ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Discrimination of regulatory DNA by SVM on the basis of over- and under-represented motifs

Rene te Boekhorst [1], Irina Abnizova[2], Lorenz Wernisch[2]

[1]University of Hertfordshire - Dept of Computer Science College Lane, Hatfield - UK
[2]MRC Biostatistics Unit Institute of Public Health, Robinson Way, CB2 2SR, Cambridge - UK

Abstract. In this paper we apply three pattern recognition methods (support vector machine, cluster analysis and principal component analysis) to distinguish regulatory regions from coding and non-coding non regulatory DNA sequences. Using a new feature representation (the degree by which motifs are over- and under-represented) we demonstrate the remarkable power of this methodology in identifying regulatory regions of *Drosophila melanogaster*.

1 Introduction and related work

Variation in the complexity of organisms appears to be due to differences in the regulation of gene activity rather than to differences in the genetic specifications for protein coding per se. Whereas the general principles underlying the translation of the coding regions of genes (exons) into their protein products are largely comprehended, the relationship between a gene’s expression and the information contained in (non-coding) regulatory regions of the genome is not so well understood. These regulatory regions contain Transcription Factor Binding Sites (TFBS), short sequences of DNA which are often located upstream or downstream of the position where gene transcription begins (although regulatory activity may also occur within a gene, or in the case of enhancers, far removed from genes). In turn, these binding sites are “recognized” by transcription factors, proteins that - upon binding to them - act as repressors or activators, thus controlling the rate of transcription. According to [1] as much as 50% of the metazoan genome is regulatory. However, most of this is not yet deciphered as it is extremely difficult to identify regulatory regions computationally. Part of the problem lies in defining appropriate attributes (features) that characterize regulatory regions, but also in the choice of an effective discriminating algorithm. We want to find solutions to this identification problem by considering both a newly developed feature set and exploring the application of various machine learning techniques.

Our feature space is based on the assumption that regulatory regions stand out by the clustering and frequency of TFBSs. Because these TFBSs in turn are short strings of particular nucleotide compositions (so-called “motifs”), we used the statistical over- and under-representation of all possible motifs of a given stretch of DNA as the input and feature vector representing that stretch of DNA. Some classifier methods have recently been employed to predict regulatory motifs [2, 3] and gene regulatory networks [4] as well as to detect functionally similar proteins [5-7]. Among these classification methods, Support Vector Machines (SVM) enjoy an increasing popularity. However, although SVM has been used to identify TFBSs up till now there are no applications of SVM to distinguish regulatory sequences from other types of functional DNA. In addition, we use two non-supervised techniques (hierarchical cluster analysis and principal component analysis) to back up the performance and visualize the results of the supervised SVM.

2 Data

To train and test our classifiers we use three data sets. The positive training set is a collection of 60 experimentally verified functional *Drosophila melanogaster* regulatory regions [8] located far from gene coding sequences and transcription start sites (i.e. enhancers instead of promoters). It contains the most significant clusters of binding sites for five transcription factors (Bicoid, Hunchback, Kruppel, Knirps and Caudal) involved in the regulation of developmental genes. The total size of the positive training set comprises about 68 Kb of sequence data. The two negative training sets are: (i) 60 randomly picked *Drosophila* internal exons, and (ii) 60 randomly picked *Drosophila* non-coding, non-regulatory (NCNR) sequences using the Ensembl Genome Browser (http://www.ensembl.org/). For the latter, we left out exons and, to exclude possible promoters, regions 1Kb upstream and downstream of genes. Each training set contains 68 Kb of sequences in total. A detailed description of the selection procedure can be found in [9].

3 Methods

3.1 Feature representation

We represent each of the $i = 1, 2, \ldots, 180$ sequences in our training sets by the *n*-dimension vector

$$F(seq_i) = (Z^j_1, Z^i_2, \ldots, Z^i_n) \qquad (1)$$

The elements Z^i_j of this vector measure the degree of over- or under-representation ("Conspicuousness") of all possible "words" of a length of *m* nucleotides (that is, all the $j = 1, 2, \ldots, 4^m$ permutations of A, C, T and G) for sequence *i*. In this paper we fixed word length at three (implying $n = 64$ possible words), and allowed for at most one mismatch. The conspicuousness of a word was assessed as the normalized difference between the observed and expected number of occurrences of that word given single nucleotide independence, i.e.

$$Z = \frac{N(X) - E(X)}{Var(X)} \qquad (2)$$

where $N(X)$ is the frequency of a word X and $E(X)$ and $Var(X)$ are its expectancy and variance respectively. For a formal derivation of the expectancy, see [10].

