Programming Research Group

FINDING TRANSCRIPTION FACTOR BINDING SITES IN DNA SEQUENCES: A TEMPLATE BASED APPROACH

Sumedha Gunewardena
Peter Jeavons

PRG-RR-03-RR-03-21

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford OX1 3QD

Abstract

Motivation: A problem faced by many algorithms for finding transcription factor binding sites is the high number of false positive hits that result with the increased sensitivity of their prediction. A main contributing factor to this is the short and degenerate nature of these sites which results in a low signal to noise ratio. In order to counter this problem one needs to look beyond the base independence assumption. We propose a model based on templates designed to capture not only the vertical consensus but also the correlation of individual bases with the other bases of the site.
Results: We use the *Escherichia coli* catabolite gene activator protein (CAP) as our primary test system. Results are also shown for $NF-\kappa B$, $AP-1$, $NF-1$ and C/EBP families of binding sites.
Availability: Contact the author at the address below.
Contact: Sumedha.Gunewardena@comlab.ox.ac.uk

1 Introduction

In this paper, we describe a template-based approach for finding transcription factor binding sites in DNA sequences.

The binding of transcription factors (TF) to their cognate sites is an important part of the transcription regulation process, a process important for the control of many cellular functions including cell differentiation, the cell cycle, and carcinogenesis. Understanding how these transcriptional regulatory proteins recognize specific DNA sequences is of great interest in molecular biology, and one of the major challenges in bioinformatics is to develop effective computational methods for identifying these gene regulatory regions. Transcription factor binding sites are relatively short (10-20bp) and highly degenerate sequences, which makes their effective identification a computationally challenging task.

Initial methods for identifying TF binding sites were based on consensus sequences [15]. (A critical comparison of several different methods for generating such consensus sequences can be found in [4].) Consensus sequences are good for representing well-conserved sites such as restriction endonucleases, but prove to be less effective for identifying the more variable TF binding sites. Position specific weight matrices, initially used for identifying E.coli transcription and translation initiation sites [20, 8] and later used for finding TF binding sites [14, 3, 6, 17], have been shown to be more effective than simple consensus sequences. Some other approaches which have been used to identify TF binding sites include rule-based systems [19], neural networks [5, 12, 9], expectation maximisation [1, 7], and Gibbs sampling [10].

Experimental studies carried out on various protein-binding domains in DNA sequences have shown that in many cases there are context-dependent effects and interactions between different nucleotide positions [24, 11, 22]. This has led many researchers to question the independence assumption on which techniques such as consensus sequences and weight matrices for identifying TF binding sites are based [2, 11, 22]. This issue has been addressed by various authors with a number of more sophisticated techniques including improved weight matrices with prior information on correlated nucleotide positions [25, 18], neural networks [9], and principal coordinates analysis [22].

In this paper we propose a new method based on *templates*, which is designed to provide a representation (of the binding sites) that captures the dependencies between different nucleotide positions. The templates we use are not modelled directly on the nucleotide sequences of the binding sites; instead, we work with structural representations of these sequences obtained by translating the original sequences of nucleotides into numerical sequences derived from a number of different di-nucleotide parameters. The di-nucleotide parameters we use represent various geometric and structural properties associated with each possible pair of consecutive nucleotide bases (e.g., propeller twist, stacking energy, protein induced deformability, etc.) The values of these parameters are obtained from molecular dynamic simulation experiments [13]. By translating our DNA sequences into numerical sequences based on di-nucleotide parameters, we obtain numerical representations which already capture certain correlations between consecutive nucleotide positions.

Using these numerical representations we compute a number of different *templates* to describe the binding sites in a training set containing several examples of known binding sites for a particular TF. These templates are modelled to capture the linear interactions present between different nucleotide positions in the site (see the Method section for details). Once we have these templates we can compare any unknown sequence of nucleotides with each template and obtain a measure of the goodness of fit. By doing this for each of the templates, we obtain a vector which can be used to classify the unknown sequence as a binding sequence or a non-binding sequence. This classification is obtained by using a standard minimum distance classifier based on Mahalanobis distance. To train the classifier we use further positive examples of binding sites, together with a set of negative examples, that is, sequences which are not binding sites. An overview of the entire process is shown in Figure 1.

