Identification of Conserved Structural Features at Sequentially Degenerate Locations in Transcription Factor Binding Sites

Heather E. Burden[1,2] hburden@bu.edu
Zhiping Weng[1,2] zhiping@bu.edu

[1] Bioinformatics Program, Boston University, Boston, MA 02215, USA
[2] Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

Abstract

Many locations within transcription factor binding sites are not sequentially conserved and appear to be degenerate. We hypothesize that some of these positions contain essential structural codes that are recognized by the transcription factors that bind to them. The structural codes can be defined by base-pair step parameters that describe the relative displacement and orientation of two adjacent base pairs in a nucleic acid structure. We have developed a method, Identification of Conserved Structural Features (ICSF), which uses base-pair step parameters obtained from a collection of high-resolution DNA crystal structures to discover structural conservation that exists in the sequentially degenerate areas within a binding site and produce profiles of the structural features along the entire site. We have focused our study on the transcription factor binding sites in the JASPAR database and have found that one-third (P-value ≥ 0.05) of the binding sites contain sequentially degenerate locations with highly conserved structural features as described by the base-pair step parameters. These results will help us to gain a better understanding of the process by which transcription factors recognize their binding sites and possibly lead to an improvement in our ability to find these sites in genomic sequences.
Availability: ICSF is freely available to academic users at `http://zlab.bu.edu/ICSF`
Contact: `zhiping@bu.edu`
Supplementary information: `http://zlab.bu.edu/ICSF`

Keywords: transcription factor binding sites, conserved structure, base-pair step parameters, JASPAR, position-specific scoring matrices

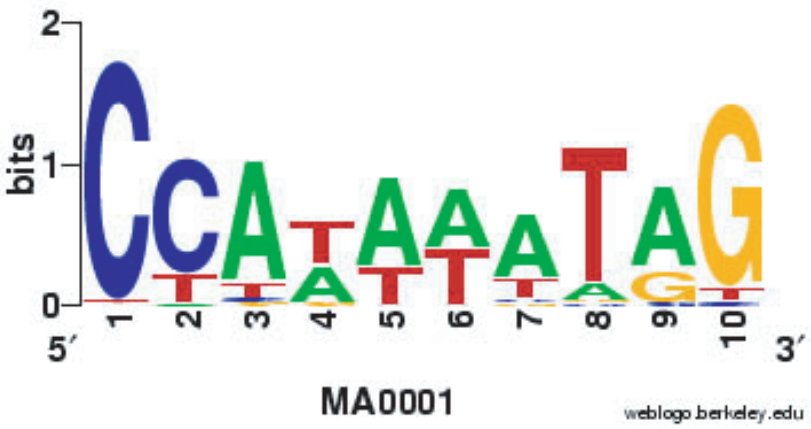
1 Introduction

Transcription factor binding sites are short (usually between 5 and 15 base pairs) and degenerate sequences found in the promoter regions of genes. Transcription factors are proteins that bind to these sequences in order to regulate gene expression. Each transcription factor binds to a unique set of sequences. When the binding sites of a particular transcription factor have been experimentally determined, the nucleotides at each position along the site can be counted and compiled into a position-specific scoring matrix [31] (PSSM) as shown in Figure 1. The PSSM reveals the level of sequence conservation, and hence information content, at each position of the binding site. Some positions within a site are sequentially conserved and have high information content while others allow more than one type of nucleotide and have low information content.

PSSMs are widely used in algorithms to identify transcription factor binding sites in genomic DNA [6, 7, 8, 11, 13, 27]. While there has been success with these algorithms, they are limited by the information content available in the PSSMs [19, 32]. The base-pair step parameters that describe the relative displacement and orientation of two adjacent base pairs in a nucleic acid structure provide

additional information about the binding site [3, 14]. These parameters are measured with respect to coordinate frames embedded in the base pairs as opposed to a global coordinate frame of the entire nucleic acid structure [21]. They consist of three translational and three rotational parameters as described in Figure 2. The x, y and z axes of each base-pair step is a mean or middle frame of the associated base pairs. The translational base-pair step parameters - Shift, Slide and Rise - describe the displacement between two adjacent base pairs along the x, y and z axes, respectively. The rotational base-pair step parameters - Tilt, Roll and Twist - describe the orientation between two adjacent base pairs around the x, y and z axes, respectively. El Hassan and Calladine(1997) argue that the dinucleotide, as opposed to a trinucleotide or tetranucleotide, step is the fundamental unit for a local structural description of DNA [5]. In this study, we use the six base-pair step parameters to describe the structure of the dinucleotide steps in the transcription factor binding sites available in JASPAR, an open-access database of transcription factor binding sites of multicellular eukaryotes [25].

