Proteome Science

BioMed Central

Review

Open Access

From sequence to structure and back again: approaches for predicting protein-DNA binding

Annette Höglund* and Oliver Kohlbacher

Address: Department for Simulation of Biological Systems, Eberhard Karls University Tübingen, Sand 14, D-72076 Tübingen, Germany

Email: Annette Höglund* - hoeglund@informatik.uni-tuebingen.de; Oliver Kohlbacher - oliver.kohlbacher@uni-tuebingen.de

* Corresponding author

Published: 17 June 2004

Proteome Science 2004, **2**:3

Received: 15 April 2004

Accepted: 17 June 2004

This article is available from: http://www.proteomesci.com/content/2/1/3

© 2004 Höglund and Kohlbacher; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Gene regulation in higher organisms is achieved by a complex network of transcription factors (TFs). Modulating gene expression and exploring gene function are major aims in molecular biology. Furthermore, the identification of putative target genes for a certain TF serve as powerful tools for specific targeting of rational drugs.

Detecting the short and variable transcription factor binding sites (TFBSs) in genomic DNA is an intriguing challenge for computational and structural biologists. Fast and reliable computational methods for predicting TFBSs on a whole-genome scale offer several advantages compared to the current experimental methods that are rather laborious and slow. Two main approaches are being explored, advanced sequence-based algorithms and structure-based methods.

The aim of this review is to outline the computational and experimental methods currently being applied in the field of protein-DNA interactions. With a focus on the former, the current state of the art in modeling these interactions is discussed. Surveying sequence and structure-based methods for predicting TFBSs, we conclude that in order to achieve a sound and specific method applicable on genomic sequences it is desirable and important to bring these two approaches together.

Introduction

A complex network of gene regulatory signals allows each cell in both single- and multicellular organisms to flexibly respond to environmental factors. In 1967, Ptashne realized that gene expression is regulated by protein switches that bind to target sequences in the DNA [1]. Understanding the mechanisms underlying sequence-specific binding of proteins to DNA and the resulting gene expression, holds great promise for targeting numerous diseases through rational drug development [2].

The sequencing of whole genomes alongside with experimental studies of the control of gene expression has revealed some fundamental mechanisms. Each gene is regulated by at least one, but often multiple transcription factor (TFs). The TFs bind to specific transcription factor binding sites (TFBSs) within the regulatory regions (promoters) of the genes. The functional arrangement, i.e. the presence, combination, and order of the TFBSs in a regulatory region, form promoter modules [3] that control the spatial and temporal expression of genes [4].

The analysis of individual TFBSs can provide important clues in deducing regulatory networks in a cell and the functional context of specific genes. Over time, several experimental methods have been developed for studying

TFBSs. *In vitro* analysis is complicated by two facts typical of TFs: TFs usually bind to multiple target sequences with varying affinity and they often regulate multiple genes. *In silico* analysis is not straight-forward either, but presents a necessary extension to current *in vitro* methods. The main obstacles are that TFBSs are often located in non-coding DNA, degenerate in their sequence, and relatively short (5–12 nucleotides). Searching for such low-information content sites within huge amounts of genomic DNA using computational methods typically yields a large number of randomly occurring false positive sites. Reducing the number of these false positives has been the goal of many efforts. Currently, most successful sequence-based algorithms are context-sensitive and account for the presence of other TFBSs [5], relative positioning to transcription start site (TSS) [6], and evolutionary conservation of functional regulatory elements [7]. Seen from a structural point of view, the recognition of a nucleotide sequence by a DNA-binding protein is determined by the interactions between the DNA base pair (bp) edges and the amino acid side chains. Structure-based methods use either statistical information obtained from structural data, or models for representing the steric and chemical complementarity, for evaluating the affinity of a protein-DNA complex [8].

