Accounting for Structural Properties and Nucleotide Co-variations in the Quantitative Prediction of Binding Affinities of Protein-DNA Interactions

Sumedha Gunewardena and Zhaolei Zhang

Pacific Symposium on Biocomputing 11:379-390(2006)

ACCOUNTING FOR STRUCTURAL PROPERTIES AND NUCLEOTIDE CO-VARIATIONS IN THE QUANTITATIVE PREDICTION OF BINDING AFFINITIES OF PROTEIN-DNA INTERACTIONS

SUMEDHA GUNEWARDENA AND ZHAOLEI ZHANG

Charles H. Best Institute, University of Toronto,
112 College Street, Toronto, Ontario,
M5G 1L6, Canada
E-mail: Sumedha@cantab.net

We describe a quantitative model for predicting the binding affinity of protein-DNA interactions. The described model is based on *templates* capable of providing a global representation of the modelled transcription factor (TF) binding sites. Templates can capture non independent nucleotide variations and structural properties present in these sites. Tests carried out on the *p50p50* and *p50p65* variants of the transcription factor *NF-κB* demonstrate a high correlation between the observed binding affinities and the binding affinities predicted by the templates. Only a small subset of training data spanning the space of the binding sites is required to train the templates.

1. Introduction

In human and other higher eukaryotes, gene expression is regulated by the binding of various modulatory transcription factors (TF) onto *cis*-regulatory elements near genes. Binding of different combinations of transcription factors may result in a gene being expressed in different tissue types or at different developmental stages. To fully understand a gene's function, therefore, it is essential to identify the transcription factors that regulate the gene and the corresponding TF binding sites.

TF binding sites are relatively short (10-20bp) and highly degenerate sequences, which makes their effective identification a computationally challenging task. Early methods for identifying TF binding sites were mainly non-quantitative binary classifiers. They ranged from consensus sequences[24] and position specific weight matrices[20,27] to approaches such as rule-based systems[28], Gibbs sampling[12], expectation maximisation[8], neural networks[14,10] and comparative genomics[35,33].

Transcription factors, unlike restriction enzymes for example, display a wide variation of sequence specific binding affinities characterizing the strength of their interaction with different *cis*-regulatory elements that control the transcriptional mechanism[32,27]. A need for quantitative models for predicting the strength of protein-DNA interactions arise from this variation of binding affinities displayed by different sites to a given transcription factor. There is evidence to suggest that the binding energy of a protein-DNA interaction is to some extent intrinsic in the base composition of the operator DNA. Berg and von Hippel[2] for example showed, using statistical-mechanical theory, that given a set of regulatory sites, the logarithm of the base frequencies of those sites were proportional to their binding affinity. A more refined version of their calculation, taking into account the base composition of the genome in question, was introduced by Stormo & Fields[27] and Sarai & Takeda[22] showed that the binding energy of a site was additive in the free energy changes of individual bases. The binding affinity of a protein-DNA interaction can be measured experimentally as an equilibrium constant of its binding reaction[25,21].

Studies carried out on various domains of TF binding sites have shown correlated nucleotide variations to exist between different nucleotide positions of those sites[15,30,34]. This has led many researchers to question the base independence assumption on which methods such as consensus sequences and weight matrices for identifying TF binding sites are based[4,15]. This issue has been addressed by different authors with different techniques ranging from non-quantitative models such as improved weight matrices with prior information on correlated nucleotide positions[36,37], biophysical approaches[5], non-parametric models[11] and neural networks[14,10] to quantitative models such as principal coordinates analysis[31]. The method introduced in this paper, among other things, accounts for nucleotide co-variations to improve the quantitative prediction of binding affinities of protein-DNA interactions.

