This information has not been peer-reviewed. Responsibility for the findings rests solely with the author(s).

Deposited research article

All motifs are not created equal: structural properties of transcription factor - dna interactions and the inference of sequence specificity

Michael B Eisen

Addresses: Center for Integrative Genomics, Division of Genetics and Development, Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, USA. Department of Genome Sciences, Genomics Division, Ernest Orlando, Lawrence Berkeley National Lab, Berkeley, USA. E-mail: MBEISEN@LBL.GOV

Posted: 31 March 2005

Genome ***Biology*** 2005, **6:**P7

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/5/P7

© 2005 BioMed Central Ltd

Received: 30 March 2005

This is the first version of this article to be made available publicly.

.deposited research

AS A SERVICE TO THE RESEARCH COMMUNITY, GENOME **BIOLOGY** PROVIDES A 'PREPRINT' DEPOSITORY TO WHICH ANY ORIGINAL RESEARCH CAN BE SUBMITTED AND WHICH ALL INDIVIDUALS CAN ACCESS FREE OF CHARGE. ANY ARTICLE CAN BE SUBMITTED BY AUTHORS, WHO HAVE SOLE RESPONSIBILITY FOR THE ARTICLE'S CONTENT. THE ONLY SCREENING IS TO ENSURE RELEVANCE OF THE PREPRINT TO GENOME **BIOLOGY**'S SCOPE AND TO AVOID ABUSIVE, LIBELLOUS OR INDECENT ARTICLES. ARTICLES IN THIS SECTION OF THE JOURNAL HAVE **NOT** BEEN PEER-REVIEWED. EACH PREPRINT HAS A PERMANENT URL, BY WHICH IT CAN BE CITED. RESEARCH SUBMITTED TO THE PREPRINT DEPOSITORY MAY BE SIMULTANEOUSLY OR SUBSEQUENTLY SUBMITTED TO GENOME **BIOLOGY** OR ANY OTHER PUBLICATION FOR PEER REVIEW; THE ONLY REQUIREMENT IS AN EXPLICIT CITATION OF, AND LINK TO, THE PREPRINT IN ANY VERSION OF THE ARTICLE THAT IS EVENTUALLY PUBLISHED. IF POSSIBLE, GENOME **BIOLOGY** WILL PROVIDE A RECIPROCAL LINK FROM THE PREPRINT TO THE PUBLISHED ARTICLE.

All Motifs are NOT Created Equal:

Structural Properties of Transcription Factor – DNA Interactions and the Inference of Sequence Specificity

Michael B. Eisen

Affiliations:

Center for Integrative Genomics
Division of Genetics and Development
Department of Molecular and Cell Biology
University of California Berkeley
Berkeley, CA

Department of Genome Sciences
Genomics Division
Ernest Orlando Lawrence Berkeley National Lab
Berkeley, CA

Contact:

Michael B. Eisen
Mailstop 84-171
One Cyclotron Road
Berkeley, CA 94720

Email: MBEISEN@LBL.GOV
Tel: +1 (510) 486-5214
FAX: +1 (786) 549-0137

Abstract

The identification of transcription factor binding sites in genome sequences is an important problem in contemporary sequence analysis, and a plethora of approaches to the problem have been proposed, implemented and evaluated in recent years. Although the biological and statistical models, descriptions of binding sites and computational algorithms used vary considerably amongst these methods, most share a common assumption – that all motifs are equally likely to be transcription factor binding sites. Here we argue that this simplifying assumption is incorrect – that the specific nature of transcription factor-DNA interactions imposes constraints on the types of motifs that are likely to be transcription factor binding sites and on the relationships between motifs recognized by members of structurally similar transcription factors. We propose that our structural and biochemical understanding of the interactions between transcription factors and DNA can be used to guide *de novo* motif detection methods, and, in a series of related papers introduce several methods that incorporate this idea.

Introduction:

Of the myriad ways that cells control the abundance and activity of the proteins encoded by their genomes, regulation of mRNA synthesis is perhaps the most general and significant. Transcriptional regulation plays a central role in a multitude of critical cellular processes and responses, and is a central force in the development and differentiation of multicellular organisms. There has thus been considerable interest in understanding how genome sequences specify when and where genes should be

transcribed, and the availability of a wide range of genome sequences has greatly accelerated research to decipher the genomic regulatory code.

Although they are only part of the complex networks that regulate transcription, sequence specific DNA binding proteins (transcription factors) provide a crucial link between DNA sequence and the cellular machinery that controls and carries out mRNA synthesis. Transcription factors regulate gene expression by binding to sequences flanking a gene (*cis*-sequences), interacting with each other and with other proteins (e.g. cofactors, chromatin-remodeling enzymes, and general transcription factors) to modulate the rate of transcription initiation at the appropriate promoter. To a large extent, the specific temporal, positional and conditional pattern of expression of each gene is a function (albeit a very complicated one) of the arrangement of transcription factor binding sites in its *cis*-DNA.

Thus, in analyzing the transcription regulatory content of a genome, it is of paramount importance to know the binding specificities of all the organism's transcription factors. Although methods exist to experimentally determine the *in vitro* [1, 2] and *in vivo* [3-5] binding specificities of transcription factors, it is not yet feasible to routinely apply these methods to the hundreds or thousands of transcription factors encoded by most organisms' genomes.