Research during the past decades has focused on understanding the mechanisms underlying protein-DNA interactions and aiming towards expressing these using general sets of rules. First attempts to define such a recognition code arose in 1976 through the work of Seeman and Rosenberg [9], who identified a specific pattern of hydrogen bond (H-bond) acceptors and donors on the DNA bp edges. More detailed studies of protein-DNA structural complexes soon concluded that the interactions could not be explained by a simple one-to-one correspondence [10,11]. However, specific amino acid-base preferences do exist [12,13], which comes as no surprise given their chemical and structural characteristics.

Current sequence-based algorithms and structure-based models will benefit from a mutual integration, when the primary aim is to develop fast and reliable prediction methods for TFBSs and an understanding for how DNA recognition is facilitated. Experimental techniques for studying protein-DNA interactions and the physical characteristics of such interactions will be explained in the first two sections. In the final section, accurate computational modeling of the binding sites of regulatory proteins will be discussed in the light of experimental and theoretical implications.

Experimental methods

In order to be able to analyze differences and commonalities of how binding takes place, examples of binding sites are required. Experimental methods used in the determination of binding sites for transcription factors are important for creating a sound description of each TFBS.

There are a several methods available for producing interaction data. Nitrocellulose-binding assay [14], electrophoretic mobility shift assay (EMSA) [15], enzyme-linked immunosorbent assay (ELISA) [16], DNase 1 footprinting [17], DNA-protein crosslinking (DPC) [18], and reporter conducts [19] are examples of *in vitro* techniques that are used for determining DNA binding sites and analyzing the difference in binding specificity for different protein-DNA complexes. They are all currently in use, but suffer from major drawbacks: they are not suited for high-throughput experiments and information on optimal vs. suboptimal protein binding sites is lost.

Chromatin immunoprecipitation (ChIP) is a recent microarray-based assay developed for genome-wide determination of protein binding sites on DNA [20]. Systemic evolution of ligands by exponential enrichment (SELEX) [21] and Phage Display (PD) [22] represent another type of experiments and offer a high-throughput possibility to select high-affinity binders, DNA and protein targets respectively. Both SELEX and PD suffer from the same drawback, the fact that the multitude of sequences obtained from these experiments are all good binders, but it is hard to say anything about their relative affinities. The assumption that the best binders occur more frequently, from purely statistical reasons, is commonly adopted. The differences between individual mutants have to be measured one at a time by other and more laborious methods (discussed above).

In 1999, Bulyk *et al*. presented dsDNA microarrays for exploring sequence specific protein-DNA binding [23]. The major advantage over the methods discussed above is that it is a high-throughput method resulting in data with associated relative binding affinities, which is of high importance in protein-DNA interaction studies.

Finally, there is X-ray crystallographic and NMR spectroscopic data providing a base for studying the structural details of protein-DNA interactions. Protein-DNA complexes have successfully been co-crystallized [24], and the data has been deposited into the Protein Data Bank (PDB) and Nucleic Acid Database (NDB). Each complex is a 3D representation of all intermolecular interactions participating in protein-DNA recognition, however, the experiments are very time-consuming.

Characteristics of protein-DNA interactions

Double-stranded DNA forms the famous double helix [25], where pairs of complementary bases on opposing strands are stabilized by intermolecular H-bonds. The chemical composition of the DNA sugar-phosphate