While a PSSM is a useful description of a binding site, it is based only on sequence and therefore the sequentially degenerate positions have low information content. We hypothesize that transcription factors recognize both the sequential and the structural features of these sites and that some of the sequentially degenerate locations have conserved structural features that are represented by the base-pair step parameters. We have developed a method called Identification of Conserved Structural Features (ICSF) and implemented it as a Perl program that uses the experimentally determined binding sites of the transcription factors provided by JASPAR to extract the structural information that resides in the binding sites and identify the structural features that are conserved at sequentially degenerate locations. ICSF is available online at `http://zlab.bu.edu/ICSF`.



A	0	3	79	40	66	48	65	11	65	0
C	94	75	4	3	1	2	5	2	3	3
G	1	0	3	4	1	0	5	3	28	88
T	2	19	11	50	29	47	22	81	1	6

Figure 1: The sequence logo [2, 26] and position-specific scoring matrix of the transcription factor binding site MA0001 (AGL3) [12] of the JASPAR database. When the binding sites of a transcription factor have been experimentally determined, the nucleotides at each position along the site can be counted and compiled into a position-specific scoring matrix (PSSM) that reveals the level of sequence conservation at each position of the binding site.

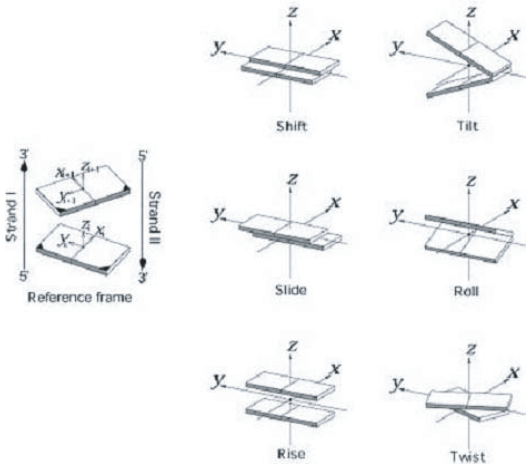


Figure 2: The base-pair step parameters describe the relative displacements and orientations of two adjacent base pairs in a dinucleotide step. Shift, Slide and Rise are the translational parameters along the x, y and z axes, respectively and Tilt, Roll and Twist are the rotational parameters around the x, y and z axes, respectively. This image was kindly provided by Dr. Xiang-Jun Lu and Dr. Wilma K. Olson [15].

2 Materials and Methods

2.1 Transcription Factor Binding Sites

Of the 111 transcription factor binding sites in the JASPAR database, 95 of them contain the experimentally determined binding sites that are used to construct the PSSM of the transcription factor. Our data set contains these 95 sites: MA0001, MA0002, MA0004, MA0005, MA0006, MA0007, MA0008, MA0009, MA0010, MA0011, MA0012, MA0013, MA0014, MA0015, MA0016, MA0017, MA0018, MA0019, MA0020, MA0021, MA0022, MA0023, MA0024, MA0025, MA0026, MA0027, MA0028, MA0029, MA0031, MA0034, MA0036, MA0037, MA0038, MA0040, MA0041, MA0043, MA0044, MA0045, MA0046, MA0047, MA0048, MA0049, MA0051, MA0052, MA0053, MA0054, MA0056, MA0057, MA0058, MA0059, MA0060, MA0061, MA0062, MA0063, MA0064, MA0065, MA0066, MA0067, MA0069, MA0070, MA0071, MA0072, MA0073, MA0074, MA0075, MA0076, MA0077, MA0078, MA0079, MA0080, MA0081, MA0082, MA0083, MA0084, MA0085, MA0086, MA0087, MA0088, MA0089, MA0091, MA0092, MA0093, MA0094, MA0095, MA0096, MA0097, MA0098, MA0101, MA0102, MA0103, MA0104, MA0105, MA0106, MA0107 and MA0111.