Using this novel template-based technique on the binding site data for the particular transcription factor investigated in [21], we were able to achieve remarkably high accuracy in the identification of binding sites in the test set (see the Results section). Furthermore, the same method was found to give very good results for a number of other transcription factors, from both eukaryotes and prokaryotes, using data on binding sites obtained from the SAMPLES database [23], as shown in Table 2. Hence we suggest that this method is a very promising new approach to the identification of transcription factor binding sites.

2 Method

2.1 Preparation of Training Sequences

Given a set of nucleotide sequences which are known binding sites for a particular TF (possibly of varying length), we use the alignment program in Target Explorer [16] to obtain an aligned set of sequences of a fixed length, $(L+1)$.

We also need a further set of DNA sequences of the same fixed length, $(L+1)$, which are not binding sites. These negative examples are generated using the random sequence generator in the SAMPLES system in the Molecular Biology Server [23].

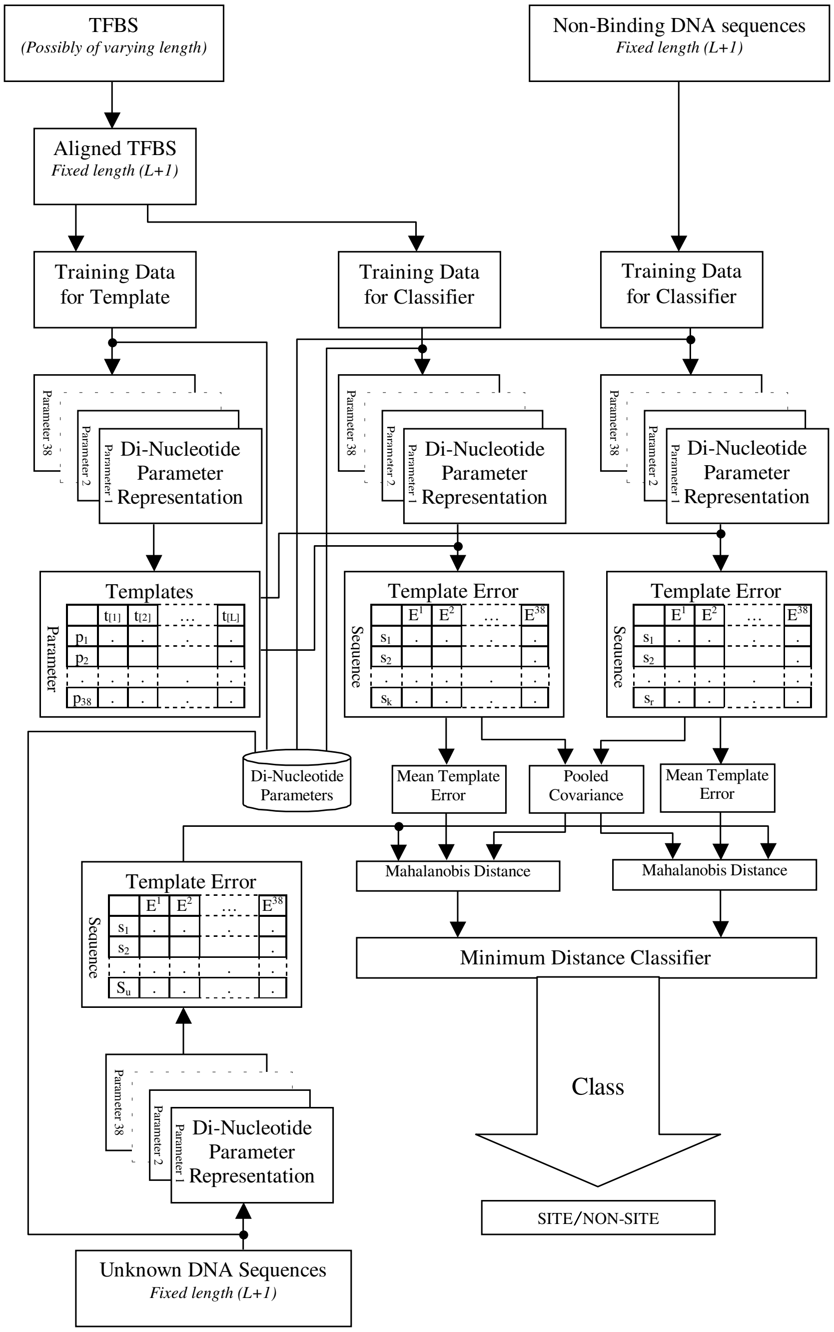

Figure 1: Data Flow Diagram.