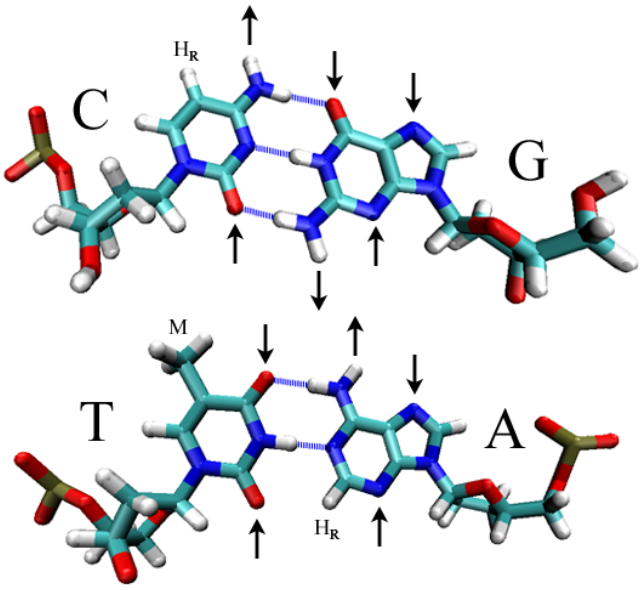

Figure 1
Characteristics of C-G and T-A base pairs Intermolecular H-bonds (dotted lines) in the C-G and T-A bp, stabilize the DNA double helix. The bp edges form a pattern of H-bond acceptors and donors that can be recognized by amino acid side chains of proteins. This pattern is unique for each bp (C-G, G-C, T-A, and A-T) in the major groove (up), whereas it is only possible to distinguish a C-G bp (top) form an T-A bp (bottom) in the minor groove (down) [9]. H-bond acceptors and donors are indicated by outward and inward pointing arrows respectively. The letter M is the methyl group of the base T and H_R is a ring hydrogen donor. The chemical composition of the DNA sugar-phosphate backbone (not shown) is constant and independent of the bp sequence.

backbone is independent of the bp sequence and thus not involved in the specificity of sequence recognition. Only the edges of the bp are exposed in the grooves of the helical DNA, where they form a pattern of H-bond acceptors and donors [9] that can be recognized by the amino acid side chains, see Figure 1 for an illustration. Specific recognition of DNA has to rely on the interactions with these exposed patches. TFs typically contain a DNA-binding domain and one or multiple interaction domains that bind to other TFs. It is common to group the TFs into families according to the structure of these DNA-binding domains [26], where each family employs a different mechanism for recognizing the DNA sequence of the target site [12].

The energetics and mode of protein-DNA interactions differ from those of protein-protein interactions. The main differences are that the protein-DNA interfaces are much more polar, have many more intermolecular H-bonds, and a higher abundance of buried water molecules [27,28]. The most important biochemical interactions in protein-DNA complexes are van der Waals contacts, H-bonds, and water-mediated contacts [29]. About two-thirds of all contacts are non-specific and made with the sugar-phosphate backbone of the DNA, leaving one-third of all interactions for the specificity [30]. Nonspecific interactions (protein-DNA backbone) are extremely important for the overall stability of the complex, and are mainly mediated through van der Waals contacts. About two-thirds of the specific interactions (protein-DNA base edges) involve complex H-bond patterns [29]. The distribution of H-bonds clearly demonstrates particular amino acid-base preferences, but no generalizable code can be deduced [13]. It is important to note that each amino acid can interact with more than one bp simultaneously, and several different amino acids can interact with the same bp. Interdependence between both bases and amino acids is an important feature of the interaction scheme. Very specific contact patterns can be achieved in this way and enable subtle but crucial differences in binding affinities [31].

Water molecules act as contact-mediators and space-fillers at the protein-DNA interface and play a key role in complex formation. As suggested in [32], an atomic description of water molecules at the interface is required for a complete formulation of protein-DNA interactions. Important water bridges can be identified in crystal structures or using molecular modeling [33].

The helical DNA structure is often distorted when bound to a protein [34,35]. Enforced bending of the DNA strand occurs through kinks at the base steps, leading to unstacking and unwinding of the helix. Several types of structural changes have been detected, including shift, slide, twist, rise, roll, and tilt [36]. The stiffness of the DNA helix is determined by the background bp composition [37], i.e. C-G bp are more rigid since they have one additional H-bond compared to A-T bp. The side chains of the protein are flexible and can re-arrange upon complex formation in order to achieve complementarity.

