac.uk (Y. Sun).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.07.004

based representations of DNA sequences has made possible the application of a wide range of powerful computational algorithms to the analysis of DNA sequences. A limitation common to many if not all algorithmic approaches is that they 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 limit the utility of prediction algorithms for their use in regulatory network reconstruction. Reduction of the false positive rates is therefore a high priority.

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 datasets: for our major study we use a set of annotated yeast promoters taken from the SCPD (Zhu & 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, Alba, Messeguer, & Guigo, 2006) and ORegAnno (Montgomery et al., 2006) databases.

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 the 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, along with references. 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; once again they represent a range of different algorithmic approaches along with some additional sources of relevant evidence. Table 2 lists the sources of evidence used with the mouse dataset. Each of these sources was extracted from the UCSC genome browser (Karolchik et al., 2003) for the promoter regions of interest. Details of the algorithms can also be found in Karolchik et al. (2003).

Table 1
The 12 Prediction Algorithms used with the yeast dataset

Strategy	Algorithm
Scanning algorithms	Fuzznuc
	MotifScanner (Thijs et al., 2001)
	Ahab (Rajewsky et al., 2002)
Statistical algorithms	PARS
	Dream (2 versions) (Abnizova et al., 2006)
	Verbumculus (Apostolico et al., 2000)
Co-regulatory algorithms	MEME (Bailey & Elkan, 1994)
	AlignACE (Hughes et al., 2000)
	Sampler
Evolutionary algorithms	SeqComp (Brown et al., 2002)
	Footprinter (Blanchette & Tompa, 2003)

Note Dream was run using two different modes of operation.

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, *Saccharomyces 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 & Zhang, 1999) for training and testing the algorithms. For each promoter, 500 base pairs (*bp*) of the sequence taken immediately upstream from the transcriptional start site were considered sufficient to typically allow full regulatory characterisation in yeast (Zhu & Zhang, 1999). 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 et al. (2006).

The dataset for the second experiment uses annotated transcription factor sites for the mouse, *Mus 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. Seven sources of evidence were used as input in this study. *MotifLocator* uses the PHYLOFACTS matrices from the JASPAR database (Wasserman & Sandelin, 2004) to scan for good matches in the sequences. *EvoSelex* uses motifs from Rajewsky, Vergassola, Gaul, and Siggia (2002) and the *Fuzznuc* algorithm to search for consensus sequences. A number of sources of evidence were extracted from the UCSC genome browser (Karolchik et al., 2003): *Regulatory Potential*