Another feature that plays a role in protein-DNA interactions is nucleotide structure[1,16,7]. It is reasonable to expect, given the multitude of binding sites recognized by the transcription machinery, that there exists other factors beside sequence similarity that influence the binding process. It has been shown that the binding of transcription factors cause a significant distortion to the regular twist and bending of the DNA double helix[9,23,26]. This often results in the bound DNA strand changing conformation from its B-form to A- or Z- forms[13]. It is conceivable that the binding affinity of a site to a given transcription factor will depend, at least to some extent, on its ability to tolerate such structural distortion from the classical B-form.

Nussinov[16], for example, demonstrated the presence of structural homology in regions with weak sequence homology at sites -10, -35 and -16, of the Escherichia coli promoter. Structural properties of a DNA helix can be expressed in terms of its conformational parameters in di-nucleotide and tri-nucleotide models. There are many different such parameters reported in the literature. The Property database[18], for example, lists 38 such parameters. Many non-quantitative algorithms have been developed to analyze binding sites based on their structural homology[19,17,29]. The method described in this paper uses a combination of different structural parameter representations of a site to account for sequence structure in the prediction of its binding affinity to a given transcription factor.

2. Method

The key to our method is the use of numerical *templates* to capture certain key features of TF binding sites. One of the principle drawbacks of base-independent models of TF binding sites is their inability to account for non-independent nucleotide variations. One problem of modelling nucleotide substitutions in a general model of TF binding sites is that the nucleotide positions that exhibit such correlations vary from factor to factor. As the exact positions on the TF binding sites which are correlated are unknown in the general case, one would need a model that accounts for all pairs of positions on the sites to fully represent them, which will need a very large number of parameters (e.g. a fully connected HMM). Templates present a compromise between the base independent model and the fully connected model. They model the correlation of an individual position relative to the rest of the positions on the site. By restricting the expression of correlation of a given position on the sites to all the other positions, instead of individual pairs of positions, templates are able to reduce the number of parameters required to the length of the sites, while still capturing a global representation of the positional correlations present in them.

In the template model, each template (defined by its template parameters $\mathbf{t}$) is modelled on a given numerical encoding of the nucleotides forming the training set of binding sites. The numerical encoding can be some value assigned to individual nucleotides or a value assigned to a combination of them. Values can be assigned to single nucleotides to capture sequence properties (e.g. sequence homology) of the sites. Values can be assigned to di- and tri- nucleotides to capture geometric and structural properties (e.g. propeller twist, stacking energy, protein induced deformability, DNAse I sensitivity, etc.) of the sites. For a given nucleotide sequence, s, and a given nucleotide parameter p, the resulting numerical vector will be denoted $\mathbf{r}^p(s)$ (see Figure 1.). Each nucleotide sequence is first converted into a table of numerical representations. For each nucleotide sequence s, the representations of s that we work with will be denoted $\mathbf{r}^1(s), \mathbf{r}^2(s), \ldots, \mathbf{r}^m(s)$ where m is the number of parameters selected.

$p =$ Slide

aa	at	ag	ac	ta	tt	tg	tc	ga	gt	gg	gc	ca	ct	cg	cc
0.1	-0.7	-0.3	-0.6	0.1	0.1	0.4	0.1	0.1	-0.6	-0.1	-0.3	0.4	-0.3	0.7	-0.1

$s =$ g g c g t g g c   ($\mathbf{r}^p(s) = -0.1, -0.3, +0.7, -0.6, +0.4, -0.1, -0.3$)

Figure 1. The figure shows the encoding $\mathbf{r}^p(s)$ of sequence s by the dinucleotide step parameter values $p =$'Slide'.