Computational methods

Computational approaches present an attractive solution for modeling and discovering TFBSs on a genomic scale. Several different computational approaches for predicting TFBSs have been explored, which has lead to considerable progress during recent years. The main approaches are sequence and structure-based, where the difference is that sequence-based methods consider only the primary structure of DNA, whereas structure-based methods aim at describing the physical and chemical complementarity

Table 1: Representation of an example TFBS. Two sequence-based representations of the same TFBS, a consensus sequence and a position specific scoring matrix (PSSM). The example used here is the binding site of the early growth response protein 1 (EGR-1, Zif268), which is a zinc finger protein.

CONSENSUS		T	G	C	G	T	G	G	G	C	G
POSITION		1	2	3	4	5	6	7	8	9	10
SCORING MATRIX	A	5	7	0	2	0	31	0	0	13	0
	C	3	0	**98**	0	2	0	0	0	**76**	0
	G	5	**93**	0	**98**	14	**69**	**100**	**100**	0	**100**
	T	**87**	0	2	0	**84**	0	0	0	11	0

between a TF and its binding site. We will now briefly discuss some selected sequence and structure-based computational methods for predicting TFBSs.

Experimentally verified binding sites can be used for constructing a consensus sequence motif of the binding site of a TF. A consensus sequence can be obtained from a multiple alignment of known binding sites [38], and can be used for scanning genomic sequences in the search for TFBSs [39,40]. However, methods using scoring matrices for describing the binding sites [41,42] offer great advantages over consensus sequence methods. Position specific scoring matrices (PSSMs) are based on experimentally verified binding sites and represent the relative distribution and conservation of all nucleotides in the binding site. PSSMs exist for almost all types of TFBSs [43] and are widely used for predicting binding sites [41]. For an excellent review on PSSMs, see [44]. Table 1 is an illustration of a consensus sequence and a PSSM for an example TFBS. Sequence logos can be used for graphically describing the PSSMs [45]. The main advantage of PSSMs is that a qualitative measure can be obtained rather than the yes/no type of answer obtained from consensus models. Accounting for interdependence [46] between bases in the TFBS is not trivial, thus treating the binding energy contribution of each position in the binding site as independent ("independent binding hypothesis") is a frequently adopted approximation [47]. However, some improvement in performance has been achieved using higher order PSSM models [48,49].

False positive hits are detected with high frequencies [50], when using consensus or PSSMs for scanning genomes for putative binding sites. Bringing genetic context into the models has improved the specificity of the prediction methods. Limiting the search to predicted promoter regions [6,51], combining a set of functionally related TFs [4], and searching for their co-abundance has increased the specificity significantly [40]. The inclusion of spacing-rules between the TFs [52], limitations of the number of each contributing TF [53], and combinatorial aspects of TF positioning [54,55] has further reduced the number of false hits.

Several TFs bind their target sequences as homo- or heterodimers, leading to co-occurring binding sites. The number of nucleotides in the gaps between the two half-sites may vary, even for the same TF binding to two different sites [56]. Accounting for varying half-site spacing in computational search algorithms is not trivial, nevertheless essential. Synergy, or cooperative binding is another reason for co-occurring motifs. Per definition, classical cooperative binding is when protein-protein interactions lead to a more efficient control of the promoter. Biological experiments have shown that synergistic activation can also occur when two regulatory proteins have no physical contact [57]. Computer simulations indicate that this might be an effect of the protein first binding changing the tension in the DNA strand [58]. Several computational methods predicting TFBS have been developed that take such putative synergy effects into account. BioProspector [59] and Co-Bind [5] are examples of methods that can be used for discovering co-occurring motifs.