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

3.2 Support Vector Machine (SVM)

Essentially, SVMs [11, 12], like perceptrons and neural networks, allows the classification of a new sequence based on a training model. A core component of a SVM [13] is the kernel function, which takes a similar role as the activation function in the perceptron. However, whereas the activation function is based on a linear combination of the coordinates of each separate object (input) to be classified, the kernel takes into account relationships among objects as the pair wise similarity between them. A kernel function is derived by first choosing an appropriate feature space, representing each sequence as a vector in this space, and then by taking the inner product (or a function derived from it) between these vector representations. By defining non-linear kernels, the SVM is able to achieve non-linear separability and in this sense outperforms the perceptron. Furthermore, apart from constructing a separating hyper plane, the SVM finds the (support) vectors that define the maximal margin, the distance to two parallel hyper planes on each side of the hyper plane. The novelty of our approach is the *Z*-score representation of a DNA sequence as the feature vector and input space in a two-step approach. First, we separated coding from non-coding DNA with the help of the model ‘coding versus rest' DNA. Next, within the class of non-coding DNA, we made a further distinction by means of a ‘regulatory versus non-regulatory' model. To train these models, we submitted half of our experimentally verified sequences to the models, using other half for testing. Training and testing was carried out by the package Libsvm [15], with a default Gaussian RBF kernel function and the soft margin option. The parameters were adapted by 5-fold cross-validation.

3.3 Cluster analysis

Given the vector representation of each sequence (1), a similarity matrix was created by calculating the pair wise Euclidian distances in the 64-dimensional space between all the 180 sequences. Those sequences that are closest together are unified in initial clusters. Ward’s algorithm was used to combine these initial clusters into clusters at higher levels of dissimilarity. Ward’s method is an unsupervised hierarchical clustering procedure by which those clusters are combined that minimize the variance of the distance between its members. If over- and under-representation are indeed important attributes of regulatory regions, we expect these regions to be into taken together in one cluster and to be separated from the other, non-regulatory sequences of DNA.

3.4 Principal Component Analysis

PCA is a classification method that reduces dimensionality by finding the principle components that explain a predefined proportion of the variance of the data. Principle components are linear combinations of the original variables, the 64 features in our sequence representation (1).
Ideally, applied to our data, three PCs (representing the three types of DNA) should explain a major proportion of the variance of the Z-scores.

4 Results

4.1 SVM results

With the two-step procedure described in the **Methods**, we obtained a very good separation of coding DNA from other DNA types with an overall accuracy 97 % at the first step (see the ROC-curve at the left in Figure 1):

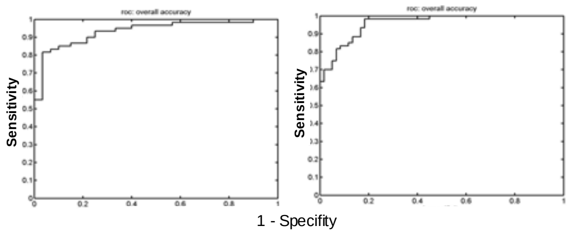

Fig. 1: ROC (overall threshold-independent accuracy) curves for coding and non-coding DNA separation (left), and regulatory and NCNR separation (right).

The second step predicted regulatory DNA with a 95 % overall accuracy (see the ROC-curve at the right in Figure 1)

The performance of the SVM classifier is summarized in Table 1.

MODEL/ TEST SET PREDICTIONS	SENSITIVITY Exons	SENSITIVITY Regulatory	SENSITIVITY NCNR	OVERALL ACCURACY
Exons vs Non-Coding	85.3	95.3	96.5	97
Regulatory vs NCNR		93.3	85.6	95

Table 1. Sensitivity and overall accuracy (in percentage) of the SVM prediction results for different models and training sets.

The predictions depend on a threshold, pre-computed by the SVM classifier. Here the intersection between specifity and sensitivity was taken as the optimal threshold value. Overall accuracy is the area under the ROC curve and is threshold independent.

4.2 Cluster analysis

The dendrogram in Figure 2 clearly shows two distinct clusters and separates coding from regulatory regions well. The first cluster contains only 5 of all 60 coding regions, whereas the second virtually lacks regulatory regions. Non-coding non-regulatory regions are not separated and are equally spread over the two clusters

4.3 PCA

Similar to clustering analysis results, PCA separates coding from regulatory DNA.

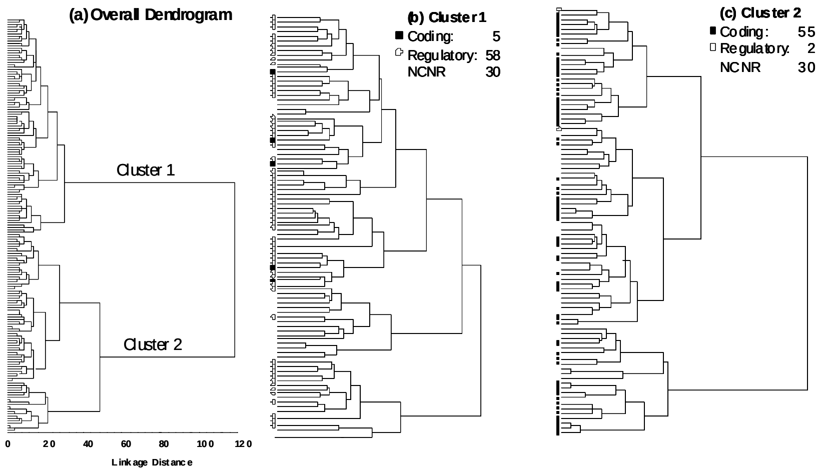

Fig. 2. Dendrograms of the cluster analysis (Euclidean distance, Ward's method) of the 180 sequences.