2.2 Conserved Structural Features at Sequentially Degenerate Locations

ICSF proceeds as follows for each of the transcription factor binding sites:

1. The six base-pair step parameters of each dinucleotide step in each of the experimentally determined binding sites of a transcription factor are assigned the average values of that type of dinucleotide as determined by Olson *et al.* [22]. In that study, the authors used an ensemble of B-DNA and protein-DNA nucleic acid crystal structures obtained from the Nucleic Acid Database [1] to obtain the average values and standard deviations of the base-pair step parameters for each of the ten unique dinucleotide steps.

2. In order to identify the dinucleotide steps in the binding site that are sequentially degenerate and have a low standard deviation of a base-pair step parameter, a control group is created for comparison. The control group is made up of all combinations of dinucleotides that are possible

from the base frequencies making up the dinucleotide step. For example, if the base frequencies of a dinucleotide step are A= 1, C = 4, G = 3 and T = 0 for base 1 and A = 2, C = 1, G = 1 and T = 4 for base 2, then the control group contains the following dinucleotides: 2 AA, 1 AC, 1AG, 4 AT, 8 CA, 4 CC, 4 CG, 16 CT, 6 GA, 3 GC, 3 GG, 12 GT, 0 TA, 0 TC, 0 TG and 0 TT.

3. The Moses rank-like test is a nonparametric test for comparing differences in variability between two groups in which the medians are not equal [28]. The observations in the two groups are randomly divided into subsets of equal size with any extra data discarded. The subsets are ranked according to their dispersion indexes and the rank sums of each group are calculated. The sampling distribution is approximately that of the normal distribution and a z-score and a corresponding level of significance of the rank sums can be determined from this approximation. The Moses rank-like test is performed on each base-pair step parameter of each dinucleotide step in the binding site to determine if the standard deviation of the base-pair step parameter is lower in the group of observed dinucleotide steps than in the control group and, if so, whether the difference is significant. A z-score and a corresponding level of significance is produced for each base-pair step parameter of each dinucleotide step in the binding site as shown in Figure 3. Since there is an element of randomness in the test, the z-scores presented are the average values of 1,000 tests and presented as absolute values. Z-scores ≥ 2.326 and 1.645 correspond to significance levels ≥ 0.01 and 0.05, respectively. The elegance of the Moses rank-like test is that a high z-score is produced only if the base-pair step parameter is conserved and the dinucleotide step is sequentially degenerate, i.e., structural conservation due to sequence conservation does not receive a high z-score.

(c)

TILT	0.35	0.55	0.64	1.11	1.64	0.64	1.00	0.53	0.69
ROLL	0.47	0.63	0.37	0.54	2.07	0.43	0.51	0.43	0.60
TWIST	0.40	0.56	0.77	2.77	*2.86*	1.22	1.04	0.59	0.93
SHIFT	0.45	0.38	0.38	0.45	0.77	0.49	0.35	0.52	1.04
SLIDE	0.53	0.47	0.58	2.57	1.99	1.57	0.96	0.40	0.37
RISE	0.36	0.33	0.97	0.49	*2.50*	0.51	0.59	0.39	1.58

(b)

CG	0	2	1	0	0	0	0	1	3
[TG	0	1	0	0	0	0	2	22	1
[CA	2	59	1	2	1	2	1	1	0
TA	1	17	7	26	5	29	5	56	0
[CT	18	11	2	1	0	0	4	0	0
[AG	0	0	3	0	0	5	1	3	62
[CC	74	3	0	0	0	0	0	0	0
[GG	0	0	0	1	0	0	0	2	22
[TT	1	0	4	23	24	15	15	0	0
[AA	0	3	31	35	41	34	4	7	0
[TC	0	1	0	1	0	3	0	3	0
[GA	0	0	1	3	1	0	1	1	0
AT	0	0	42	5	23	7	58	1	0
[GT	0	0	2	0	0	0	4	0	6
[AC	0	0	3	0	2	2	2	0	3
GC	1	0	0	0	0	0	0	0	0

(a)