Computational *de novo* discovery of overrepresented motifs has been used for finding putative and functionally related TFBSs. Detecting short and degenerate binding sites in genomic sequences is a very hard task. Limiting the search to promoters and conserved non-coding regions where TFBSs are enriched [60] has improved the performance. Gibbs Sampling [61], Ann-SPEC [62], and LOGOS [63] are examples of algorithms that have proven helpful in detecting TFBSs [64,65]. Further improvement has been made by assuming that co-expressed genes are co-regulated [66], at least to some extent. Inferring co-expression in order to detect overrepresented motifs in regulatory sequences has frequently been adopted [67,68]. Phylogenetic footprinting is a computational method commonly applied as a filter for pointing towards conserved, possibly functional regions of non-coding regulatory sequences [7,69]. Several successful examples have

been reported [70,71], and the computational methods have been reviewed in [72,73].

Alongside with an increasing number of genomic sequences, the amount of structural information on protein-DNA complexes has been increasing rapidly. Careful structural analyses of protein-DNA complexes obtained from PDB and NDB have identified the characteristics of such interactions [13,27,29]. Examination of the relationship between amino acid sequence conservation and role in DNA sequence recognition in protein-DNA complexes has revealed a strong correlation across all protein structural families [74,75].

Structure-based models offer promising extensions to the sequence-based models. These provide a way to qualitatively analyze DNA deformation, cooperativity, and other structural properties of protein-DNA interactions. There are mainly two categories of structure-based approaches. The first one is based on statistical potentials and the second one on potentials obtained from molecular mechanics simulations. Statistical potentials are derived from systematic analysis of structural protein-DNA complexes. Pairwise potentials are extracted from distributions of C_α atoms around DNA bases of known protein-DNA complexes, which reflect the statistical occurrence of specific interactions. They have proven to be sufficiently sensitive to evaluate the affinities of sequences obtained in a combinatorial fashion by threading them onto the fold of the original complex [8,76]. Computer simulations have been used to derive free-energy interaction maps between pairs of bases and amino acids [77,78], which can be used for prediction of TFBSs in a similar fashion as described above. In order to fully address structural flexibility of both protein and DNA, and interaction redundancy, intensive computation is needed. Observing processes during appropriate simulation periods and accounting for whole-system interactions are the two main limiting factors. Despite the required computing power, free energies have been analyzed in larger biological systems, see [79] for a review. Encoding the structural properties of specific DNA sequences and using these in combination with sequence-based methods can improve the specificity of the predictions [28,80].

The direct interactions between amino acids and DNA bases are mainly specific hydrogen bonds, which are fairly well understood. The non-specific interactions, constituting the majority of all interactions involved, are less well understood yet nevertheless, indications exist that these will provide important clues in understanding the complete picture of protein-DNA recognition. Structure-based approaches for modeling protein-DNA interactions are expensive regarding computing power, however, they provide valuable insights into the physical interactions at an atomic level.

Conclusion

Protein-DNA interactions have been under intense research during recent years, which has resulted in numerous valuable finding as well as computational methods for the prediction of TFBSs. While sequence-based methods are amenable to analyses on a whole-genome scale, the computational costs for structure-based methods are currently still prohibitively high. The required computation time ranges up to several days for one single protein-DNA complex, due to the complexity of the interactions. At the same time, structure-based methods provide deep insights into the mechanisms and features of the protein-DNA interaction. These insights allow us to validate – or falsify – some of the assumptions and approximations underlying some of the sequence-based methods. Sequence-based algorithms also provide a fast and flexible system for analyzing and reducing the search space in genomic sequences, whereas computationally intensive structure-based approaches can then be used in a final step with the specificity needed for a final evaluation of the predicted binding sites.

We hence observe both a need and a recent tendency to use structure-based methods for validation of sequence-based methods. We conclude that advanced sequence-based methods and detailed structure-based methods will make a strong combination in the search for putative binding sites for regulatory proteins in genomic sequences.