We do not work directly with the nucleotide sequences; instead we work with numerical representations of them. Each numerical representation is a sequence of numbers obtained by translating each successive di-nucleotide pair in the original DNA sequence into a corresponding numerical value, using a table of values defining a particular di-nucleotide parameter. Note, that the representation obtained in this way is a numerical vector of length L. For a given nucleotide sequence, s, and a given di-nucleotide parameter p, the resulting vector will be denoted $r^p(s)$.

In fact, we use all 38 di-nucleotide parameter tables available from the Property database [13], and hence obtain 38 different numerical representations for each nucleotide sequence in the training set. For each nucleotide sequence s, the representations of s that we work with will be denoted $r^1(s), r^2(s), \ldots, r^{38}(s)$.

2.2 Computing the Templates

The key to our method is the use of numerical templates to capture certain key features of the binding sites.

Definition 2.1 *A* template, *t, of length L, is simply a numerical vector,* $(t[1], t[2], \ldots, t[L])$.

For any numerical vector $r = (r[1], r[2], \ldots, r[L])$, the prediction error *of r with respect to the template t, denoted $E(r,t)$, is defined as follows:*

$$E(r,t) = \frac{1}{L} \sum_{j=1}^{L} \left\{ \left(\sum_{\substack{i=1 \\ i \neq j}}^{L} t[i] r[i] \right) - r[j] \right\}^2$$

Note that the prediction error of a vector r with respect to a template t is zero exactly when each co-ordinate value in r can be expressed as a linear combination of the other co-ordinate values in r, with fixed coefficients which are given by the values in the template t.

For any set of numerical vectors, $\{r_1, r_2, \ldots, r_n\}$, the mean value of the prediction error with respect to a fixed template t is given by

$$\frac{1}{n} \sum_{k=1}^{n} E(r_k, t)$$

Hence the template which minimises this mean error value for this set of vectors can be obtained by calculating the partial derivatives of this equation with respect to $t[1], t[2], \ldots, t[L]$ and setting each of these equal to zero. This gives the following set of L linear equations

$$\begin{pmatrix} (L-1)C_{1,1} & (L-2)C_{1,2} & \cdots & (L-2)C_{1,L} \\ (L-2)C_{2,1} & (L-1)C_{2,2} & \cdots & (L-2)C_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ (L-2)C_{L,1} & (L-2)C_{L,2} & \cdots & (L-1)C_{L,L} \end{pmatrix} \begin{pmatrix} t[1] \\ t[2] \\ \vdots \\ t[L] \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_L \end{pmatrix} \tag{1}$$

where $C_{i,j} = \sum_{k=1}^{n} r_k[i] r_k[j]$ $(1 \leq i, j \leq L)$ and $b_i = \sum_{\substack{j=1 \\ j \neq i}}^{L} C_{i,j}$ $(1 \leq i \leq L)$.

These equations are symmetric and can be solved efficiently to find the template values $t[1], t[2], \ldots, t[L]$ which minimise the mean prediction error for this set of vectors.

To obtain a suitable collection of templates for identifying the binding sites of a given transcription factor, we randomly select a set $\{s_1, s_2, \ldots, s_n\}$ of nucleotide sequences of fixed length $L+1$ from the training set of known binding sites for this transcription factor. Then, for each di-nucleotide parameter, p, we construct a set of n numerical vectors of length L, $\{r^p(s_1), r^p(s_2), \ldots, r^p(s_n)\}$, as described above. For each of these sets of vectors, we calculate a corresponding template, t^p, using Equation 1. Hence we obtain 38 different templates, $t^1, t^2, \ldots, t^{38}$, which are used to classify unknown nucleotide sequences, as described below.

2.3 Training the Classifier

Given any nucleotide sequence x of length $L+1$, we can represent it as described above by using 38 numerical vectors of length L, $r^1(x), r^2(x), \ldots, r^{38}(x)$. For each of these numerical vectors we compute the prediction error with respect to the corresponding template, and hence obtain a vector of 38 prediction errors, which we call the *error vector* for x, and denote by $v(x)$. That is,

$$v(x) = \left(E(r^1(x), t^1), E(r^2(x), t^2), \ldots, E(r^{38}(x), t^{38})\right). \tag{2}$$

The error vector $v(x)$ is used to classify the nucleotide sequence x as a binding site or a non-binding site, using a standard 2-class classifier based on the notion of Mahalanobis distance, as described below.

