BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 10 2005, pages 2254–2263
doi:10.1093/bioinformatics/bti361

Sequence analysis

Non-additivity in protein–DNA binding

R. A. O'Flanagan[1], G. Paillard[2], R. Lavery[2] and A. M. Sengupta[1,*]

[1]Department of Physics and Astronomy and BioMaps Institute, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854-8019, USA and [2]Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris 75005, France

Received on September 29, 2004; revised on February 9, 2005; accepted on February 26, 2005
Advance Access publication March 3, 2005

ABSTRACT

Motivation: Localizing protein binding sites within genomic DNA is of considerable importance, but remains difficult for protein families, such as transcription factors, which have loosely defined target sequences. It is generally assumed that protein affinity for DNA involves additive contributions from successive nucleotide pairs within the target sequence. This is not necessarily true, and non-additive effects have already been experimentally demonstrated in a small number of cases. The principal origin of non-additivity involves the so-called indirect component of protein–DNA recognition which is related to the sequence dependence of DNA deformation induced during complex formation. Non-additive effects are difficult to study because they require the identification of many more binding sequences than are normally necessary for describing additive specificity (typically via the construction of weight matrices).

Results: In the present work we will use theoretically estimated binding energies as a basis for overcoming this problem. Our approach enables us to study the full combinatorial set of sequences for a variety of DNA-binding proteins, make a detailed analysis of non-additive effects and exploit this information to improve binding site predictions using either weight matrices or support vector machines. The results underline the fact that, even in the presence of significant deformation, non-additive effects may involve only a limited number of dinucleotide steps. This information helps to reduce the number of binding sites which need to be identified for successful predictions and to avoid problems of over-fitting.

Availability: The SVM software is available upon request from the authors.

Contact: anirvans@physics.rutgers.edu

1 INTRODUCTION

Identifying protein binding sites within DNA sequences remains a major goal of genomic annotation. In the case of transcription factors, identification of their binding sites is a vital step towards an understanding of transcription regulation networks. Unfortunately, transcription factors often have multiple roles and bind to many different sequences, making it difficult to describe their binding preferences with a simple consensus sequence or even with a degenerate consensus (Stormo, 2000). Weight matrices, whose entries reflect the observed base frequencies at each nucleotide position for a given set of binding sites [also termed position weight matrices (PWM) or position specific score matrices (PSSM)], are able to partially overcome such limitations, but they still suffer from the assumption that the overall binding affinity of a given protein is made up of additive contributions from interactions at each nucleotide position within the binding site. The same limitation applies to simple hidden Markov models (HMM), which represent a possible alternative to weight matrix approaches (Stormo *et al*., 1982). This assumption need not be valid and indeed a number of studies of specific proteins, notably involving the Mnt repressor (Man and Stormo, 2001), and the wild-type and variants of the EGR1 Zn-finger protein (Bulyk *et al*., 2001), have shown that correlations do exist between neighboring nucleotide positions.

It is possible to take into account such correlations by various extensions of the methods described above, e.g. by replacing mononucleotide PWM representations with those based on dinucleotides or longer sequence elements, or alternatively, by adding hidden layers to HMM formulations (Benos *et al*., 2002; Stormo, 2000). Support vector machines (SVMs) also offer an interesting route toward a more generalized coding of binding site information (Djordjevic *et al*., 2003). However, the main hindrance to work in this area remains the lack of experimental data. For most transcription factors, only a few binding sites have been experimentally characterized (Wingender *et al*., 2001). Although this situation is likely to change with the development of new high-throughput techniques, such as DNA microarrays (Bulyk *et al*., 1999, 2001), genomic SELEX (Gold *et al*., 1997), the so-called chip–chip approach (microarray-based chromatin immunoprecipitation assays) (Ren *et al*., 2000) or SELEX SAGE (Roulet *et al*., 2002), we are not yet in a position to make a comprehensive analysis of correlation effects. This is particularly true if we wish to analyze correlation within a complete binding site (typically containing 10–20 nt positions), rather than limiting the study to two or three adjacent nucleotides, as in the pioneering experimental studies of Mnt and EGR1 cited above. We would like to get some idea about the minimal number of known sites required for obtaining good classifiers of binding sites. It is easy to convince oneself that the existing datasets are too small for training reliable classifiers with dinucleotide correlations built in. However, that exercise is not enough to suggest how many more sites are needed, a crucial piece of information for further experimental endeavor.

We propose to overcome this difficulty by using a theoretical approach to obtain the necessary binding site data. Our approach uses a recently developed methodology for analyzing protein–DNA recognition mechanisms termed ADAPT (Lafontaine and Lavery,

*To whom correspondence should be addressed.

2000a,b; Paillard and Lavery, 2004). ADAPT allows the principal sequence-dependent components of the protein–DNA complexation energy to be calculated sufficiently fast that it becomes possible to scan the full combinatorial set of potential binding sequences for a given protein. Since protein binding sites typically range from 10 to 20 nt positions, this implies studying 4^{10}–4^{20} (i.e. $\sim 10^5$–10^{11}) sequences. Two of the present authors have already shown that by extracting the set of sequences corresponding to the most stable complexes (typically those within 5 kcal/mol of the best sequence), we can generate a simple weight matrix that can be compared with experimental results. Results published recently on 18 different proteins (Paillard and Lavery, 2004), suggest that ADAPT yields results which are very close to experimental consensus sequences.

The energy calculations performed by ADAPT take into account the two terms within the binding free energy which are likely to be the most sequence dependent, namely, the protein–DNA interaction energy (E_{int}) and the DNA deformation energy (E_{def}), i.e. the energy necessary to deform a free DNA segment to the structure it adopts when bound to the protein. It is possible to equate these two energy terms to the so-called direct and indirect components of protein–DNA recognition: the protein–DNA interaction energy accounts for 'direct' recognition due to the formation of specific hydrogen bonds, steric contacts or other interactions at the interface between the two macromolecules, while the DNA deformation energy accounts for the 'indirect' recognition linked to the ease with which the protein can induce the bound conformation of the double helix.

Until recently, direct recognition was thought to dominate protein binding specificity, except in cases where binding induced severe DNA deformation, a good example being the TATA-box binding protein (TBP), which must open up the minor groove and bend the double helix away from the approaching protein in order to establish a large binding interface (Kim *et al*., 1993; Nikolov *et al*., 1996). It has been shown experimentally that prebending DNA could enhance TBP binding (Starr *et al*., 1995) and also that binding was related to the flexibility of the targeted sequences (Singer *et al*., 1990). The results obtained with ADAPT confirm the importance of the indirect term (E_{def}) for TBP, and also suggest that this term plays a significant role in determining the specificity of almost all the protein complexes studied (Paillard and Lavery, 2004).

This surprising result is particularly interesting in the light of analyzing non-additive effects in protein binding since these effects are expected to be linked to protein-induced DNA deformation and to reflect changes in the interactions between neighboring nucleotide pairs (base stacking energies, in particular) following complexation. Since the formulation of ADAPT ignores sequence-dependent conformational changes in unbound DNA (by using a single, sequence-averaged B-DNA reference conformation) and also uses a pairwise additive force field for its energy calculations, correlations within these theoretical results can indeed arise only from the DNA deformation term. It should be added, however, that correlation can only arise for nucleotide positions where there is some degeneracy in the sequence preference, therefore significant DNA deformation is a necessary but an insufficient condition for correlation to occur.

In the present study, we look at the correlation effects on the binding specificity of some prototypical protein–DNA complexes and also investigate how effectively correlation can be incorporated into binding site prediction methods. The fact that ADAPT data agree well with the available experimental results for the complexes we study, encourages us to believe that although we are dealing with a theoretical model of binding affinity, the results should be close to those which will become accessible in the future using high-throughput experimental techniques.

2 METHODOLOGY

2.1 Calculating protein–DNA binding energies

Binding energies as a function of DNA base sequence were calculated using an all-atom representation of the complex derived from available high-resolution crystallographic data (see below). The protein–DNA interaction energy (E_{int}) and the DNA deformation energy (E_{def}), corresponding to the passage from a sequence-averaged B-DNA conformation to the bound conformation, were calculated using the AMBER parm98 force field (Cheatham *et al*., 1999). A distance-dependent dielectric function with a sigmoidal form was used to represent solvent damping of electrostatic interactions (Hingerty *et al*., 1985; Lavery *et al*., 1995). A detailed description of the ADAPT methodology can be found in Paillard and Lavery (2004).

Present tests were carried out on three protein complexes: the human TBP (Nikolov *et al*., 1996), the endonuclease BamH1 (Newman *et al*., 1995) and the bZIP protein GCN4 bound to its ATF/CREB site (Keller *et al*., 1995). These three complexes are henceforth referred to as TBP, BamH1 and GCN4, respectively. In each case, the binding energy was calculated for the full combinatorial set of base sequences within the DNA fragment employed. Those sequences falling within 5 kcal/mol of the energy of the optimal sequence were considered to constitute potential binding sites. Sequences with energies between 5 and 10 kcal/mol were considered to constitute a set of non-binding sites. For TBP, 881 binding sites and 6515 non-binding sites were selected using these criteria. Similarly, for BamH1 there were 368 binding sites and 1582 non-binding sites and, for GCN4 there were 476 binding sites and 4091 non-binding sites.

It may be noted that, it is possible to convert the number of sequences selected by the energy cutoff into an effective binding site length (expressed as a number of base pairs). This length, denoted by L_{tot}, is simply obtained as $L_{\text{tot}} = N - (\log M/\log 4)$, where N is the total length of the DNA fragment in the complex studied and M is the number of sequences with energies less than the cutoff energy. This expression can be derived by noting that N base pairs are associated with 4^N possible base sequences. Therefore, if M sequences fall below the cutoff, it is equivalent to B base pairs remaining undefined after protein binding (where $4^B = M$). Given $B = \log M/\log 4$, subtracting this number from N gives the effective length of the protein binding site (Paillard and Lavery, 2004).

2.2 Analyzing correlation between nucleotide positions

We will limit our analysis to the correlation between neighboring nucleotide positions within the protein binding site. Any such correlation will show up as the difference in the probability of given dinucleotide base combinations compared with the sum of the corresponding mononucleotide bases. This can be calculated as a change in entropy

$$\Delta S_{i,i+1} = S_{i,i+1} - (S_i + S_{i+1}),$$

where the corresponding mononucleotide and dinucleotide entropies are defined as:

$$S_i = -\sum_{\alpha} p_{i\alpha} \log_2(p_{i\alpha})$$

$$S_{i,i+1} = -\sum_{\alpha}\sum_{\beta} p_{i\alpha,(i+1)\beta} \log_2(p_{i\alpha,(i+1)\beta}).$$

Here $p_{i\alpha}$ is the probability of base α appearing at position i in the set of binding sequences selected using their binding energies computed with ADAPT and $p_{i\alpha,(i+1)\beta}$ is the probability of base α appearing at position i followed by base β in position $i+1$. Note that α and β are indices running over the four nucleic acid bases (A, C, G and T) and that, $0 \leq S_i \leq 2$ and $0 \leq S_{i,i+1} \leq 4$.

We also introduce two other overall measures quantifying the correlation, namely the binding site lengths, L_m and L_d. These are related to the length L_{tot} defined above, but are calculated, by assuming that there is no correlation between neighboring sites or that only nearest-neighbor correlation exists.

$$L_m = l + \sum_{i=1}^{l} \sum_{\alpha} p_{i\alpha} \log_4(p_{i\alpha})$$

$$L_d = l + \sum_{\alpha} p_{1\alpha} \log_4(p_{1\alpha}) + \sum_{i=1}^{l-1} \sum_{\alpha\beta} p_{i\alpha,(i+1)\beta} \log_4 \left(\frac{p_{i\alpha,(i+1)\beta}}{p_{i\alpha}} \right).$$

Note that, L_{tot} measures the binding site length taking into account correlations between any nucleotides within the target oligomer. By definition $L_m \leq L_d \leq L_{tot}$. These three lengths yield a quantitative measure of correlation effects.

2.3 Extracting weight matrix parameters from binding site data

A popular method of characterizing binding motifs is the information theoretic weight matrix. Given a set of sequences to which a given protein binds, a simple mononucleotide weight matrix can be constructed and used to score a chosen sequence σ as follows (Stormo, 2000):

$$W_m(\sigma) = -\sum_{i=1}^{l} \sum_{\alpha} \omega_{i\alpha}\sigma_{i\alpha} - C_0,$$

where $W_m(\sigma)$ is the negative of the 'information score' of the sequence σ (of length l) and $\sigma_{i\alpha}$ is 1 if the i-th base is of type α and 0 otherwise. The $\omega_{i\alpha}$ are given by $\log_2(p_{i\alpha}/p_\alpha)$, where $p_{i\alpha}$ is the probability of base α appearing at position i in the set of binding sequences and p_α is the background probability of finding the base α. A constant shift, C_0, is chosen so that the best binding site scores zero (poorer sites having positive scores).