References

1. Ptashne M: **Specific binding of the lambda phage repressor to lambda DNA.** *Nature* 1967, **214:**232-234.
2. Bartsevich VV, Miller JC, Case CC, Pabo CO: **Engineered zinc finger proteins for controlling stem cell fate.** *Stem Cells* 2003, **21:**632-637.
3. Werner T: **Models for prediction and recognition of eukaryotic promoters.** *Mamm Genome* 1999, **10:**168-175.
4. Wasserman W, Fickett J: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:**167-181.
5. GuhaThakurta D, Stormo GD: **Identifying target sites for cooperatively binding factors.** *Bioinformatics* 2001, **17:**608-621.
6. Levy S, Hannenhalli S: **Identification of transcription factor binding sites in the human genome sequence.** *Mamm Genome* 2002, **13:**510-514.
7. Wasserman W, Palumbo M, Thompson W, Fickett J, Lawrence C: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26:**225-228.
8. Kono H, Sarai A: **Structure-based prediction of DNA target sites by regulatory proteins.** *Proteins* 1999, **35:**114-131.
9. Seeman NCRA, Rosenberg JM: **Sequence-specific recognition of double helical nucleic acids by proteins.** *Proc Natl Acad Sci U S A* 1976, **73:**804-808.
10. Pabo C, Sauer R: **Protein-DNA recognition.** *Annu Rev Biochem* 1984, **53:**293-321.
11. Matthews B: **Protein-DNA interaction. No code for recognition.** *Nature* 1988, **335:**294-295.