To train the classifier we take the remaining positive examples of binding sites in the training data (those not used to compute the templates), and calculate the mean value of the error vector over these sequences. This gives a single vector to represent the class of binding sites, which we denote m_S.

Similarly, we take the set of negative examples, as described above, and compute the mean value of the error vector for these sequences. This gives a single vector to represent the class of non-binding sites which we denote m_N

Finally, we compute the covariance matrix for all of these error vectors, which is denoted Q.

The Mahalanobis distance from $v(x)$ to the class of binding sites is given by

$$\left\{(v(x) - m_S)^T Q^{-1} (v(x) - m_S)\right\}^{\frac{1}{2}} \tag{3}$$

The Mahalanobis distance from $v(x)$ to the class of non-binding sites is given by

$$\left\{(v(x) - m_N)^T Q^{-1} (v(x) - m_N)\right\}^{\frac{1}{2}} \tag{4}$$

The unknown nucleotide sequence x is classified as a binding site or a non-binding site according to whichever of these distances is smaller.

Model	Sensitivity (Test set)	Sensitivity (Test+Training)	Specificity (Test set)
PWM	0.67	0.88	0.96
MDS-HMM		0.92	0.90
Template	1.00	1.00	0.96

Table 1: Results for CAP binding sites.

TF	Train	Classify		Test		Sensitivity	Sensitivity	Specificity	Specificity
		T	F	T	F	Test	Test+Tra	Test	Test+Tra
NF-kB	10	10	130	16	130	0.93 (0.05)	0.95 (0.03)	0.97 (0.01)	0.98 (0.01)
CAP	8	8	130	9	110	0.90 (0.11)	0.95 (0.04)	0.95 (0.10)	0.97 (0.10)
NF-1	15	15	130	41	130	0.85 (0.05)	0.95 (0.02)	0.87 (0.03)	0.96 (0.01)
CEBP	15	15	130	49	130	0.81 (0.06)	0.85 (0.04)	0.90 (0.03)	0.93 (0.02)
AP-1	10	10	130	21	130	0.76 (0.09)	0.84 (0.05)	0.92 (0.04)	0.95 (0.02)

Table 2: Statistics for 5 different transcription factors.

3 Results and Discussion

3.1 Results for CAP Binding Sites

We use the *Escherichia coli* catabolite gene activator protein (CAP) as our primary test system. CAP is a DNA binding protein involved in bacterial regulation that triggers the transcription of catabolite operons in the absence of glucose. It recognises a 22bp two-fold-symmetric DNA site with the consensus sequence AAA*TGTGA*TCT AGA*TCACA*TTT. This sequence has a highly conserved *TG TGA* motif and its inverted repeat *TCACA*, symmetrically placed from the centre of the polymer. Data for the CAP binding sites were obtained from Thayer & Beveridge [21].

Thayer & Beveridge describe a model for identifying CAP binding sites that incorporates a molecular dynamic description of sequence-dependent structure into a hidden Markov model. We will refer to this as the MDS-HMM model.

To enable a straightforward comparison with the results obtained by Thayer & Beveridge we used exactly the same training data and test data used by them. Their data consisted of 25 examples of CAP binding sites and 110 eukaryotic non-binding sites. As the 110 eukaryotic non-binding sites were required for testing, we generated an extra 130 eukaryotic non-binding sites to train our classifier. Our training set comprised of the 16 positive examples used by Thayer & Beveridge for their training. From this training set, 8 sequences were used to compute the templates, and the other 8 were used to train the classifier. Following Thayer & Beveridge, we used all 25 positive examples (16 from the training set and 9 unseen sequences) together with the 110 eukaryotic non-binding sites for testing.

As a further comparison, we investigated the use of a positional weight matrix (PWM) for this data. We used Target Explorer [16] for this purpose. The 16 positive examples used for training the other two models were used to build the PWM. This matrix was then used to search the same 25 known binding sites and 110 eukaryotic non-binding sites.

The results for the three models are shown in Table 1. Our template model was able to correctly identify all 25 binding sites in the test data. It was also able to correctly classify 106 of the 110 non-binding sites. These figures correspond to a *sensitivity*[1] of 100% and a *specificity*[2] of 96.36% which compares very favourably with the sensitivity of 92.0% and specificity of 90.0% reported by Thayer & Beveridge for their MDS-HMM model.

The PWM model was able to correctly identify 22 of the 25 binding sites in the test data, but only 6 of these were from the unseen data (i.e., from the 9 TF binding sites not included in the training set). The PWM model was also able to correctly classify 106 of the 110 non-binding sites. These results correspond to a sensitivity of 88.0% and a specificity of 96.36%, which is again poorer than the template model.

3.2 Results for Other Binding Sites

We further tested our model on the following transcription factors using data on binding sites obtained from the SAMPLES system in the Molecular Biology Server [23]:

- Nuclear factor kappa B ($NF-\kappa B/Rel$) ($p50/p65$)

  The $NF-\kappa B/Rel$ complex is a heterodimer composed of two DNA-binding subunits ($NF-\kappa B1$ and $relA$). It plays a key role in the regulation of many genes that code for mediators of the immune, acute phase and inflammatory responses. Homo- and heterodimers of members of the $NF-\kappa B/Rel$ family recognize nucleotide sequences with the consensus $GGGACTTTCC$.

- Activating protein-1 ($AP-1$)

  The dimeric transcription factor $AP-1$ is a multiprotein complex made up of Fos and Jun proteins together with $ATF/CREB$ proteins. AP-1 controls both basal and inducible transcription of several genes. It binds to the consensus sequence $TGACTCA$ (TPA-responsive element (TRE)).

- Nuclear factor-1 ($NF-1$)

  $NF-1$ is a family of transcription factors that have been shown to play important roles in tissue specific transcription of differentiation associated genes in a number of tissues. $NF-1$ proteins bind as dimers to the duplex DNA sequence with consensus $TTGGC$-$NNNNNGCCAA$.

- $CCAAT$ box/enhancer binding protein (C/EBP)

  C/EBP forms dimers and consists of an activation domain, a DNA-binding basic region, and a leucine-rich dimerization domain (leucine zippers). It binds to the $CCAAT$-box with the consensus sequence $GGYCAATCT$.

The results for these additional transcription factors, together with additional results for the CAP binding sites, are shown in Table 2.

To investigate the robustness of our method, we ran every experiment 100 times, randomly selecting the training sequences for the templates and the classifier from the available positive examples on each trial. We performed this rigorous evaluation for the CAP data of the first experiment as well. The results in Table 2 show the mean values for the different statistics, with the standard deviation in brackets.

[1]Sensitivity is defined as $\frac{True\ Positives}{True\ Positives+False\ Negative}$

[2]Specificity is defined as $\frac{True\ Negative}{True\ Negative+False\ Positive}$

3.3 Efficiency and Effectiveness

We have implemented the template model described above in the software system $Matlab^{\circledR}$. The total time required to compute the templates and train the classifier was around five to six seconds for each transcription factor on a PC running at 466MHz. The time required to run the classifier on the test set was around one to two seconds for each transcription factor.

The time complexity of searching a whole genome for possible transcription factor binding sites using the method described here depends on three factors: the length of the genome G, the length of the template L, and the number of structural parameters used P ($P = 38$ in our case). The error vector for a sequence of length $L+1$, defined in Equation 2, can be computed in $\bigcirc(PL^2)$ time. The two Mahalanobis distance measures for sites and non-sites used in the classifier can be computed in $\bigcirc(P^2)$ time. This gives an overall time complexity of $\bigcirc(PL^2 + P^2)$ for processing a single site in the genome. This has to be done for $G - L$ sites in the whole genome being searched. The entire process will therefore have a time complexity of $\bigcirc\left(GP(L^2 + P)\right)$. While G can be relatively large, L and P are generally small. The length of the templates, L, is similar to the length of the binding sites, and would typically be around 10 to 12 (although for the CAP data above it is 32). The number of di-nucleotide parameters, P, plays an important part in the classification phase. In the experiments reported here we used 38 different di-nucleotide representations – all the values available in the SAMPLES database. However, we have found in further trials that this value can be reduced to around 10 to 15 with a minimal effect on the quality of the output, by working with the 10 to 15 principal components of the 38 parameter vectors.

Finally, we speculate on why the template-based method we propose here is so effective for the identification of several families of TF binding sites. There are two contributing factors for the effectiveness of this approach: the template and the classifier trained on that template. To compute the template we make a simplifying assumption of direct proportionality in the variation of the di-nucleotide values between different di-nucleotide positions (i.e, values at different di-nucleotide positions increase or decrease proportionally to each other). This is not necessarily the case between all di-nucleotide positions of a site, but it does capture more information about possible interactions than other methods such as positional weight matrices. Any deviations from this assumption are partly addressed by the classifier.