The information theoretic weight matrix has been formulated as a maximum-likelihood estimation of parameters, by assuming that the probability of binding to a sequence is proportional to the exponential of the information score. In our study, the training set sequences are sampled from those with a binding free energy below a cutoff. This distribution need not be well approximated by the aforementioned exponential distribution. In fact the maximum-likelihood method for distributions with sharp cutoffs has been described in Djordjevic *et al.* (2003). This method is an SVM. A mononucleotide SVM can be used to determine the binding energy of the protein to the sequence σ, as:

$$E_m(\sigma) = \varepsilon \cdot \sigma = \sum_{i=1}^{l} \sum_{\alpha} \varepsilon_{i\alpha}\sigma_{i\alpha},$$

where $\varepsilon_{i\alpha}$ is the free energy contribution from the i-th base. The parameters $\varepsilon_{i\alpha}$ are chosen to minimize the variance of $\varepsilon \cdot \sigma$ over the background distribution of sequences, subject to the constraints $\varepsilon \cdot \sigma^{(j)} \leq -1$ for the set of binding sequences $\sigma^{(j)}$, $j = 1, \ldots, N$. Sequences satisfying $\varepsilon \cdot \sigma \leq -1$ are then declared to be binding sites, although for the purposes of comparison with the weight matrix method, one can consider a more general threshold $\varepsilon \cdot \sigma \leq \mu$. In practice, the distribution of free energies is taken to be Gaussian and the quantity to be minimized is given by,

$$\chi^2 \equiv \sum_{i=1}^{l} \sum_{\alpha} p_\alpha \varepsilon_{i\alpha}^2$$

subject to the constraints, $\sum_\alpha p_\alpha \varepsilon_{i\alpha} = 0$, for each i.

Generalization of the weight matrix and SVM approaches to include dinucleotide terms yields the following two expressions:

$$Z_d(\sigma) = \sum_{i=1}^{l} \sum_{\alpha} \omega_{i\alpha}\sigma_{i\alpha} - \sum_{i=1}^{l-1} \sum_{\alpha\beta} \omega_{i\alpha\beta}\sigma_{i\alpha}\sigma_{i+1,\beta} - C_0$$

$$E_d(\sigma) = \sum_{i=1}^{l} \sum_{\alpha} \varepsilon_{i\alpha}\sigma_{i\alpha} - \sum_{i=1}^{l-1} \sum_{\alpha\beta} J_{i\alpha\beta}\sigma_{i\alpha}\sigma_{i+1,\beta}$$

where $\omega_{i\alpha\beta}$ is $\log_2(p_{i\alpha(i+1)\beta}/(p_{i\alpha}\, p_{(i+1)\beta}))$ and $p_{i\alpha,(i+1)\beta}$ is the probability of base α appearing at position i followed by base β in position $i+1$. The energies $\varepsilon_{i\alpha}$ and the $J_{i\alpha\beta}$ are chosen to minimize the variance:

$$\chi^2 = \sum_{i=1}^{l} \sum_{\alpha} p_\alpha \varepsilon_{i\alpha}^2 + \sum_{i=1}^{l-1} \sum_{\alpha\beta} p_\alpha p_\beta J_{i\alpha\beta}^2$$

subject to the constraints, $\sum_\alpha p_\alpha \varepsilon_{i\alpha} = 0, \sum_\alpha p_\alpha J_{i\alpha\beta} = 0$ and $\sum_\beta p_\beta \times J_{i\alpha\beta} = 0$ for each i.

3 RESULTS AND DISCUSSION

3.1 Evidence for non-additivity in binding

We begin by asking whether the binding energies calculated with ADAPT show significant non-additivity effects. The first test was performed for TBP which, given the strong DNA deformation induced by the protein (Nikolov *et al.*, 1996), is a likely candidate for correlations to be observed between adjacent nucleotide pair positions. After calculating the binding energy E_{tot} ($= E_{int} + E_{def}$) for the full combinatorial set of sequences within a 12 nt pair fragment bound to the protein, we extracted the sequences lying within a 5 kcal/mol interval of the optimal sequence (880 cases) and constructed a mononucleotide weight matrix, W_m. This matrix, the corresponding sequence logo (Schneider and Stephens, 1990) and the experimental sequence logo, based on the binding sites listed in the TRANSFAC data (Wingender *et al.*, 2001), are shown in Figure 1. We then used the W_m matrix to score all possible sequences within our TBP complex, that is to say the 4^{12} possible sequences that can fit in the 12 bp DNA target used here. As an example, using the data in Figure 1A and the first equation in Section 2.3, the sequence AGTATAATTAAA gives an initial sum of $-[\log_2(0.512/0.25) + \log_2(0.323/0.25) + \cdots] = -15.5$. Since this sequence has the most negative sum, C_0 in the equation is set to this value and the score of this sequence becomes zero. All other sequences consequently give positive scores.

The result is plotted against the calculated binding energies of these sequences in Figure 2. If the TBP complex was characterized by additive contributions to the binding energy, one would expect all sequences plotted in this figure to lie close to a single diagonal line. A diagonal, implying a perfect correlation between W_m scores and binding energies, would indicate that dinucleotide (or higher) dependencies are not required to explain the variations in total binding energies. In reality, we see that identical W_m scores correspond to multiple binding energies, leading to distinct clusters of sequences lying along vertically shifted diagonals. For well-separated clusters, it is possible to generate sequence logos (shown on the right-hand side of Fig. 2) which reveal that each group of sequences is associated with a particular combination of bases, mainly involving the four nucleotide positions starting from the last A of the TATA-box.

Figure 3A shows a more detailed view of the poor predictions made using the W_m score for TBP for the subset of 880 sequences lying below the energy cutoff, and defined in this study as good binding sites. We can now go further and isolate the origin of these effects by recalculating the ADAPT binding energies excluding all interactions between neighboring nucleotides within the term E_{def}. This leads to the plot shown in Figure 3B which is much closer to the single diagonal line expected. If we reintroduce energetic interactions only between neighboring nucleotides we recover almost exactly the same results as in Figure 3A (data not shown). This confirms that non-additivity arises from DNA deformation during binding

A.

pos	T	C	A	G
1	23.4	11.8	51.2	13.6
2	29.2	15.7	22.8	32.3
3	100.0	0.0	0.0	0.0
4	0.0	0.0	100.0	0.0
5	100.0	0.0	0.0	0.0
6	27.4	0.0	72.6	0.0
7	24.6	0.0	75.4	0.0
8	54.7	0.0	45.3	0.0
9	50.7	0.0	49.3	0.0
10	5.7	10.2	61.9	22.2
11	3.1	24.0	52.9	20.1
12	14.0	28.3	33.7	24.1