12. Pabo C, Sauer R: **Transcription factors: structural families and principles of DNA recognition.** *Annu Rev Biochem* 1992, **61:**1053-1095.
13. Mandel-Gutfreund Y, Schueler O, Margalit H: **Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles.** *J Mol Biol* 1995, **253:**370-382.
14. Woodbury CJ, von Hippel P: **On the determination of deoxyribonucleic acid-protein interaction parameters using the nitrocellulose filter-binding assay.** *Biochemistry* 1983, **22:**4730-4737.
15. Garner M, Revzin A: **A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system.** *Nucleic Acids Res* 1981, **9:**3047-3060.
16. Choo Y, Klug A: **A role in DNA binding for the linker sequences of the first three zinc fingers of TFIIIA.** *Nucleic Acids Res* 1993, **21:**3341-3346.
17. Galas D, Schmitz A: **DNAse footprinting: a simple method for the detection of protein-DNA binding specificity.** *Nucleic Acids Res* 1978, **5:**3157-3170.
18. Molnar G, O'Leary N, Pardee A, Bradley D: **Quantification of DNA-protein interaction by UV crosslinking.** *Nucleic Acids Res* 1995, **23:**3318-3326.
19. Hanes S, Brent R: **A genetic model for interaction of the homeodomain recognition helix with DNA.** *Science* 1991, **251:**426-430.
20. Ren B, Robert F, Wyrick J, Aparicio O, Jennings E, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert T, Wilson C, Bell S, Young R: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290:**2306-2309.
21. Choo Y, Klug A: **Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions.** *Proc Natl Acad Sci U S A* 1994, **91:**11168-11172.
22. Choo Y, Klug A: **Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage.** *Proc Natl Acad Sci U S A* 1994, **91:**11163-11167.
23. Bulyk M, Gentalen E, Lockhart D, Church G: **Quantifying DNA-protein interactions by double-stranded DNA arrays.** *Nat Biotechnol* 1999, **17:**573-577.
24. Kim J, Burley S: **1.9 A resolution refined structure of TBP recognizing the minor groove of TATAAAAG.** *Nat Struct Biol* 1994, **1:**638-653.
25. Watson J, Crick F: **The structure of DNA.** *Cold Spring Harb Symp Quant Biol* 1953, **18:**123-131.
26. Luscombe N, Austin S, Berman H, Thornton J: **An overview of the structures of protein-DNA complexes.** *Genome Biol* 2000, **1:**. Review
27. Jones S, van Heyningen P, Berman H, Thornton J: **Protein-DNA interactions: A structural analysis.** *J Mol Biol* 1999, **287:**877-896.
28. Woda J, Schneider B, Patel K, Mistry K, Berman H: **An analysis of the relationship between hydration and protein-DNA interactions.** *Biophys J* 1998, **75:**2170-2177.
29. Luscombe NM, Laskowski RA, Thornton JM: **Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level.** *Nucleic Acids Res* 2001, **29:**2860-2874.
30. Steffen N, Murphy S, Tolleri L, Hatfield G, Lathrop R: **DNA sequence and structure: direct and indirect recognition in protein-DNA binding.** *Bioinformatics* 2002, **18:**S22-S30.
31. Isalan M, Klug A, Choo Y: **Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers.** *Biochemistry* 1998, **37:**12026-12033.
32. Schwabe J: **The role of water in protein-DNA interactions.** *Curr Opin Struct Biol* 1997, **7:**126-134.
33. Tsui V, Radhakrishnan I, Wright P, Case D: **NMR and molecular dynamics studies of the hydration of a zinc finger-DNA complex.** *J Mol Biol* 2000, **302:**1101-1117.
34. Steffen NR, Murphy SD, Lathrop RH, Opel ML, Tolleri L, Hatfield GW: **The role of DNA deformation energy at individual base steps for the identification of DNA-protein binding sites.** *Genome Inform Ser Workshop Genome Inform* 2002, **13:**153-162.
35. Chen S, Vojtechovsky J, Parkinson G, Ebright R, Berman H: **Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: DNA binding specificity based on energetics of DNA kinking.** *J Mol Biol* 2001, **314:**63-74.
36. Gromiha M, Siebers J, Selvaraj S, Kono H, Sarai A: **Intermolecular and intramolecular readout mechanisms in protein-DNA recognition.** *J Mol Biol* 2004, **337:**285-294.
37. Hogan M, Austin R: **Importance of DNA stiffness in protein-DNA binding specificity.** *Nature* 1987, **329:**263-266.
38. Pribnow D: **Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter.** *Proc Natl Acad Sci U S A* 1975, **72:**784-788.
39. Prestridge D: **SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements.** *Comput Appl Biosci* 1991, **7:**203-206.
40. Frech K, Herrmann G, Werner T: **Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids.** *Nucleic Acids Res* 1993, **21:**1655-1664.
41. Chen Q, Hertz G, Stormo G: **MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices.** *Comput Appl Biosci* 1995, **11:**563-566.
42. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23:**4878-4884.
43. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S: **The TRANSFAC system on gene expression regulation.** *Nucleic Acids Res* 2001, **29:**281-283.
44. Stormo G: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
45. Schneider T, Stephens R: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18:**6097-6100.
46. Bulyk M, Johnson P, Church G: **Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors.** *Nucleic Acids Res* 2002, **30:**1255-1261.
47. Benos PV, Bulyk ML, Stormo GD: **Additivity in protein-DNA interactions: how good an approximation is it?** *Nucleic Acids Res* 2002, **30:**4442-4451.
48. Benos PV, Lapedes AS, Fields DS, Stormo GD: **SAMIE: statistical algorithm for modeling interaction energies.** *Pac Symp Biocomput* 2001:115-126.
49. Barash Y, Elidan G, Friedman N, Kaplan T: **Modeling dependencies in protein-DNA binding sites.** *RECOMB'03* 2003.
50. Wasserman W, Krivan W: **In silico identification of metazoan transcriptional regulatory regions.** *Naturwissenschaften* 2003, **90:**156-166.
51. Werner T: **The state of the art of mammalian promoter recognition.** *Brief Bioinform* 2003, **4:**22-30.
52. Frith M, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17:**878-889.
53. Kielbasa S, Korbel J, Beule D, Schuchhardt J, Herzel H: **Combining frequency and positional information to predict transcription factor binding sites.** *Bioinformatics* 2001, **17:**1019-1026.
54. Pilpel Y, Sudarsanam P, Church G: **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 2001, **29:**153-159.
55. Zhu Z, Pilpel Y, Church G: **Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm.** *J Mol Biol* 2002, **318:**71-81.
56. Suckow MHC, Kisters-Woike B: **A novel feature of DNA recognition: a mutant Gcn4p bZip peptide with dual DNA binding specificities dependent of half-site spacing.** *J Mol Biol* 1999, **286:**983-987.
57. Vashee S, Willie J, Kodadek T: **Synergistic activation of transcription by physiologically unrelated transcription factors through cooperative DNA-binding.** *Biochem Biophys Res Commun* 1998, **247:**530-535.
58. Rudnick J, Bruinsma R: **DNA-protein cooperative binding through variable-range elastic coupling.** *Biophys J* 1999, **76:**1725-1733.
59. Liu X, Brutlag D, Liu J: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Pac Symp Biocomput* 2001:127-138.
60. Levy S, Hannenhalli S, Workman C: **Enrichment of regulatory signals in conserved non-coding genomic sequence.** *Bioinformatics* 2001, **17:**871-877.