The global representation of positional correlations of a TF binding site s, encoded with parameter p (where p can be for example a mono-, di- or tri-nucleotide parameter), having encoded length L, is captured by a template t and is given by the following equations. As we do not know the contribution of individual bases towards the binding energy of a site, we make a simplifying assumption that the bases make a uniform contribution towards it, hence in these equations, the binding energy of a site is modelled as an external potential f equally distributed across its bases.

$$(\ \mathbf{Q}\ \mathrm{diag}(\mathbf{r}^p(s))\ )\ \mathbf{t} \ = \ \mathbf{r}^p(s) \ - \ \mathbf{f} \ - \ \mathbf{e} \tag{1}$$

Where $\mathbf{t} = (t[1], t[2], \ldots, t[L])^T$, is the vector of template parameters, $\mathbf{r}^p(s)$, is the vector representing the encoding of nucleotide sequence s with parameter p, $\mathbf{f} = (f, f, \ldots, f)^T_{(1\times L)}$, is an equally distributed vector of the binding affinity of site s, $\mathbf{e} = (e[1], e[2], \ldots, e[L])^T$, the residual error and $\mathbf{Q}_{(L\times L)}$ a square matrix with zeros on the diagonal and ones every where else.

For any numerical vector $\mathbf{r}^p = (r^p[1], r^p[2], \ldots, r^p[L])$, the *template error* of $\mathbf{r}^p$ with respect to a template $\mathbf{t}^p$, denoted as $E(\mathbf{r}^p, \mathbf{t}^p)$, is defined as the sum of squared residual errors.

$$E(\mathbf{r}^p, \mathbf{t}^p) = e[1]^2 + e[2]^2 + \ldots + e[L]^2 \tag{2}$$

Given a numerical vector $\mathbf{r}^p$, we can find a set of template parameters

$\mathbf{t}^p$ that minimises the template error $E(\mathbf{r}^p, \mathbf{t}^{\mathbf{P}})$ for that vector. This minimisation process is referred to as *'training the template'*. The template $\mathbf{t}^p$ that minimises $E(\mathbf{r}^p, \mathbf{t}^p)$ for the vector $\mathbf{r}^p$ is obtained as follows:

$$\begin{aligned} E(\mathbf{r}^p, \mathbf{t}^p) &= \arg\min_{\mathbf{t}^p} \left(e[1]^2 + e[2]^2 + \ldots + e[L]^2\right) \\ &= \arg\min_{\mathbf{t}^p} \left(\mathbf{e}^T \mathbf{e}\right) \\ &\textit{making the substitution } \mathbf{e} = (\mathbf{r}^p \; - \; \mathbf{f}) \; - \; (\; \mathbf{Q} \; \mathrm{diag}(\mathbf{r}^p) \;) \; \mathbf{t}^p \\ &= \arg\min_{\mathbf{t}} \left(\; (\; \mathbf{r}^p - \mathbf{f} - \mathbf{Q_r} \; \mathbf{t}^p \;)^T \; (\; \mathbf{r}^p - \mathbf{f} - \mathbf{Q_r} \; \mathbf{t}^p \;) \; \right) \end{aligned}$$

Where $\mathbf{Q_r} = (\; \mathbf{Q} \; \mathrm{diag}(\mathbf{r}^p) \;)$.

For any **set** of numerical vectors, $\{\mathbf{r}_1^p, \mathbf{r}_2^p, \ldots, \mathbf{r}_n^p\}$, the mean value of the template error with respect to a **fixed** template $\mathbf{t}^p$ is given by

$$\frac{1}{n} \sum_{k=1}^{n} E(\mathbf{r}_k^p, \mathbf{t}^p) \tag{3}$$

The template that minimises this mean error value for this set of vectors can be obtained by calculating the partial derivatives of Equation 3 with respect to $t^p[1], t^p[2], \ldots, t^p[L]$ and setting each of these equal to zero. This gives the following set of L linear equations:

$$\mathbf{t} \; = \; \left[\; \sum_{k=1}^{n} \mathbf{Q_r}_k^T \; \mathbf{Q_r}_k \; \right]^{-1} \; \left[\; \sum_{k=1}^{n} \mathbf{Q_r}_k^T \; (\mathbf{r}_k^p \; - \; \mathbf{f}_k) \; \right] \tag{4}$$

Where $\mathbf{Q_r}_k = \mathbf{Q} \; \mathrm{diag}(\mathbf{r}_k^p)$.

These equations are symmetric and can be solved efficiently to find the set of template parameters $t^p[1], t^p[2], \ldots, t^p[L]$ that minimises the mean template error for the set of vectors. These parameters represent the template that best describe the relationship between the encoded sites and their binding affinity. Templates are created for all the different parametric encodings p, of the sites in the training data. Given a template $\mathbf{t}^p$ and any site of the appropriate length $\mathbf{x}$, we can compute the projected binding affinity $\tilde{f}^p$ of that site from

$$\tilde{\mathbf{f}}^p = \mathbf{r}^{\mathbf{P}}(\mathbf{x}) \; - \; (\mathbf{Q} \; \mathrm{diag}(\mathbf{r}^{\mathbf{P}}(\mathbf{x}))) \; \mathbf{t}^p \tag{5}$$

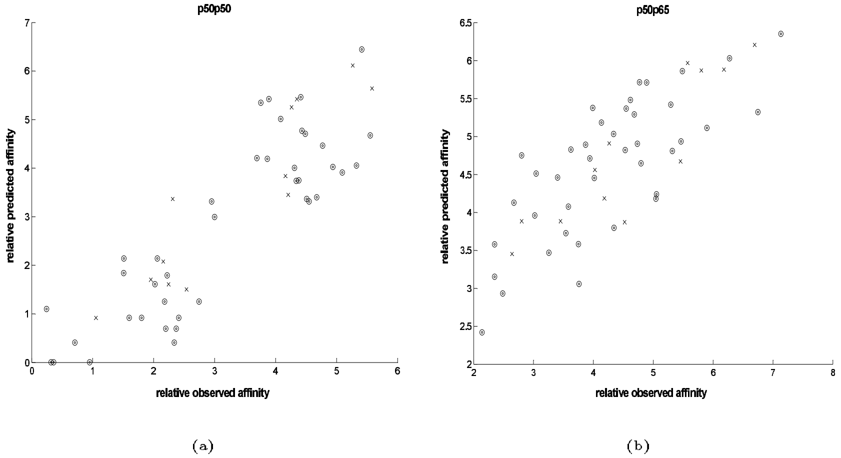
Where the projected binding affinity $\tilde{f}^p$ is taken as the mean value of the vector $\tilde{\mathbf{f}}^p$.

The predictive power of a set of templates depends on the specific nucleotide parameters the templates are modelled on. As mentioned before, for a given set of training data, we first create templates from all available nucleotide parameters. We then use a greedy approach for selecting the best subset of templates from this set for the final predictor. The predictor output will be the average over all selected templates. The templates are selected based on the degree of correlation displayed between the predicted and observed values of binding affinity of the training data. The correlation coefficient between the predicted and observed values of binding affinity is computed for each template representing a specific nucleotide parameter. Then, starting with the template with the highest correlation coefficient of prediction vs. observed, we add templates, one at a time in descending order of their correlation coefficient of prediction with the observed values to an expanding set of templates. This process is continued until the correlation coefficient of prediction of the combine set of templates drop as a result of the addition of a new template to the set. In that case we select all the templates excluding the last template added, as our set of templates.

3. Results

The di-nucleotide parameters, representing structural properties of the sequences, were obtained from the Property database[18]. In its current release this database lists *38* different parameter values. We used all *38* of these parameters. The tri-nucleotide parameters were obtained from Brukner et al.[3]

There isn't much published quantitative experimental data available on the binding affinities of different protein-DNA interactions. One study is reported in Udalova *et. al.*[31] which reports on the binding affinities of *NF-κB* binding sites. We tested the predictive capability of the templates for predicting the binding affinities of these sites. The authors list the binding affinities of *52* of the possible *256* variants of the *'GGRRNNYYCC'* *NF-κB* motif to the recombinant *p50p50* homodimer and *p50p65* heterodimer complexes. There were two estimations of the experimental binding affinities listed for each oligo varying in rang from *0* to *2431* normalised to the control sequence *'GGGGTTCCCC'* which was given the value *227*. We used the average of these two measurements as the observed binding affinities of the sites. Templates were modelled on the log binding affinities of the sites. Figure 2 (a) shows the predicted log binding affinities of the *p50p50* variant of the *NF-κB* sites plotted against the observed log binding affinities of those sites. The templates were trained on the first *12* sites listed in Udalova *et. al.*[31] (the data had no particular arrangement so can be considered random). The correlation coefficient of the test data (i.e. observed



(a)

(b)

Figure 2. The predicted binding affinity (y axis) plotted against the observed values (x axis) of the (a) *p50p50* binding sites (b) *p50p65* binding sites. These predictions are based on templates modelled using *12* sites (the sites shown with crosses).

vs predicted binding affinities of sites *13* to *52*) was *0.8104* $(p = 10^{-9})$. The templates selected by the predictor in obtaining the above results were the two twist parameters (P0000018, P0000026), the two roll parameters (P0000014, P0000028), the propeller twist parameter (P0000030) and the tilt parameter (P0000016). The values listed in brackets refer to the Property database[18] ID of these parameters.

Figure 2 (b) shows the predicted log binding affinities of the *p50p65* variant of the *NF-κB* sites plotted against the observed log binding affinities of those sites. As with the *p50p50* sites, the templates were trained on the first *12* sites listed in Udalova *et. al.*[31]. For this variant, the correlation coefficient of the test data (i.e. sites *13* to *52*) was *0.7547* $(p = 10^{-6})$. The templates selected by the predictor in obtaining the above results were the two twist parameters (P0000018, P0000026), the two roll parameters (P0000014, P0000028), and the propeller twist parameters (P0000030). It is interesting that the predictor selected templates modelled on similar parameters in both cases except for the tilt parameter in the prior case (i.e. for the *p50p50* sites). It could be that the di-nucleotide step values of roll and twist express the flexibility of the DNA strand which facilitates its orientation when binding to proteins. We believe, as observed from these results, that this flexibility of the DNA strand, to orient itself around the binding protein, in tern may have a direct correlation to the binding affinity of the protein-DNA interaction. It is also interesting to note that templates modelled on sequence were not selected by the predictor.

To further ascertain the robustness of the templates to predict the binding affinity of a given site, we performed a recursive randomised prediction experiment for both the *p50p50* and *p50p65* variants of the *52 NF-κB* binding sites. We forecasted the binding affinities of the *52* sites for both these variants *100* times with each time seeing its prediction based on a new template created from a set of *12* randomly selected sites and tested on the remaining sites. The results of these experiments are listed in Table 1. Also shown in Table 1 are the results of the above experiment carried out with an increase from *12* to *21* of the number of sites used to train the templates. The results produced by the *template* method described here are compared with three other methods. The first is the matrix similarity score computed as described in Quandt *et. al.*[20]. The second is the logarithm of the base frequencies as described by Berg and von Hippel[2] and the third is the non-parametric model described by King *et. al.*[11](with default parameters). These tests were done exactly as the tests carried out for the templates where predictions were performed *100* times with a new frequency matrix created from the training sites (both for *12* and *21* sites) selected randomly from the *52* sites and tested on the remaining sites, for each test. These results are also listed in Table 1. The standard deviation of each experiment is given in brackets.

Table 1. Performance statistics of templates for the *NF-κB* sites with training sets of *12* and *21* examples. The mean values are taken over *100* randomised trials. The standard deviations are given in brackets.

Mean correlation coefficient (Predicted vs. Observed)		
	$p50p50$	
Number of training examples	*12*	*21*
Templates	0.7809 (0.0469)	0.8124 (0.0370)
Matrix similarity score[20]	0.3535 (0.3167)	0.4559 (0.2851)
Logarithm of base frequencies[2]	0.2033 (0.3087)	0.2880 (0.2781)
Non-parametric model[11]	0.2686 (0.1691)	0.2758 (0.1727)
	$p50p65$	
Number of training examples	*12*	*21*
Templates	0.6444 (0.0869)	0.6941 (0.0883)
Matrix similarity score[20]	0.1896 (0.2732)	0.3338 (0.2121)
Logarithm of base frequencies[2]	0.1302 (0.2583)	0.1719 (0.2397)
Non-parametric model[11]	0.2413 (0.1589)	0.2414 (0.1601)

4. Discussion

We have described a novel approach for predicting the binding affinity of protein-DNA interactions. The approach described is based on templates that are sensitive to positional co-variations. These can be co-variations expressing sequence or structural polymorphisms as described by the different parametric encodings of the nucleotide sequence. Templates work in sets, usually containing more than one element, with each template characterising a different sequence or structural property of the sites. The amalgamation of different templates optimally selected to work in unison endows a synergic effect on the predictive capabilities of the system.

The training phase of the system requires a subset of binding sites along with their experimentally verified binding affinities. One advantage of the method described above, unlike other approaches such as, for example, those based on base frequencies which are susceptible to small-sample uncertainties[2], is its ability to learn quite well from a minimal number of training data. This is a feature that has many practical advantages when we a dealing with a dearth of properly annotated examples.

Binding assays of transcription factors such as *NF-κB*, *Zif268* zinc fingers and *Mnt* repressor-operator proteins suggest strong evidence to the existence of non-independent effects on positional interactions when at least some proteins bind to DNA[30,15,4]. The exact positions that exhibit such interdependent effects vary from one factor to another, and there is no evidence that all transcription factors exhibit a similar pattern of behaviour. This makes it difficult to capture such properties in a general model. The requirement is for models that can learn such variations from a set of training data.

The sensitivity of templates described above to positional co-variations is not based on any prior knowledge of which positions exhibit correlated behaviour. This is an important characterisation, especially in the absence of such prior knowledge individualising a family of binding sites, which is usually the case. It is not always practical to build exhaustive models detailing the different co-variations present between individual positions. Models such as neural networks and HMMs that are able to account for such information suffer from the practical drawback of balancing between the complexity of the systems and the number of examples required to train them well. In these systems, the complexity of the model architecture imposes lower bounds on the number of examples required to form a good training set. These bounds usually increase exponentially with the increase in complexity of the system.

There is evidence[1,16,6] that suggests the presence of structural homologies in DNA sequences that interact with some transcription factors. What these structural homologies are and exactly what geometric features play a part in them is not always very clear or easy to ascertain. Programs that incorporate such features do so with an implicit assumption of the presence of these properties in the sequences that they analyse. This is a weak assumption that may be tentative in the absence of specific knowledge of their presence and would not hold for the general case. It is possible for different binding sites to exhibit different structural properties intrinsic to the particular factor that they bind to. It is also possible for some binding sites not to display any significant structural homology for any of the known structural parameters. In such cases, one has only got sequence homology to rely on.

Templates used here for predicting binding affinity can model both sequence and structural homology. The important fact when modelling templates for a particular family of TF binding sites is that we do not make any prior decision on which structural parameters to use. The selection of the best set of parameters is done automatically during the training phase of the system, though in a greedy fashion.

References

1. T. Aoyama and M. Takanami. Essential structure of E. coli promoter II. Effect of the sequences around the RNA start point on promoter function. *Nucleic Acids Res.*, 13 (11):4085–4096, 1985.
2. O. G Berg and P. H von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–750, Feb 20 1987.
3. I. Brukner, R. Sanchez, D. Suck, and S. Pongor. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO Journal*, 14:1812–1818, 1995.
4. 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.
5. M Djordjevic, A. M Sengupta, and B. I Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Res.*, 13(11):2381–90, Nov 2003.
6. M. A. El Hassan and C. R. Calladine. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *Journal Molecular Biology*, 259(1):95–103, 1996.
7. M. A. El Hassan and C. R. Calladine. Two distinct modes of protein-induced bending in DNA. *Journal Molecular Biology*, 282(2):331–343, 1998.