A	0	3	79	40	66	48	65	11	65	0
C	94	75	4	3	1	2	5	2	3	3
G	1	0	3	4	1	0	5	3	28	88
T	2	19	11	50	29	47	22	81	1	6

Figure 3: The Moses rank-like test identifies the base-pair step parameters that are conserved at sequentially degenerate locations in a binding site. (a) The position-specific scoring matrix of the transcription factor binding site MA0001 (AGL3) of the JASPAR database. (b) The position-specific scoring matrix of the dinucleotides of the binding site. Brackets indicate identical dinucleotides. (c) The absolute values of the z-scores (averaged from 1,000 tests) of the base-pair step parameters of each dinucleotide step. The underlined and *italicized* z-scores have significance levels ≥ 0.01 and 0.05, respectively.

2.3 Profiles of the Structural Features

A profile, or parameter matrix, of each of the base-pair step parameters is also produced by ICSF. For each feature, a weight is derived for each type of dinucleotide at each position of the binding site, even if a dinucleotide is not seen in a given position in the experimentally determined binding sites. The parameter matrices take into account the average values as well as the standard deviations of the parameters of each type of dinucleotide step from the ensemble of crystal structures [22] and the z-scores from the Moses rank-like test. The weights in each matrix indicate the extent to which a type of dinucleotide contributes to a conserved feature at that position.

The base-pair step parameters of the dinucleotide steps gathered from the ensemble of crystal structures tend to form normal distributions [22]. Using the average value and standard deviation of a parameter, we can model its normal distribution according to [10]:

$$Y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-\mu)^2}{2\sigma^2}} \tag{1}$$

in which μ is the average value of the parameter, σ is the standard deviation of the parameter and Y is the frequency associated with a value X that falls along the normal distribution. To calculate a parameter weight for a particular type of dinucleotide at a given position, we must determine where the average value of that dinucleotide falls on the normal distributions of each of the dinucleotides that were seen at the position in the experimentally determined binding sites. In this case, X is the average value of the dinucleotide step in question and Y is the frequency associated with this value on the normal distribution of each of the dinucleotides at the position.

To find the weight of a dinucleotide at a given position, compute the sum of the Y values for each of the 16 types of dinucleotides multiplied by the number of times the dinucleotide is seen at that position in the experimentally determined binding sites. An element of a Twist matrix $\mathrm{Twist}_{i,j}$ (the weight of dinucleotide type j at position i) is computed as:

$$\mathrm{Twist}_{i,j}(1) = \sum_{k=1}^{16} N_{i,k} \frac{1}{\sqrt{2\pi \mathrm{Twist}(k)\mathrm{StdDev}^2}} e^{-\frac{(\mathrm{Twist}(j)\mathrm{Average} - \mathrm{Twist}(k)\mathrm{Average})^2}{2\mathrm{Twist}(k)\mathrm{StdDev}^2}}. \tag{2}$$

$N_{i,k}$ is the number of dinucleotide type k seen at position i in the experimentally determined binding sites. Twist(x)Average and Twist(x)StdDev are the average and one-third of the standard deviation, respectively, of the Twist values of dinucleotide type x from the ensemble of crystal structures.

To find the average weight, divide by the number of experimentally determined binding sites:

$$\mathrm{Twist}_{i,j}(2) = \frac{\mathrm{Twist}_{i,j}(1)}{\text{number of binding sites}}. \tag{3}$$

To normalize the parameter matrices with respect to each other, divide by the highest value in the matrix:

$$\mathrm{Twist}_{i,j}(3) = \frac{\mathrm{Twist}_{i,j}(2)}{\text{highest value in matrix}}. \tag{4}$$

To give the more conserved positions more weight than the less conserved positions, multiply each column of the matrix by the parameter's z-score at that position:

$$\mathrm{Twist}_{i,j}(4) = \mathrm{Twist}_{i,j}(3) \times z(\mathrm{twist})_i. \tag{5}$$

Figure 4 gives an example of the Twist matrix for binding site MA0001 in the JASPAR database. The parameter matrices for each of the 95 binding sites is available online at `http://zlab.bu.edu/ICSF`.