61. Lawrence C, Altschul S, Boguski M, Liu J, Neuwald A, Wootton J: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262:**208-214.
62. Workman CT, Stormo GD: **ANN-Spec: a method for discovering transcription factor binding sites with improved specificity.** *Pac Symp Biocomput* 2000:467-478.
63. Xing E, Wu W, Jordan M, Karp R: **LOGOS: A modular Bayesian model for de novo motif detection.** *IEEE Computer Society Bioinformatics Conference, CSB2003* 2003.
64. Hughes J, Estep P, Tavazoie S, Church G: **Computational identification of cis-regulatory elements associated with functionally coherent groups of genes in Saccharomyces cerevisiae.** *J Mol Biol* 2000, **296:**1205-1214.
65. Sandelin A, Höglund A, Lenhard B, Wasserman W: **Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes.** *Funct Integr Genomics* 2003, **3:**125-134.
66. Ideker T, Thorsson V, Ranish J, Christmas R, Buhler J, Eng J, Bumgarner R, Goodlett D, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001, **292:**929-934.
67. Wang T, Stormo G: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs.** *Bioinformatics* 2000, **19:**2369-2380.
68. Bussemaker H, Li H, Siggia E: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27:**167-171.
69. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome Res* 2002, **12:**739-748.
70. McCue L, Thompson W, Carmack C, Ryan M, Liu J, Derbyshire V, Lawrence C: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29:**774-782.
71. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman W: **Identification of conserved regulatory elements by comparative genome analysis.** *J Biol* 2003, **2:**.
72. Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting.** *J Comput Biol* 2002, **9:**211-223.
73. Zhang Z, Gerstein M: **Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements.** *J Biol* 2003 in press.
74. Luscombe N, Thornton J: **Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity.** *J Mol Biol* 2002, **320:**991-1009.
75. Mirny LA, Gelfand MS: **Structural analysis of conserved base pairs in protein-DNA complexes.** *Nucleic Acids Res* 2002, **30:**1704-1711.
76. Mandel-Gutfreund Y, Baron A, Margalit H: **A structure-based approach for prediction of protein binding sites in gene upstream regions.** *Pac Symp Biocomput* 2001:139-150.
77. Pichierri F, Aida M, Gromiha M, Sarai A: **Free-energy maps of base-amino acid interactions for protein-DNA recognition.** *J Am Chem Soc* 1999, **121:**6152-6257.
78. Yoshida T, Nishimura T, Aida M, Pichierri F, Gromiha MMnSA: **Evaluation of free energy landscape for base-amino acid interactions using ab initio force field and extensive sampling.** *Biopolymers* 2001, **61:**84-95.
79. Lazaridis T: **Binding Affinity and Specificity from Computational Studies.** *Current Organic Chemistry* 2002, **6:**1319-1332.
80. Karas H, Knuppel R, Schulz W, Sklenar H, Wingender E: **Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements.** *Comput Appl Biosci* 1996, **12:**441-446.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral