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

Sumedha Gunewardena

Oxford University Computing Laboratory
Parks Road, Oxford
OX1 3QD, UK

Zhaolei Zhang

C. H. Best Institute, University of Toronto
112 College Street, Toronto, ON
M5G 1L6, Canada

ABSTRACT

The short and highly degenerate nature of transcription factor (TF) binding sites makes their identification a challenging task. We propose a new method based on *templates* for identifying TF binding sites. Templates account for sequence structure and nucleotide polymorphisms present in TF binding sits providing them with a greater discriminatory capability to methods based on sequence homology.

1. INTRODUCTION

A TF binding site is the sequence of nucleotides on a DNA strand onto which a transcription factor binds and either aid or inhibit gene transcription. The binding of transcription factors to their cognate sites forms an important part in the transcription regulation process, a process important for the control of many cellular functions including cell differentiation, the cell cycle and carcinogenesis. One of the major challenges in bioinformatics is to develop computational methods for recognizing these gene regulatory regions. TF binding sites are relatively short (10-20bp) and highly degenerate sequences, which makes their efficient prediction a computationally challenging task.

The early methods for identifying TF binding sites were based on consensus sequences, degenerate consensus sequences and position specific weight matrices [1]. Some other approaches for finding TF binding sites include rule-based systems [2], General Neural networks [3] and Gibbs Sampling [4].

Studies carried out on various domains of TF binding sites have shown context-dependent effects to be present between different nucleotide positions of the site [5, 6, 7]. This has led many researchers to question the base independence assumption on which techniques such as consensus sequences and weight matrices for identifying TF binding sites are based on [8, 5]. This issue has been addressed by different authors with different techniques ranging from Neural networks [9] to principal coordinates analysis [6]. Templates introduced in this paper, among other things, exploits the presence of nucleotide polymorphisms to improve prediction specificity.

Studies have also shown that nucleotide structure plays a role to some extent in protein-DNA interactions [10, 11]. Nussinov [10] for example, demonstrated structural homology in the three sites at -10, -35 and -16 regions of the Escherichia coli promoter recognized by its polymerase. Structural homology has been used as a discriminatory feature for identifying TF binding sites. Lisser and Margalit [12], for example, describes how the helix stability, helix flexibility and the two conformational parameters represented by the DNA tendencies for B-DNA to Z-DNA and B-DNA to A-DNA transitions describe the E-coli promoter regions. They used linear discriminant analysis based on the relative contributions of these properties towards promoter structure to discriminate between promoter and random sequences. Ponomarenko et al. [13] described a statistical discriminent approach that incorporates dinucleotide conformational angles of direction, wedge, helical twist, roll and tilt to analyse promoters. McPROMOTER [14] is a probabilistic promoter recognition tool that incorporates both sequence information and structural parameters of DNA such as DNA bendability, protein induced deformability and GC content into the recognition process. Thayer and Beveridge [15] describes a model for identifying E-coli catabolite gene activator protein (CAP) binding sites that incorporates both sequence and sequence-dependent structural information into a hidden Markov model.

2. METHOD

We propose a novel approach for finding TF binding sites based on *'templates'*. Templates encapsulate the discriminatory features of nucleotide polymorphism and structural homology along with sequence homology present in TF binding sites for discriminating them from non-binding sites. 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.

Base-independent models of TF binding sites do not account for dependencies that might be present between nucleotides in different positions of the site when interacting with proteins. One problem of modelling nucleotide polymorphisms 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 expression of the positional correlation present in them.

The global expression of positional correlations of a TF binding site, of encoded length L, is captured by a template in the following equation:

$$(\mathbf{Q} \ \mathrm{diag}(\mathbf{r})) \ \mathbf{t} \ = \ \mathbf{r} \ - \ \mathbf{e}$$

Where $\mathbf{t} = (t[1], t[2], \ldots, t[L])^T$, is the vector of template parameters, $\mathbf{r} = (r[1], r[2], \ldots, r[L])^T$, is the vector representing the encoded nucleotide sequence, $\mathbf{e} = (e[1], e[2], \ldots, e[L])^T$, 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 $\mathbf{r} = (r[1], r[2], \ldots, r[L])$, the *template error* of $\mathbf{r}$ with respect to a template $\mathbf{t}$, denoted as $E(\mathbf{r}, \mathbf{t})$, is defined as the sum of squared residual errors.