3 Results and Discussion

3.1 Conserved Structural Features at Sequentially Degenerate Locations

Our method ICSF identifies locations in transcription factor binding sites in which the sequence is degenerate but the base-pair step parameters describing the structure of a dinucleotide step are conserved. These are positions where a transcription factor is likely to identify the site by the structure rather than the sequence of the DNA. We have found that one-third (32 out of 95, P-value ≥ 0.05) of the transcription factor binding sites in the JASPAR database that were included in this study contain at least one sequentially degenerate dinucleotide step with structurally conserved features, thus proving our hypothesis that some of the sequentially degenerate locations in binding sites are structurally conserved. The sequentially degenerate locations that do not contain structurally conserved features with a P-value ≥ 0.05 are either actually degenerate or are conserved in a manner in which our test is unable to detect. The transcription factor binding sites that contain at least one parameter with a z-score with a significance level ≥ 0.05 are listed in Table 1. Details of the results of each of the 95 binding sites is available online at `http://zlab.bu.edu/ICSF`.

The significance of the similarities of the dinucleotide steps that are found in structurally conserved positions is based upon which dinucleotides could potentially have formed from the base frequencies of the two positions making up the dinucleotide, i.e., the control group. For example, in MA0001 there are 9 dinucleotide steps and the 4^{th} and 5^{th} steps are identified as degenerate with conserved features. In the 4^{th} step, Twist and Slide are conserved and in the 5^{th} step, Roll, Twist, Slide and Rise are conserved. The significance lies in which dinucleotide steps are formed compared to which dinucleotide steps could have formed (the control group). In each of these positions, the degree to which the standard deviation of these parameters is lower than in the control group is statistically significant. This implies that the dinucleotide steps that are observed are present for a biologically significant reason and are not random.

	CG	0.07	0.32	0.32	2.19	1.96	0.86	0.26	0.37	0.08
⌜	TG	0.02	0.37	0.15	1.30	0.76	0.56	0.14	0.44	0.04
⌞	CA	0.02	0.37	0.15	1.30	0.76	0.56	0.14	0.44	0.04
	TA	0.02	0.36	0.11	1.04	0.46	0.47	0.11	0.44	0.03
⌜	CT	0.33	0.10	0.19	0.20	0.35	0.16	0.30	0.04	0.93
⌞	AG	0.33	0.10	0.19	0.20	0.35	0.16	0.30	0.04	0.93
⌜	CC	0.36	0.12	0.17	0.60	0.68	0.28	0.20	0.07	0.80
⌞	GG	0.36	0.12	0.17	0.60	0.68	0.28	0.20	0.07	0.80
⌜	TT	0.15	0.24	0.40	2.48	2.53	0.95	0.31	0.25	0.19
⌞	AA	0.15	0.24	0.40	2.48	2.53	0.95	0.31	0.25	0.19
⌜	TC	0.06	0.34	0.29	2.04	1.75	0.82	0.24	0.39	0.07
⌞	GA	0.06	0.34	0.29	2.04	1.75	0.82	0.24	0.39	0.07
	AT	0.05	0.02	0.39	0.17	0.77	0.12	0.73	0.01	0.20
⌜	GT	0.29	0.09	0.22	0.15	0.38	0.15	0.38	0.04	0.88
⌞	AC	0.29	0.09	0.22	0.15	0.38	0.15	0.38	0.04	0.88
	GC	0.32	0.14	0.24	1.22	1.31	0.49	0.21	0.10	0.58

Figure 4: Structural profile of the base-pair step parameter Twist of the transcription factor binding site MA0001 (AGL3) of the JASPAR database. The profile gives a Twist weight for each type of dinucleotide step at each position of the binding site even if the dinucleotide is not seen in the experimentally determined binding sites. The weights indicate the extent to which a type of dinucleotide step contributes to a conserved Twist score at a given position.

Table 1: Transcription factor binding sites with structurally conserved parameters at sequentially degenerate locations. One-third (32) of the 95 transcription factor binding sites in the JASPAR database that were included in this study contain at least one base-pair step parameter that is structurally conserved at a sequentially degenerate location with a z-score from the Moses rank-like test with a significance level ≥ 0.05. Of these 32 binding sites, 10 have at least one parameter with a z-score with a significance level ≥ 0.01.