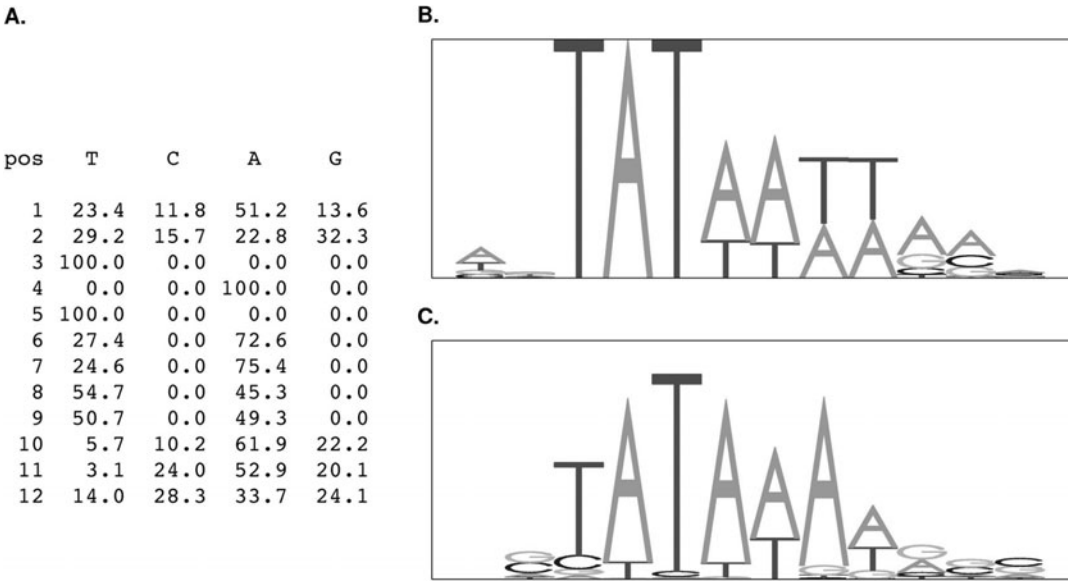

Fig. 1. (**A**) The W_m for human TBP (Nikolov *et al*., 1996) deduced from the binding sites with energies within 5 kcal/mol of the minimum, calculated using ADAPT; (**B**) the corresponding sequence logo; (**C**) the experimental sequence logo for human TBP from TRANSFAC (Wingender *et al*., 2001).

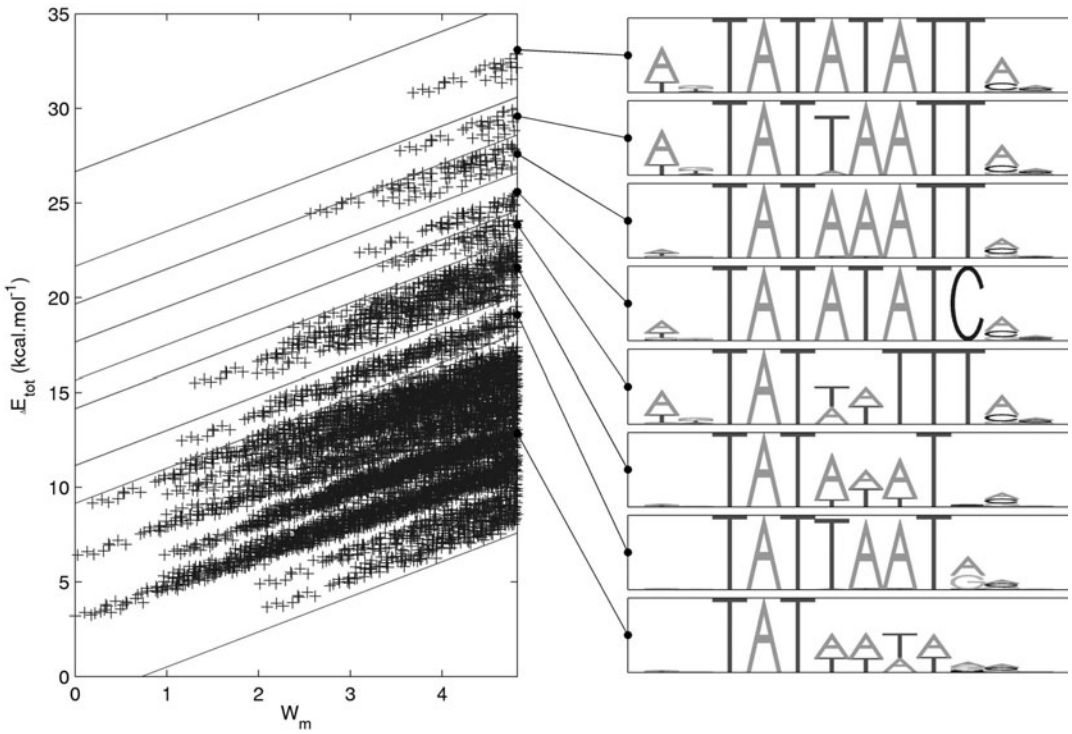

Fig. 2. TBP binding energies plotted against the corresponding W_m score. The sequence logos corresponding to well-defined clusters of sites are shown on the right-hand side of the figure.

and is dominated by the interactions between adjacent nucleotides. Figure 3C confirms this conclusion in another way by scoring the full ADAPT binding energies using the dinucleotide weight matrix W_d. The inclusion of nearest-neighbor effectively eliminates virtually all of the clustering seen in Figure 3A.

3.2 Analyzing non-additivity within the binding site

We can get an overall view of non-additivity for our three test proteins by calculating the binding sites lengths formulated in Section 2. The results, given in Table 1, confirm that TBP exhibits the strongest correlations, leading to an increase in the effective binding site length

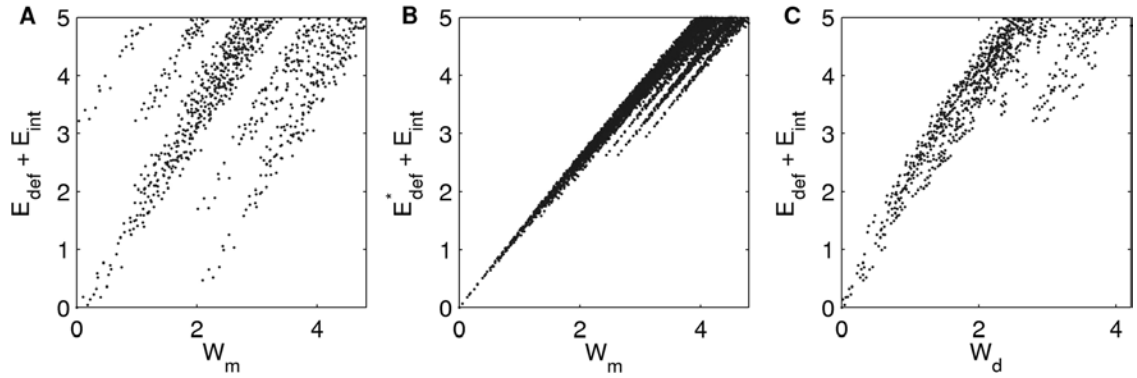

Fig. 3. (**A**) TBP binding energies, for sequences falling below the energy cutoff, plotted against the corresponding W_m score; (**B**) TBP binding energies ($E_{def} \rightarrow E_{def^*}$), excluding nearest-neighbor contributions, plotted against the corresponding W_m score; (**C**) TBP binding energies plotted against the corresponding W_d score.