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

Given a numerical vector $\mathbf{r}$, we can find a set of template parameters $\mathbf{t}$ that minimises the template error $E(\mathbf{r}, \mathbf{t})$ for that vector. This minimisation process is referred to as *'training the template'*. The template $\mathbf{t}$ that minimises $E(\mathbf{r}, \mathbf{t})$ for the vector $\mathbf{r}$ is obtained as follows:

$$\begin{aligned} E(\mathbf{r}, \mathbf{t}) &= \arg\min_{\mathbf{t}} \left(e[1]^2 + e[2]^2 + \ldots + e[L]^2 \right) \\ &= \arg\min_{\mathbf{t}} \left(\mathbf{e}^T \ \mathbf{e} \right) \end{aligned}$$

making the substitution $\mathbf{e} = \mathbf{r} \ - \ (\mathbf{Q} \ \mathrm{diag}(\mathbf{r})) \ \mathbf{t}$

$$= \arg\min_{\mathbf{t}} \left((\mathbf{r} \ - \ \mathbf{Q_r} \ \mathbf{t})^T \ (\mathbf{r} \ - \ \mathbf{Q_r} \ \mathbf{t}) \right)$$

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

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

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

The template that minimises this mean error value for this set of vectors can be obtained by calculating the partial derivatives of Equation 1 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:

$$\mathbf{t} \ = \ \left[\ \sum_{k=1}^{n} \mathbf{Q}_{\mathbf{r}_k}^T \ \mathbf{Q}_{\mathbf{r}_k} \ \right]^{-1} \ \left[\ \sum_{k=1}^{n} \mathbf{Q}_{\mathbf{r}_k}^T \ \mathbf{r}_k \ \right]$$

Where $\mathbf{Q}_{\mathbf{r}_k} = \mathbf{Q} \ \mathrm{diag}(\mathbf{r}_k)$.

These equations are symmetric and can be solved efficiently to find the set of template parameters $t[1], t[2], \ldots, t[L]$ that minimises the mean template error for the set of vectors.

2.1. Classification

Once the templates have been trained, we use linear discriminant analysis (LDA) to distinguish binding sites from non-binding sites based on their template errors. For this, a linear discriminant analyser is trained on the template errors of known positive and negative examples to optimally separate the two classes.

Let $\mathbf{m_s}_{(m \times 1)}$ be the vector of mean template errors from m templates for a set $\mathbf{X_s}_{(n_s \times m)}$ of template errors of positive sites. Let $\mathbf{m_n}_{(m \times 1)}$ be the vector of mean template errors of the templates for a set $\mathbf{X_n}_{(n_n \times m)}$ template errors of negative sites. Then, for LDA classification, we first compute the global mean template error $\mathbf{m_g}_{(m \times 1)}$ of the training examples for the templates as:

$$\mathbf{m_g} = \frac{(n_s - 1)\mathbf{m_s} + (n_n - 1)\mathbf{m_n}}{(n_s + n_n - 2)}$$

The within-class scatter $\mathbf{S_w}_{(m \times m)}$ i.e. the expected covariance of each class, is computed using the equation

$$\mathbf{S_w} = \frac{(n_s - 1)\,\mathrm{cov}(\mathbf{X_s}) + (n_n - 1)\,\mathrm{cov}(\mathbf{X_n})}{(n_s + n_n - 2)}$$

Where, $\mathrm{cov}(\mathbf{X})$ is the variance-covariance matrix of $\mathbf{X}$. The between-class scatter $\mathbf{S_b}_{(m\times m)}$ is computed using the equation

$$\mathbf{S_b} = (\mathbf{m_s} - \mathbf{m_g})(\mathbf{m_s} - \mathbf{m_g})^T + (\mathbf{m_n} - \mathbf{m_g})(\mathbf{m_n} - \mathbf{m_g})^T$$

The between-class scatter can be seen as the covariance of the data set whose members are the mean vectors of each class. Once we have computed $\mathbf{S_w}$ and $\mathbf{S_b}$, we can obtain the optimization criterion $\mathbf{O}_{(m\times m)}$ using the equation

$$\mathbf{O} = {\mathbf{S_w}}^{-1}\mathbf{S_b}$$

To build a transformation matrix $\tilde{\mathbf{O}}_{(m\times k)}$ of reduced dimensions from $\mathbf{O}$, we select all the eigenvectors ($k \leq m$) of $\mathbf{O}$ with non-zero eigenvalues. Given two vectors of template errors, $\mathbf{x}_{(m\times 1)}$ and $\mathbf{y}_{(m\times 1)}$, the squared distance between these two vectors in the transformed space is given by $\mathbf{x}^T\ \hat{\mathbf{O}}\ \mathbf{y}$ where $\hat{\mathbf{O}} = \tilde{\mathbf{O}}\ \tilde{\mathbf{O}}^T$.

Given any vector $\mathbf{x}_{(m\times 1)}$ of template errors, let $D_s(\mathbf{x})^2$ be the squared distance in the transformed space between a vector $\mathbf{x}$ and the mean vector $\mathbf{m_s}$ of template errors for a set of positive sites. Let $D_n(\mathbf{x})^2$ be the squared distance in the transformed space between a vector $\mathbf{x}$ and the mean vector $\mathbf{m_n}$ of template errors for a set of negative sites. These two quantities are given by

$$\begin{aligned} D_s(\mathbf{x})^2 &= (\mathbf{x} - \mathbf{m_s})^T\ \hat{\mathbf{O}}\ (\mathbf{x} - \mathbf{m_s}) \\ D_n(\mathbf{x})^2 &= (\mathbf{x} - \mathbf{m_n})^T\ \hat{\mathbf{O}}\ (\mathbf{x} - \mathbf{m_n}) \end{aligned}$$

We can simplify the above two equations to a single quantity as follows

$$D(\mathbf{x}) = D_n(\mathbf{x})^2 - D_s(\mathbf{x})^2 = \mathbf{A}\ \mathbf{x} + \mathbf{B}$$

Where $\mathbf{A} = 2\ (\mathbf{m_n} - \mathbf{m_s})^T\ \hat{\mathbf{O}}$ and $\mathbf{B} = \frac{1}{2}\ \mathbf{A}\ (\mathbf{m_n} + \mathbf{m_s})$. $D(\mathbf{x})$ is the signed distance spreading an arbitrary vector $\mathbf{x}$, and the discriminator hyper-plane, $D(\mathbf{x}) = 0$, located at the half-distance between the means $\mathbf{m_s}$ and $\mathbf{m_n}$ of the binding sites and non-binding sites used for training the classifier. A positive values of $D(\mathbf{x})$ corresponds to the vector $\mathbf{x}$ representing a binding site and a negative value corresponds to the vector $\mathbf{x}$ representing a non-binding site.

3. RESULTS

Transcription factor binding sites for our work were obtained from Vorobiev *et al.* [16] and Thayer & Beveridge [15]. Mononucleotide representations were obtained by transforming the nucleotide sequence over the 24 different mappings defined by the set $\{A, C, G, T : \{1, 2, 3, 5\} \bullet A \neq C \neq G \neq T\}$ that uniquely represented each nucleotide base for a given assignment. The dinucleotide parameters, representing structural properties of the sequences, were obtained from the Property database [17]. In its current release the database lists 38 different parameter values. We used all 38 of these parameters. The trinucleotide parameters were obtained from Brukner et al. [18]. The mononucleotide values described above, unlike the di and tri nucleotide values used are not based on any particular biophysical feature. They have been selected to provide a convenient transformation that uniquely maps each nucleotide to a numerical representation.

Table 1 shows some experimental results for TF binding-site prediction using templates. The results are for the Nuclear factor kappa B (*NF-κ B/Rel*), *Escherichia coli* catabolite gene activator protein (*CAP*), Nuclear factor-1 (*NF-1*), *CCAAT* box / enhancer binding protein (*C/EBP*) and Activating protein-1 (*AP-1*) families of binding sites. The negative examples for the experiments were generated as follows; we took genomic sequences from the different organisms that contributed towards the TF binding sites of the given factor. All sequences that matched the known TF binding sites of the factor were filtered out of these sequences. The filtered sequences were used to generate sites for the negative examples. We will refer to these sites as negative sites (although there is no guarantee that these are strictly non-binding sites for the given factor). For each experiment we used a set of 1260 negative sites extracted randomly from the filtered sequences of the different species contributing towards the TF binding sites of the given factor. To investigate the robustness of the method, we ran every experiment 100 times, randomly selecting the training sequences for the templates and the classifier from the available positive and negative examples on each trial. The results in Table 1 show the mean values for the different statistics, with the standard deviation given in brackets.

The empirical estimates of the false positive rate of the templates modelled for the TF binding sites listed in Table 1, are shown in Table 2. These results were obtained on a test set of 1130 random sequences taken as negative sites performed over 100 experiments with a new set of data randomly selected for each iteration.

4. DISCUSSION

We have described a novel approach for distinguishing TF binding sites from non-binding sites. 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 discriminative and predictive capabilities of the system.