P-value $\geq$ 0.01	**P-value $\geq$ 0.05**
MA0001	MA0005
MA0017	MA0007
MA0041	MA0008
MA0048	MA0014
MA0054	MA0019
MA0060	MA0031
MA0065	MA0037
MA0091	MA0047
MA0097	MA0049
MA0102	MA0051
	MA0061
	MA0066
	MA0067
	MA0069
	MA0072
	MA0077
	MA0078
	MA0081
	MA0082
	MA0098
	MA0106
	MA0111

3.2 Profiles of the Structural Features

The parameter matrices that we have produced provide a valuable insight into the contributions of the structural parameters to the identification of the binding sites by the transcription factors. While the z-scores from the Moses rank-like test identify which parameters are conserved at which positions, the parameter matrices go a step further to create an entire profile of each base-pair step parameter at each position of the binding site. The profile provides a quantitative assessment of the contribution of each type of dinucleotide step at each position of the binding site to a conserved structural feature even if the step type is not seen in the experimentally determined binding sites.

3.3 Applications

We tested our parameter matrices to see if they have the potential to increase our ability to identify transcription factor binding sites in genomic DNA. We used a jackknife method in which each experimentally determined binding site is left out, one at a time, and is scored using a PSSM alone and using a PSSM plus parameter matrices derived from the remaining sequences. We used a receiver operating characteristic (ROC) [18] plot to compare how the sequences scored using the PSSM alone versus using the PSSM plus parameter matrices and found the performance to be similar with the two

scoring methods. A primary reason for this may be that there is a limited amount of experimentally determined sites for each transcription factor. If all of the sequences that bind to a transcription factor were available this may change the results dramatically. Another reason could be that the high z-scores from the Moses rank-like test may be caused by a few types of dinucleotide steps that are seen most often in the experimentally determined binding sites yet the presence of a few other types of dinucleotide steps could alter the results when making predictions. Finally, there is no data available in the JASPAR database regarding the binding affinities of the transcription factors to each of the experimentally determined binding sites. This information would be invaluable in the identification of structurally conserved positions as well as in scoring algorithms for predicting binding sites.

A number of other attempts have been made to incorporate structural features of nucleic acids into the identification of transcription factor binding sites [4, 9, 16, 17, 20, 23, 24, 29, 30]. In many of these studies, the authors have sought to find what elements other than sequence are responsible for the recognition of binding sites by transcription factors. Our method should prove useful to further studies in this area by initially identifying the conserved structural features of the dinucleotide steps that may be involved in the identification of the binding site by the transcription factor as well as providing a profile of the conserved nature of each parameter.

Acknowledgments

We thank Dr. Xiang-Jun Lu and Dr. Wilma K. Olson for providing the illustration of base-pair step parameters shown in Figure 2. This work was supported by the National Institutes of Health grant NIH ENCODE R01HG03110.

References

[1] Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R., and Schneider, B., The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids, *Biophys. J.*, 63(3):751–759, 1992.

[2] Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E., WebLogo: A sequence logo generator, *Genome Res.*, 14(6):1188–1190, 2004.

[3] Dickerson, R.E., Definitions and nomenclature of nucleic acid structure parameters, *J. Biomol. Struct. Dyn.*, 6(4):627–634, 1989.

[4] Djordjevic, M., Sengupta, A.M., and Shraiman, B.I., A biophysical approach to transcription factor binding site discovery, *Genome Res.*, 13(11):2381–2390, 2003.

[5] El Hassan, M.A. and Calladine, C.R., Conformational characteristics of DNA: Empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps, *Phil. Trans. R. Soc. Lond. A*, 355:43–100, 1997.

[6] Frith, M.C., Fu, Y., Yu, L., Chen, J.F., Hansen, U., and Weng, Z., Detection of functional DNA motifs via statistical over-representation, *Nucleic Acids Res.*, 32(4):1372–1381, 2004.

[7] Frith, M.C., Hansen, U., Spouge, J.L., and Weng, Z., Finding functional sequence elements by multiple local alignment, *Nucleic Acids Res.*, 32(1):189–200, 2004.

[8] Frith, M.C., Li, M.C., and Weng, Z., Cluster-Buster: Finding dense clusters of motifs in DNA sequences, *Nucleic Acids Res.*, 31(13):3666–3668, 2003.