Table 1. Binding site length under the hypothesis of no correlation between positions (L_m), correlation with nearest neighbors (L_d) or correlation with every position (L_{tot})

Length	TBP	BamH1	GCN4
L_m	5.8	7.2	7.1
L_d	6.7	7.5	7.2
L_{tot}	7.1	7.7	7.5

Table 2. Entropy differences between mononucleotide and dinucleotide analyses as a measure of non-additive effects along the binding sites of three test proteins

Dinucleotide	TBP	BamH1	GCN4
1–2	0.01	0.02	0.00
2–3	0.00	0.27	0.00
3–4	0.00	0.00	0.00
4–5	0.00	0.00	0.00
5–6	0.00	0.00	0.00
6–7	0.52	0.00	0.11
7–8	0.17	0.00	0.01
8–9	0.79	0.00	0.00
9–10	0.18	0.00	0.00
10–11	0.02	0.17	0.00
11–12	0.02	0.01	0.01

Shading indicates dinucleotide steps showing significant correlation ($\Delta S \geq 0.1$).

by 1.3 bp (i.e. $L_{tot} - L_m$). It can also be seen that, as shown above, virtually all of these effects (0.9 bp, i.e. $L_d - L_m$) are explained by nearest-neighbor interactions. Using these same measures, BamH1 exhibits a moderate increase of 0.5 bp, of which the majority (0.3 bp) is explained by nearest-neighbor interactions, while GCN4 shows the smallest increase (0.4 bp) of which only 0.1 bp can be attributed to nearest-neighbor interactions.

Using the nearest-neighbor entropy difference ΔS defined in Section 2, we can now ask exactly which dinucleotide steps within a given binding site exhibit significant correlations. The results for the three test proteins are given in Table 2. The first thing to note is that none of the test proteins show correlation all along the binding site. As expected, TBP shows the largest entropy changes, but even for this strongly distorted binding site, there are only significant correlation effects for the last four dinucleotide steps of the site (note that the TATA-box lies at positions 3–6). This is in line with the logos for the sequence clusters shown in Figure 2, which also indicate correlations involving the last A of the TATA-box and the three following nucleotide positions (steps 6–7, 7–8, 8–9 and 9–10).

For BamH1, the six bases of the GGATCC binding site (lying at positions 4–9) are all strongly selected and consequently cannot show any correlation. However, the impact of considerable DNA deformation can be seen within the flanking positions of the binding site, which have been shown experimentally to exhibit sequence selectivity (Engler *et al*., 2001). It is thus not surprising to find some correlation outside the binding site at junctions 2–3 and 10–11. Finally, for GCN4, the TGACGT binding site lies at positions 3–8. The only significant correlation occurs at steps 6–7, where the sequence logo shown in our earlier publication (Paillard and Lavery, 2004), confirms a weaker selectivity for C and G than for the other bases within the site.

3.3 Predicting binding sites taking non-additivity into account

We begin with the example of TBP. We have chosen 200 binding sequences randomly, out of the 880 sequences with energies within 5 kcal/mol of the minimum, to be used as inputs to the weight-matrix and SVM approaches. The resulting weight matrices and energy matrices were then used to assign information scores and predicted binding energies, respectively, to any candidate site. Sites with information scores or binding energies beyond a chosen threshold are 'predicted' to be binding sites by the weight matrix or SVM algorithm. Candidate sites with binding energies calculated by ADAPT <5 kcal/mol cutoff were considered to be 'true' binding sites, while those with energies above the cutoff were considered to be true non-binding sites. A false positive (FP) arises when one of the algorithms declares a true non-binding site to be a binding site,

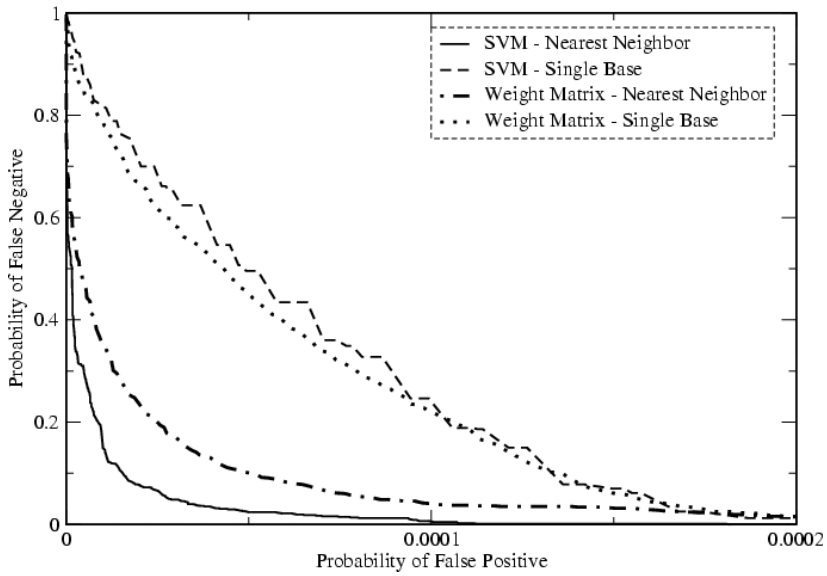

Fig. 4. Performance of mono- and dinucleotide weight matrix and SVM approaches for TBP using a training set of 200 sequences.

while a true positive (TP) arises when the algorithm correctly identifies a binding site as such. Similarly, a false negative (FN) arises when one of the algorithms declares a true binding site to be a non-binding site, while a true negative (TN) arises when the algorithm correctly identifies a true non-binding site. We estimate the rates of misclassification (either FP or FN), by using the classifier on a large number of sequences, which were not in the original training set (of size 200).

Figure 4 shows the curves representing the trade-off between the FN rate and the FP rate as the threshold varies, with or without the consideration of nearest-neighbor correlations (for all dinucleotide steps). The FP rates that are shown, indicate the fraction of the set of all ($\sim 4^{12}$) non-binding sites misclassified. The performance of either of the methods, weight matrix or SVM, clearly indicates the importance of taking nearest-neighbor non-additive effects into account for TBP. In addition, when this is done, it can be seen that the SVM approach performs considerably better, notably in terms of eliminating FPs. This is confirmed by the results in Table 3, where the weight matrix threshold was chosen to minimize the overall percentage of false attributions and the SVM threshold was the default value ($\mu = 1$). In Table 3, it may be noted that, we quantify the propensity to find TPs, rather than FPs, in terms of positive prediction value (TP/(TP+FP)). The common practice is to use the FP rate (FP/(FP+TN)). Unfortunately, in this particular context, FP rate tends to be very small for all reasonable methods (since the total number of predicted positives is a small fraction of TNs) producing a false impression of accuracy.

To some, Figure 4 might suggest that the single base model was good enough, with an FN rate of 0.1 and an FP rate of 0.0002 for some setting of the threshold. So, why bother about correlations? Unfortunately, an FP rate of 0.0002 leads to hundreds to thousands of false hits for the yeast genome depending on how much of the upstream regions are being searched. In the human genome, the number would be an order of a still larger magnitude. FN rate of 0.1 for TBP corresponds to hundreds of missed hits in yeast, with the extremely conservative estimate (Cliften *et al*., 2003) that only 15% of yeast genes have a TATA-box (based on perfect conservation in *sensu stricto Saccharomyces*). The actual number possibly runs from several hundreds to a thousand, once more. Thus, the smallness of the probabilities does not mean much, if not taken in the context of the total number of false predictions.