Further investigation is needed to determine whether the template-based approach described here can be successfully applied to the identification of binding sites for other transcription factors, and whether the method can be refined to achieve even greater accuracy.

References

[1] T. L Bailey and C Elkan. Unsupervised Learning of Multiple Motifs in Biopolymers using Expectation Maximization. *Machine Learning.*, 21(1-2):51–80, 1995.

[2] M. L Bulyk, P. L. F Johnson, and G. M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research.*, 30(5):1255–1261, 2002.

[3] Q. K Chen, G. Z Hertz, and G. D Stormo. MATRIX SEARCH 1.0: A computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, 11:563–566, 1995.

[4] W. H. E Day and F. R McMorris. Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Research.*, 20:1093–1099, 1992.

[5] B Demeler and G Zhou. Neural network optimization for E. coli promoter prediction. *Nucleic Acids Research.*, 19:1593–1599, 1991.

[6] J. W Fickett. Quantitative discrimination of MEF2 sites. *Molecular and Cellular Biology.*, 16(1):437–441, 1996.

[7] W. N Grundy, T. L Bailey, and C. P Elkan. ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool. *Computer Applications in the Biological Sciences (CABIOS).*, 12(4):303–310, 1996.

[8] R Harr, M Haggstrom, and P Gustafsson. Search algorithms for pattern match analysis of nucleic acid sequences. *Nucleic Acids Research.*, 11:2943–2957, 1983.

[9] M Horton, P. B.and Kanehisa. An assessment of neural network and statistical approaches for prediction of E. coli promoter sites. *Nucleic Acids Research*, 20:4331–4338, 1992.

[10] C. E Lawrence, S. F Altschul, M. S Boguski, J. S Liu, A. F Neuwald, and J. C Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science.*, 262(5131):208–14, 1993.

[11] T. K Man and G. D Stormo. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Research.*, 29(12):2471–8, 2001.

[12] M.C O'Neill. Training back-propogation neural networks to define conducted DNA-binding sites. *Nucleic Acids Research.*, 19:313–318, 1991.

[13] J. V Ponomarenko, M. P Ponomarenko, A. S Frolov, D. G Vorobyev, G. C Overton, and N. A Kolchanov. Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics*, 15(7/8):654–668, 1999.

[14] K Quandt, K Frech, H Karas, E Wingender, and T Werner. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research.*, 23:4878–4884, 1995.

[15] J Schug and G. C Overton. TESS: Transcription Element Search Software on the WWW. *Technical Report CBIL-TR-1997-1001-v0.0, of the Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania*, 1997.

[16] A Sosinsky, J Wildonger, K Bonin, R Mann, and B Honig. Target Explorer: An Automated Tool for Identification of New Target Genes for Specified Set of Transcription Factors (poster). *ISMB02, Edmonton, Canada.*, 2002.

[17] G. D Stormo and D. S Fields. Specificity, energy and information in DNA-protein interactions. *Trends Biochemical Sciences.*, 23:109–113, 1998.

[18] G. D Stormo, T. D Schneider, and L Gold. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic. Acids Research.*, 14:6661–6679, 1986.

[19] G. D Stormo, T. D Schneider, L Gold, and A Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research.*, 10:2997–3011, 1982a.

[20] G. D Stormo, T. D Schneider, and Gold L. M. Characterization of translational initiation sites in E. coli. *Nucleic Acids Research.*, 10(9):2971–96, 1982.

[21] K. M Thayer and D. L Beveridge. Hidden Markov models from molecular dynamics simulations on DNA. *Proceedings of the National Academy of Sciences*, 99(13):8642–8647, 2002.

[22] I. A Udalova, R Mott, D Field, and D Kwiatkowski. Quantitative prediction of NF-kB DNA-protein interactions. *Proceedings of the National Academy of Sciences USA.*, 99:8167–8172, 2002.

[23] D. G Vorobiev and O. A Ponomarenko, J.Vand Podkolodnaya. Samples and aligned databases for functional site sequences. *Proc. I Intern. Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Russia*, pages 58–61, 1998.

[24] S. A Wolfe, H. A Greisman, E. I Ramm, and C. O Pabo. Analysis of Zinc Fingers Optimized Via Phage Display: Evaluating the Utility of a Recognition Code. *Journal of. Molecular Biology.*, 285:1917–1934, 1999.

[25] Q. M Zhang and T. G Marr. A weight array method for splicing signal analysis. *Computer Applications in Biosciences.*, 9(5):499–509, 1993.