8. 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.
9. T Gustafson, A Taylor, and L Kedes. DNA bending is induced by a transcription factor that interacts with the human c-FOS and alpha-actin promoters. *Proc Natl Acad Sci U S A*, 86(7):2162–6, 1989.
10. P. B. Horton and M. Kanehisa. An assessment of neural network and statistical approaches for prediction of E. coli promoter sites. *Nucleic Acids Research*, 20:4331–4338, 1992.
11. O. D King and F. P Roth. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, 31(19):e116, Oct 2003.
12. 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.
13. S. Lisser and H. Margalit. Determination of common structural features in Escherichia coli promoters by computer analysis. *Eur J Biochem.*, 223(3):823–830, 1994.
14. I. Mahadevan and I. Ghosh. Analysis of E.coli promoter structures using neural networks. *Nucleic Acids Res.*, 22 (11):2158–2165, 1994.
15. 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.
16. R. Nussinov. Promoter helical structure variation at the Escherichia coli polymerase interaction sites. *Journal of Biological Chemistry.*, 259:6798–6805, 1984.
17. U. Ohler, H. Niemann, G. C. Liao, and G. M. Rubin. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, 17:199–206., 2001.
18. 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.
19. M. P. Ponomarenko, J. V. Ponomarenko, A. E. Kel, and N. A. Kolchanov. Search for DNA conformational features for functional sites. Investigation of the TATA box. . *In: Biocomputing: proceedings of the 1997 Pacific Symposium. (Altman, R., et al., eds.), Word Sci. Publ., Singapore*, pages 340–351., 1997.
20. 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 Res.*, 23:4878–4884, 1995.
21. E. Ragnhildstveit, A. Fjose, P. B. Becker, and J. P. Quivy. Solid phase technology improves coupled gel shift/footprinting analysis. *Nucleic Acids Research*, 25(2):453–454, 1997.
22. A Sarai and Y Takeda. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc Natl Acad Sci U S*

A, 86(17):6513–6517, Sep 1989.

23. R Schreck, H Zorbas, E. L Winnacker, and P. A Baeuerle. The NF-kappa B transcription factor induces DNA bending which is modulated by its 65-kD subunit. *Nucleic Acids Res.*, 18(22):6497–502, 1990.
24. 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.
25. S. E. Shadle, D. F. Allen, H. Guo, W. K. Pogozelski, J. S. Bashkin, and T. D. Tullius. Quantitative analysis of electrophoresis data: novel curve fitting methodology and its application to the determination of a protein-DNA binding constant. *Nucleic Acids Research*, 25(4):850–860, 1997.
26. V. Y Stefanovsky, D. P Bazett-Jones, G Pelletier, and T Moss. The DNA supercoiling architecture induced by the transcription factor xUBF requires three of its five HMG-boxes. *Nucleic Acids Res.*, 24(16):3208–15, 1996.
27. G. D. Stormo and D. S. Fields. Specificity, energy and information in DNA-protein interactions. *Trends Biochemical Sciences*, 23:109–113, 1998.
28. 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, 1982.
29. 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.
30. 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.
31. 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.
32. I. A Udalova, A Richardson, A Denys, C Smith, H Ackerman, B Foxwell, and D Kwiatkowski. Functional consequences of a polymorphism affecting NF-kappa B p50-p50 binding to the TNF promoter region. *Molecular and Cellular Biology*, 20(24):9113–9119, Dec 2000.
33. W. Wasserman and A Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet.*, 5(4):276–87, Apr 2004.
34. 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.
35. X Xie, J Lu, E. J Kulbokas, T. R Golub, V Mootha, K Lindblad-Toh, E. S Lander, and M Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–45, Mar 2005.
36. Q. M. Zhang and T. G. Marr. A weight array method for splicing signal analysis. *Computer Applications in Biosciences*, 9(5):499–509, 1993.
37. Q Zhou and J. S Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–16, Apr 2004.