Table 3. Performance of the mononucleotide and dinucleotide versions of the weight-matrix and SVM approaches for TBP

approach	Positive prediction value	FN probability
W_m	0.69	0.86
SVM_m	0.26	0.09
W_d	0.79	0.38
SVM_d	0.82	0.16

The two plots in Figure 5 show the improvement in SVM performance for training sets ranging from 2 to 200 TBP binding sites. As the size of the training set increases, the dinucleotide model clearly outperforms the mononucleotide model. In fact, as the size of the training set increases the fraction of misclassified sequences (given by the sum of FPs and FNs divided by the size of the test set) increases quite sharply unless nearest-neighbor correlations are taken into account. This can be seen in Figure 6A and is explained by the fact that the mononucleotide model must adopt a lax threshold to ensure that all sites in the training set are correctly classified as good binding sites. Note that the error bars in this figure correspond to different training sets (of a fixed size) chosen from the full set of ADAPT binding sites.

We can now contrast this situation with the results for GCN4, which has been shown to have almost no nearest-neighbor correlations within the binding site. The results in Figure 7 for a training set of 200 sites now show little gain from the inclusion

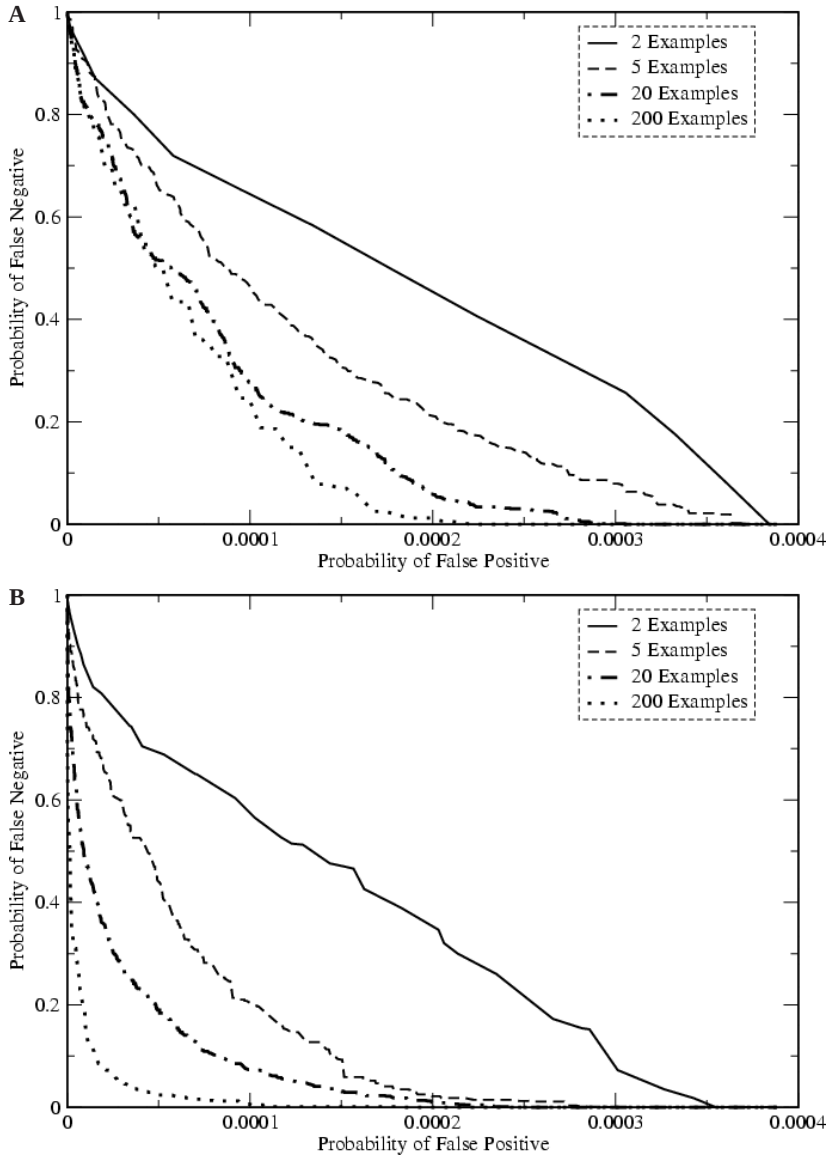

Fig. 5. Performance of the SVM approach for TBP as a function of training-set size: (**A**) mononucleotide model; (**B**) dinucleotide model.

of dinucleotide terms. This is confirmed for the fraction of misclassified sequences in Figure 6B as a function of training set size. Although the dinucleotide model outperforms the single-base approach at large training-set sizes, when fewer examples are available for training, the additional parameters in the model lead to over-fitting, with the result that the single-base model becomes significantly better. A minimum number of binding sites is therefore required before it becomes advantageous to introduce correlations. The contribution of dinucleotide terms to the overall binding energy (discussed in Section 3.2) is the most important factor determining the minimum number of sites required to justify the more complex model.

We can demonstrate this behavior clearly in the case of BamH1 which was shown to have only two dinucleotide steps with significant correlation effects (Table 2). Figure 6C contrasts the fraction of misclassified sites as a function of the size of the training set for three different models: mononucleotide (with no correlations), dinucleotide (with correlations at all dinucleotide steps) and partial dinucleotide (where correlations are only introduced where necessary, in the case of BamH1 only at junctions 2–3 and 10–11). The results show that the partial dinucleotide model outperforms both the full dinucleotide model and the mononucleotide model even at small training-set sizes.

3.4 Experimental signature of correlations from experiment

A comprehensive analysis of correlation effects using experimental data is currently impossible. Although pioneering studies have demonstrated the existence of correlation at chosen positions within protein binding sites (Bulyk *et al*., 2001; Man and Stormo, 2001), it is