TF	Train	Classify		Test		Sensitivity	Sensitivity	Specificity	Specificity
		T	F	T	F	Test	Test+Tra	Test	Test+Tra
NF-κB	7	7	130	22	1130	0.90 (0.05)	0.96 (0.03)	0.94 (0.01)	0.96 (0.01)
CAP	7	7	130	11	1130	0.85 (0.08)	0.94 (0.03)	0.91 (0.02)	0.94 (0.02)
NF-1	14	14	130	43	1130	0.90 (0.04)	0.94 (0.04)	0.93 (0.02)	0.95 (0.02)
CEBP	13	13	130	49	1130	0.86 (0.04)	0.93 (0.04)	0.90 (0.02)	0.93 (0.02)
AP-1	9	9	130	23	1130	0.77 (0.04)	0.87 (0.08)	0.91 (0.03)	0.93 (0.03)

Table 1. The mean statistics for 5 different transcription factors. The sensitivity and specificity values listed under the columns 'Test' is those values obtained from only the test data set previously unseen by the classifier and not used for training the templates. The values under the columns 'Test+Tra' are those values obtained from the whole data set, which gives an idea of, how well the classifier does on the training data. The standard deviation, over 100 runs, is given in brackets.

NF-κB	CAP	NF-1	CEBP	AP-1
0.025 (0.009)	0.050 (0.020)	0.060 (0.017)	0.095 (0.022)	0.062 (0.022)

Table 2. Empirical estimates of the expected false positive rate of the template based classifiers.

The training phase of the system requires experimental binding data, a subset representing all the potential binding sites. One advantage of templates is their ability, unlike other machine learning techniques such as neural networks or hidden Markov models to learn quite well from a minimal number of examples. This is a feature that has many practical advantages when we a dealing with a dearth of properly annotated examples. Theoretically, a single pattern is sufficient to construct a template though the resulting template may not well characterise the whole population. This is in contrast to normal regression techniques that require the cardinality of the set of training examples to be at least as great as the number of unknown parameters. Ideally, we would prefer the set of examples used for training to span the entire population.

Binding assays of transcription factors such as *NF-κ B*, *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. 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 behaviour only from a set of training data.

The sensitivity of templates to positional co-variations is not based on any prior knowledge of which positions exhibit polymorphic behaviour. This is an important characterisation, especially in the absence of such prior knowledge individualizing 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 system and the number of examples required to train it 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 [19, 10, 20] that suggests the presence of structural homologies in DNA sequences that interact with some transcription factors. This is the case in for example the E. coli catabolite gene activator protein binding sites. 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 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. This reduction in dimensionality is achieved by LDA. The feature extrac-

tion process removes redundant and irrelevant information providing a more stable representation of the data that leads to improved classification.

The time complexity of searching a whole genome for possible TF 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 (in actual fact, this comes down to the reduced dimensionality of the feature space after feature extraction). The error vector for a sequence of length L can be computed in $\bigcirc(PL^2)$ time. The two 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.

5. REFERENCES

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

[2] 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*, vol. 10, pp. 2997–3011, 1982.

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

[4] 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*, vol. 262(5131), pp. 208–14, 1993.

[5] 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*, vol. 29(12), pp. 2471–8, 2001.

[6] 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*, vol. 99, pp. 8167–8172, 2002.

[7] 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*, vol. 285, pp. 1917–1934, 1999.

[8] 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*, vol. 30(5), pp. 1255–1261, 2002.

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

[10] R. Nussinov, "Promoter helical structure variation at the Escherichia coli polymerase interaction sites.," *Journal of Biological Chemistry.*, vol. 259, pp. 6798–6805, 1984.

[11] M. A. El Hassan and C. R. Calladine, "Two distinct modes of protein-induced bending in DNA," *Journal Molecular Biology*, vol. 282(2), pp. 331–343, 1998.

[12] S. Lisser and H. Margalit, "Determination of common structural features in Escherichia coli promoters by computer analysis.," *Eur J Biochem.*, vol. 223(3), pp. 823–830, 1994.

[13] 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*, pp. 340–351., 1997.

[14] 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*, vol. 17, pp. 199–206., 2001.

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

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

[17] 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*, vol. 15(7/8), pp. 654–668, 1999.

[18] 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*, vol. 14, pp. 1812–1818, 1995.

[19] 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.*, vol. 13 (11), pp. 4085–4096, 1985.

[20] 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*, vol. 259(1), pp. 95–103, 1996.