[9] Fukue, Y., Sumida, N., Nishikawa, J., and Ohyama, T., Core promoter elements of eukaryotic genes have a highly distinctive mechanical property, *Nucleic Acids Res.*, 32(19):5834–5840, 2004.

[10] Gravetter, F.J. and Wallnau, L.B., *Statistics for the Behavioral Sciences: A First Course for Students of Psychology and Education*, West Pub. Co., 1996.

[11] Haverty, P.M., Hansen, U., and Weng, Z., Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification, *Nucleic Acids Res.*, 32(1):179–188, 2004.

[12] Huang, H., Tudor, M., Weiss, C.A., Hu, Y., and Ma, H., The Arabidopsis MADS-box gene AGL3 is widely expressed and encodes a sequence-specific DNA-binding protein, *Plant Mol. Biol.*, 28(3):549–567, 1995.

[13] Johansson, O., Alkema, W., Wasserman, W.W., and Lagergren, J., Identification of functional clusters of transcription factor binding motifs in genome sequences: The MSCAN algorithm, *Bioinformatics*, 19 Suppl.1:i169–i176, 2003.

[14] Lu, X.J. and Olson, W.K., Resolving the discrepancies among nucleic acid conformational analyses, *J. Mol. Biol.*, 285(4):1563–1575, 1999.

[15] Lu, X.J. and Olson, W.K., 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures, *Nucleic Acids Res.*, 31(17):5108–5121, 2003.

[16] Mandel-Gutfreund, Y., Baron, A., and Margalit, H., A structure-based approach for prediction of protein binding sites in gene upstream regions, *Pac. Symp. Biocomput.*, 139–150, 2001.

[17] Mandel-Gutfreund, Y. and Margalit, H., Quantitative parameters for amino acid-base interaction: Implications for prediction of protein-DNA binding sites, *Nucleic Acids Res.*, 26(10):2306–2312, 1998.

[18] Mount, D.W., *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2004.

[19] Ohler, U. and Niemann, H., Identification and analysis of eukaryotic promoters: Recent computational approaches, *Trends Genet.*, 17(2):56–60, 2001.

[20] Ohler, U., Niemann, H., Liao, G., and Rubin, G.M., Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition, *Bioinformatics*, 17 Suppl 1:S199–S206, 2001.

[21] Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., Heinemann, U., Lu, X.J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.S., Westhof, E., Wolberger, C., and Berman, H.M., A standard reference frame for the description of nucleic acid base-pair geometry, *J Mol. Biol.*, 313(1):229–237, 2001.

[22] Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M., and Zhurkin, V.B., DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, *Proc. Natl. Acad. Sci. USA*, 95(19):11163–11168, 1998.

[23] Oshchepkov, D.Y., Vityaev, E.E., Grigorovich, D.A., Ignatieva, E.V., and Khlebodarova, T.M., SITECON: A tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition, *Nucleic Acids Res.*, 32(Web Server issue):W208–W212, 2004.

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

[25] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B., JASPAR: An open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res.*, 32(Database issue):D91–D94, 2004.

[26] Schneider, T.D. and Stephens, R.M., Sequence logos: A new way to display consensus sequences, *Nucleic Acids Res.*, 18(20):6097–6100, 1990.

[27] Sharan, R., Ben-Hur, A., Loots, G.G., and Ovcharenko, I., CREME: Cis-Regulatory Module Explorer for the human genome, *Nucleic Acids Res.*, 32(Web Server issue):W253–W256, 2004.

[28] Siegel, S. and Castellan, N.J., *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, 1988.

[29] Sierk, M.L., Zhao, Q., and Rastinejad, F., DNA deformability as a recognition feature in the reverb response element, *Biochemistry*, 40(43):12833–12843, 2001.

[30] Steffen, N.R., Murphy, S.D., Lathrop, R.H., Opel, M.L., Tolleri, L., and Hatfield, G.W., The role of DNA deformation energy at individual base steps for the identification of DNA-protein binding sites, *Genome Informatics*, 13:153–162, 2002.

[31] Stormo, G.D., DNA binding sites: Representation and discovery, *Bioinformatics*, 16(1):16–23, 2000.

[32] Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z., Assessing computational tools for the discovery of transcription factor binding sites, *Nat. Biotechnol.*, 23(1):137–144, 2005.