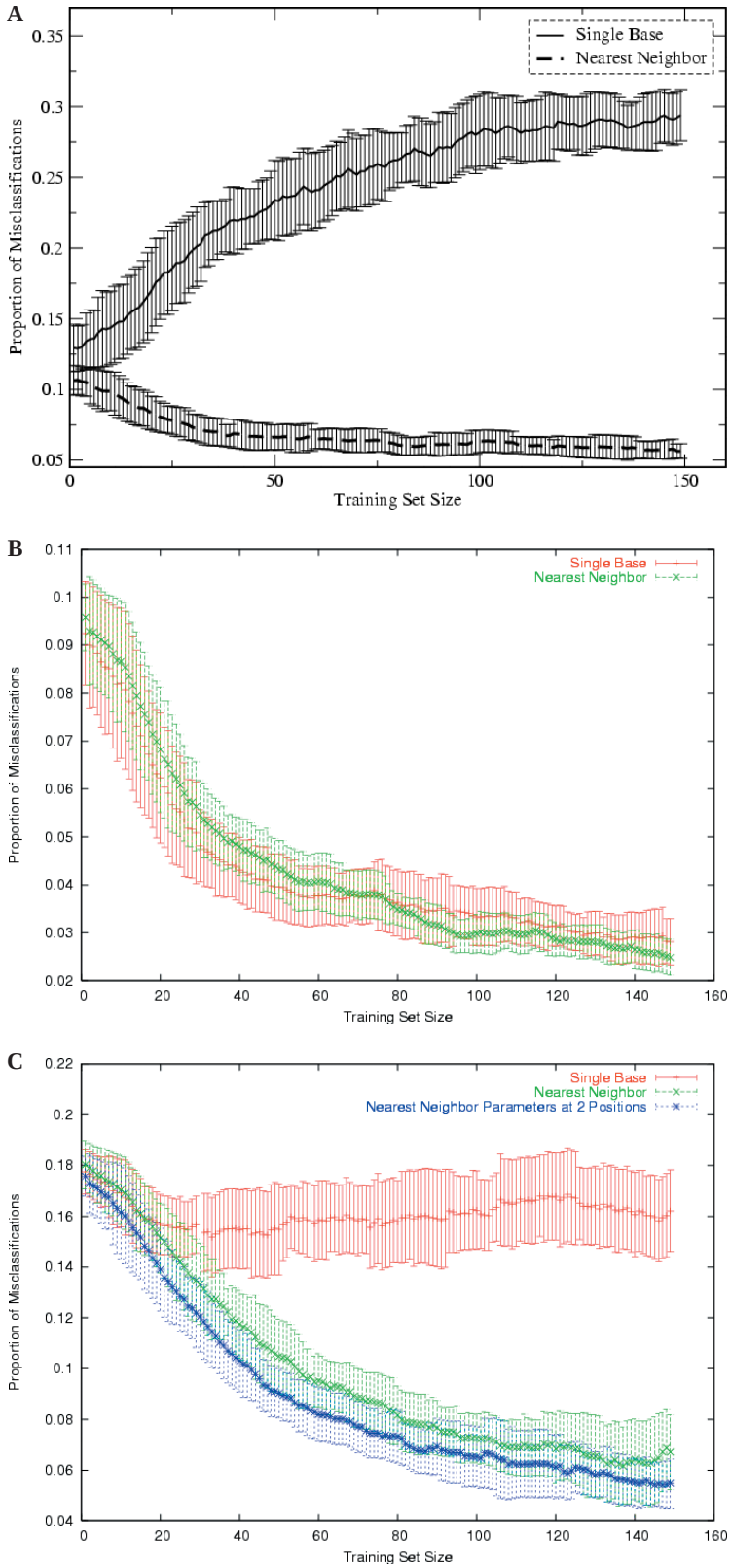

Fig. 6. Fraction of misclassified sites as a function of training-set size for the mononucleotide and dinucleotide SVM. The vertical error bars correspond to different choices of binding sites from the full set defined by the ADAPT calculations: (**A**) TBP, (**B**) GCN4 and (**C**) BamH1, including the results for a hybrid SVM treatment with dinucleotide parameters only at steps 2–3 and 10–11.

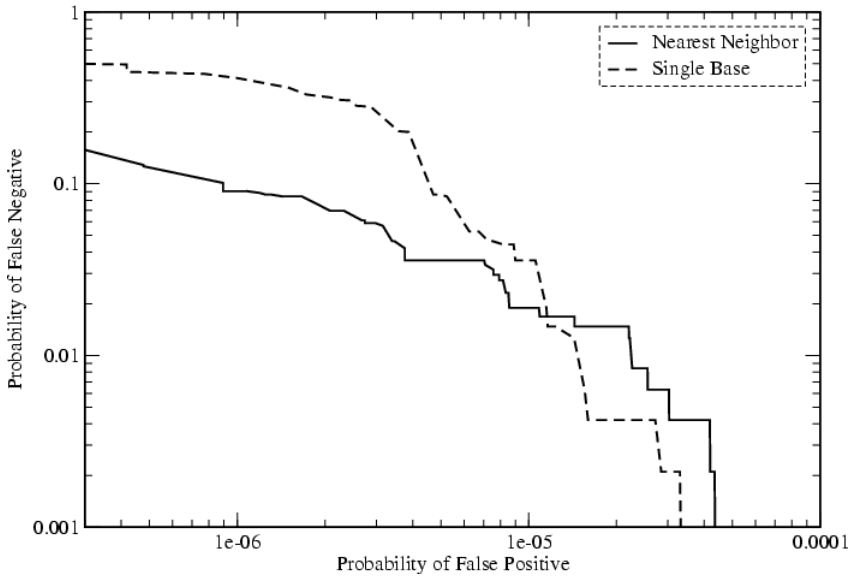

Fig. 7. Performance of mono and dinucleotide weight matrix and SVM approaches for GCN4 using a training set of 200 sequences (logarithmic plot).

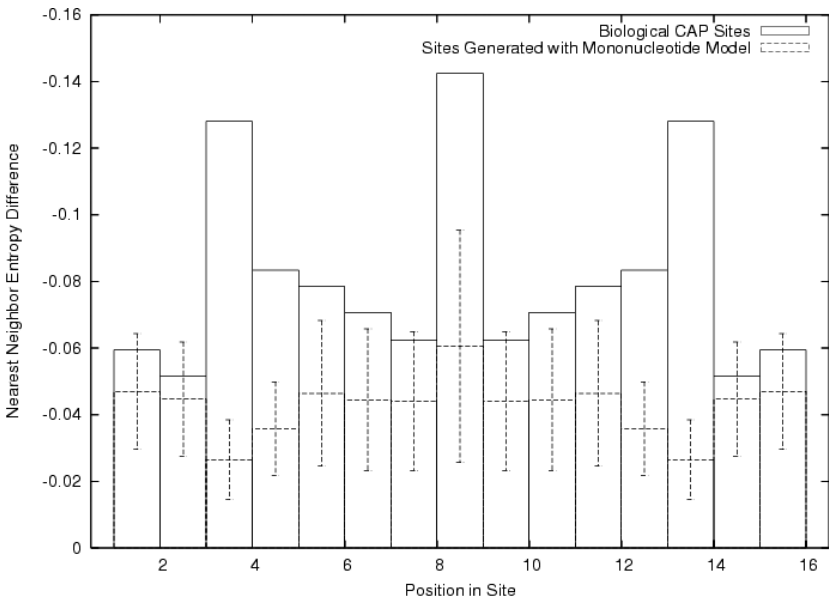

Fig. 8. Nearest-neighbor entropy differences for 76 experimentally confirmed sites and their reverse complements compared with sets of 76 artificial sites and their reverse complements, which were generated using a binding model in which mononucleotide terms account for the entire binding specificity.

difficult to find enough data to analyze the pattern of correlation within an entire site. This problem certainly applies to the three proteins we have studied above. However, it has been possible to make at least a preliminary analysis in the case of the dimeric protein CAP which binds to a 16 bp site. A careful literature study has led to the creation of a database of 76 confirmed binding sites for this protein (Thayer, 2004, http://linus.chem.wesleyan.edu/∼kthayer/capDB_index.htm).