Eleven principle components were needed to explain ~70% of the variance. However, half of this is concentrated in the first PC. Plotting the first two PCs against each other, shows a separation between regulatory regions (which have high values on the first axis) and coding regions (Figure 3).

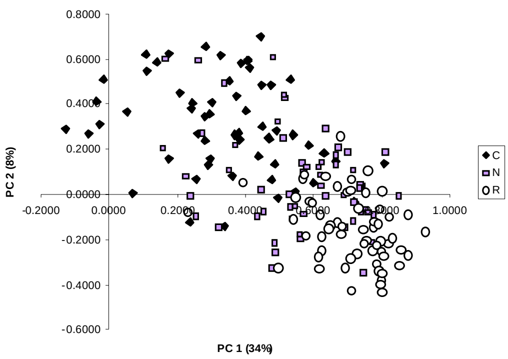

Fig.3. PCA on the 180 sequences. The first two principle components account for 42% of the variance.

5 Discussion and conclusions

Given the difficulty in distinguishing regulatory modules, as reflected in the generally poor results of a number of popular algorithms to discover TFBSs [16], our methodology yields a surprisingly good differentiation between functional DNA sequences. It outperforms SVM applications based on string [4] and mismatch kernels [5]. The latter worked well for the detection of functionally similar proteins, but achieved no more than about 50% accuracy when we trained them on our data.

Apparently, the over- and under-representation of words is a critical feature of regulatory regions that probably has to do either their mode of operation and intrinsic statistical properties. However, although our method gave a good discrimination, the number of support vectors needed to define the margins was quite large, which points to the noisy nature of the data. This is also evident from the rather low proportion of the overall variance accounted for by the first two principal components. Furthermore, because our positive training set is based solely on enhancers involved in the early development of Drosophila, the results may be species, tissue and phase specific. Experimentally verified data from a larger range of species and conditions are therefore needed to support the generality of our findings.

References

[1] M. Markstein, P. Markstein, V. Markstein and M. S. Levine, Genome-wide analysis of clustereddorsal binding sites identifies putative target genes in the Drosophila embryo. *PNAS*, 99: 763-768, 2000.

[2] J. P. Vert., R. Thurman and W. S. Noble, Kernels for gene regulatory regions, *Advances in Neural Information Processing Systems 1,* 2005.

[3] B. Jiang , M. Q. Zhang and X. Zhang, OSCAR: one-class SVM for accurate recognition of ciselements. *Bioinformatics*, 23(21): 2823-8, 2007.

[4] J. Qian , J. Lin, N. M. Luscombe , H. Yu and M. Gerstein , Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 19 (15): 1917-1926, 2003

[5] C. Leslie, E. Eskin, and W. Noble, The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput*., 564–575, 2002.

[6] C. Leslie, E. Eskin, A. Cohen, J. Weston and W. Noble, Mismatch string kernels for svm protein classification. *Adv. Neural Inf. Process. Syst*, 20: 467–476, 2003

[7] T. Jaakkola, M. Diekhans, M. and D. Haussler, A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7: 95-114, 2000.

[8] A. Nazina and D. Papatsenko, Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency. *BMC Bioinformatics* 22: 4-65, 2003.

[9] R. te Boekhorst, I. Abnizova and C. L. Nehaniv, Discriminating coding, non-coding and regulatory regions using rescaled range and detrended fluctuation analysis, *Biosystems* (in press), 2007.

[10] I. Abnizova, R. te Boekhorst and L. Wernisch, Discriminating of regulatory DNA by Support Vector machine. (Submitted to *Bioinformatics*).

[11] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York, 1998.

[12] H. Rangwala and G. Karypis, Profile-based direct kernels for remote homology detection and fold recognition, *Bioinformatics,* 21(23): 4239 – 4247, 2005.

[13] B. Scholkopf and A. Smola, *Learning with Kernels*, MIT Press, Boston, MA, 2002.

[14] A. Webb, Statistical Pattern Recognition, Wiley, Chichester, U.K, 2003

[15] C. C. Chang and C. J. Lin, LIBSVM : a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[16] M. Tompa et al, Assessing computational tools for the discovery of transcription factor binding sites, Nature Biotechnology 23 (1): 137 – 144, 2005.