These 76 sites and their reverse complements were used to construct a mononucleotide energy matrix using the procedure described in Section 2. This energy matrix was used to generate 10 000 sets of 76 theoretical binding sites, each of which would bind, at least as strongly as the weakest site in the set of 76 confirmed sites, if mononucleotide contributions to the binding energy were sufficient to describe the sequence specificity of the CAP protein. The need to incorporate dinucleotide terms into the binding model can then be established by detecting statistically significant nearest-neighbor effects in the experimental dataset which are not present in the artificially generated sites.

We can now analyze the nearest-neighbor entropy differences along the CAP binding site using the formulae given in the methodology section and either the experimental or the generated sites. For the experimental sites, nearest-neighbor entropy

differences were calculated for the set of 152 sequences consisting of the confirmed sites and their reverse complements. For the theoretical dataset, the mean and standard deviation of the nearest-neighbor entropy differences were calculated for the 10 000 sets, each consisting of 76 generated sites and their reverse complements. The results are shown in Figure 8. Significant nearest-neighbor effects can be seen at positions 3–4, 8–9 and 13–14, indicating that these are the locations at which dinucleotide terms introduced into the binding model would provide a better description of the binding specificity.

4 CONCLUSIONS

By using a theoretical approach to estimate protein–DNA binding energies, we have characterized a sufficient number of binding sites to allow an analysis of non-additive effects on binding specificity. The results confirm that DNA deformation within a protein complex can lead to significant non-additivity. These effects are shown to be almost exclusively limited to nearest-neighbor interactions. A more detailed analysis has also shown that, even in the case of significant deformation, non-additivity may only be important for a limited number of dinucleotide steps within the target site. It should be stressed that using theoretically estimated energies to obtain large enough datasets of binding and non-binding sites is clearly an approximation. The ADAPT approach only accounts for protein–DNA interaction energies and DNA deformation energies and is naturally limited by the precision of the force field employed. However, it should also be noted that ADAPT has successfully reproduced the experimental weight matrices determined for a wide variety of DNA-binding proteins. As concerns non-additivity effects, we have shown their origin is dominated by the interaction energy between neighboring base pairs, energies which are well estimated using the AMBER force field (as shown by numerous earlier modeling studies) and largely independent of the solvent, counterion and entropic effects which are excluded from our study.

Non-additivity can be taken into account within both weight matrix and SVM approaches to site prediction by the introduction of additional parameters to account for dinucleotide interactions. For the examples studied here, SVMs are shown to outperform weight-matrix techniques whether or not nearest-neighbor interactions are included. However, the improvement in predictive power is achieved only if sufficient data are available and, in this connection, it is important to take non-additivity into account only for those steps where it is really needed. Failure to do this can lead to over-fitting of dinucleotide models and consequently to poor predictive power. In general, the dinucleotide SVM with an insufficient training set results in many FNs, while its mononucleotide version will result in many FPs, even with a large training set. The present study suggests that, as a rule of thumb, the SVM approach performs well when the number of binding sites available for training is 1–1.5 times the number of parameters to be estimated. This implies that 135 sites are needed to develop a full dinucleotide description of the TBP binding site (compared with only 36 for a mononucleotide model). However, this number can be almost halved (72 sites) by only taking into account steps showing significant non-additivity.

As a corollary to the results presented here, when high-throughput techniques begin to provide much more extensive data on protein binding sites, it will become possible to detect the presence of non-additive effects both at the level of overall target sites and at the level of individual dinucleotide steps and, consequently, to deduce the existence and the distribution of significant DNA deformation.

ACKNOWLEDGEMENTS

The authors thank Dr K.M. Thayer for providing access to the CAP binding site database. G.P. and R.L. wish to thank the CNRS and the inter-organism Bioinformatics Program for contributing to the funding of this research.

REFERENCES

Benos,P.V. *et al*. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.

Bulyk,M.L. *et al*. (1999) Quantifying DNA–protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.*, **17**, 573–577.

Bulyk,M.L. *et al*. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.

Cheatham,T.E.III *et al*. (1999) A modified version of the Cornell *et al*. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.

Cliften,P. *et al*. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Djordjevic,M. *et al*. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.

Engler,L.E. *et al*. (2001) The energetics of the interaction of BamHI endonuclease with its recognition site GGATCC. *J. Mol. Biol.*, **307**, 619–636.

Gold,L. *et al*. (1997) From oligonucleotide shapes to genomic SELEX: novel biological regulatory loops. *Proc. Natl Acad. Sci. USA*, **94**, 59–64.

Hingerty,B. *et al*. (1985) Dielectric effects in biopolymers: the theory of ionic saturation revisited. *Biopolymers*, **24**, 427–439.

Keller,W. *et al*. (1995) Crystal structure of a bZIP/DNA complex at 2.2 A: determinants of DNA specific recognition. *J. Mol. Biol.*, **254**, 657–667.

Kim,Y. *et al*. (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature*, **365**, 512–520.

Lafontaine,I. and Lavery,R. (2000a) ADAPT: a molecular mechanics approach for studying the structural properties of long DNA sequences. *Biopolymers*, **56**, 292–310.

Lafontaine,I. and Lavery,R. (2000b) Optimization of nucleic acid sequences. *Biophys. J.*, **79**, 680–685.

Lavery,R. *et al*. (1995) JUMNA (junction minimization of nucleic-acids). *Comput. Phys. Commun.*, **91**, 135–158.

Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.

Newman,M. *et al*. (1995) Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science*, **269**, 656–663.

Nikolov,D.B. *et al*. (1996) Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl Acad. Sci. USA*, **93**, 4862–4867.

Paillard,G. and Lavery,R. (2004) Analyzing protein–DNA recognition mechanisms. *Structure (Camb.)*, **12**, 113–122.

Ren,B. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Roulet,E. *et al*. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **20**, 831–835.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Singer,V.L. *et al*. (1990) A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. *Genes Dev.*, **4**, 636–645.

Starr,D.B. *et al*. (1995) DNA bending is an important component of site-specific recognition by the TATA binding protein. *J. Mol. Biol.*, **250**, 434–446.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo,G.D. *et al*. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E.coli. Nucleic Acids Res.*, **10**, 2997–3011.

Thayer,K.M. (2004). http://linus.chem.wesleyan.edu/~kthayer/capDB_index.htm

Wingender,E. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.