There has, therefore, been considerable focus on methods to deduce the binding specificities of transcription factors in the absence of direct experimental data. In recent years, two largely independent approaches to this problem have emerged. In one approach, structural and biochemical rules are used to predict the binding specificity of a given transcription factor given its amino acid sequence (reviewed in [6]). In a second

approach, statistical models are used to identify from genome sequences and other information those sequences – or more precisely models of related families of sequences- that are likely to be binding sites for some biologically active transcription factor (reviewed in [7]). Surprisingly, although both of these approaches show considerable promise, there have been few efforts to combine their insights into a unified approach to the *de novo* detection and prediction of transcription factor binding sites. Here, we briefly review these two different approaches, point out the ways in which they can usefully be combined, and propose an approach to transcription factor binding site detection that incorporates aspects of both approaches. A series of related papers describe specific implementations and evaluations of this approach.

Modeling and Inference of Transcription Factor Binding Specificities

Following early structural work on protein-DNA complexes, there was considerable optimism that a protein recognition code would be discovered that would allow for the binding specificity of a factor to be directly deduced from its amino acid sequence [8]. However, as more and more structures were determined, it became clear that such a deterministic code does not exist [9], with recent studies highlighting how the detailed complexity and subtle variation of protein-DNA interactions makes such a code impossible to deduce [10].

In recent years, the idea of a deterministic code has been replaced by that of a "probabilistic code", in which the amino acid sequence of a transcription factor – in particular the identity of bases known to interact with DNA in related proteins – is used

to assess the likelihood that a given sequence will be bound by the factor or to design factors likely to bind to a given target sequence [6, 11-17].

An entirely different approach has emerged with the increased availability of genome sequence data. In particular, numerous methods have been developed and applied to infer models of transcription factor binding sites directly from sequences, often in combination with other types of information. For example, a large class of approaches seeks models of transcription factor binding sites (usually in the form of position-weight matrixes [18, 19]) that are enriched in sets of sequences that, based on experimental data, are thought to contain common transcription factor binding sites. Enriched sequences are identified in various ways, the most common based on maximum likelihood estimations of finite mixture models as implemented in MEME [20] or the Gibbs sampler [21]. Many alternate approaches have been introduced, including word counting methods [22-24], probabilistic segmentation or dictionary based approaches [25], and direct modeling of the relationship between sequences and expression data [26, 27].

Although the biological and statistical models, descriptions of binding sites and computational algorithms used vary considerably amongst these methods, they all share the assumption that all motifs are created equal; that any and all motifs have an equal *a priori* probability of being a transcription factor binding site. **Our central argument here is that this assumption is incorrect – that the biophysical and biochemical nature of transcription factor-DNA (TF-DNA) interactions imposes constraints on the types of motifs that are likely to be transcription factor binding sites, and that our structural and biochemical understanding of the interactions between**

transcription factors and DNA can be used to guide *de novo* motif detection methods.

Constraints on Sequence Specificities:

Transcription factors rarely bind exclusively to a single nucleotide sequence. Rather, they usually recognize a family of sequences that share some highly conserved bases as well as some more flexible positions (see Figure 1). These families of sequences are generally described either as consensus sequences (Figure 1B) that specify which base(s) are acceptable at each position or as position-weight matrixes (PWMs; Figure 1C) that describe the probability of observing each base at each position within bound sequences. Because consensus sequences are a special case of PWMs, and because there is solid theory relating PWMs to binding affinities [28, 29], we will limit this discussion to PWMs.

The matrix values of a PWM specify the relative preference of the transcription factor for specific bases at each position. Binding sites (and PWMs) can also be characterized by the overall tolerance of the factor for substitution at each position within the site. A common measure of this substitution tolerance is Shannon information ([30]; Figure 1D). Information (formally $I = 2 - \sum_{B=\{A,C,G,T\}} f_B \log_2 f_B$ where f_B is the frequency of base B [31]) is inversely proportional to substitution tolerance, and can be thought of as a direct measure of the selectivity of the transcription factor at each position, with higher information representing greater selectivity. Positions where only one base is ever observed have little tolerance for base substitutions and therefore contain maximal

information (2.0), while all bases are observed at equal frequency have minimal information (0.0).

Although information is a function only of observed base frequencies in sequences bound by the factor, it is natural to think of information as a measure of the importance of each base in productive transcription factor-DNA interactions as a site's tolerance for substitution should reflect the nature and extent of its contacts with the transcription factor. An important recent paper [32] provides support for this relationship. These authors analyzed five bacterial DNA binding proteins, whose structures bound to DNA had been determined by x-ray crystallography, and computed the number of contacts between each base in the bound DNA and the protein. For each factor they assembled collections of sequences known from experimental data to be bound by the protein, computed PWMs from these sequences, and showed that there is a strong correlation between the number of contacts at a position in the bound sequence and the information content of the corresponding position of the PWM. Bases that are more extensively contacted by the protein are more conserved. We have observed a similar relationship for several yeast transcription factors.

Although this observation that there is relationship between the structural footprint of a protein on DNA and the information profile of the PWM that describes sequences bound by this protein is, in some ways, fairly obvious and has been indirectly described previously [33], **it is surprising that this fundamental characteristic of protein-DNA interactions has not been incorporated into *de novo* motif detection algorithms**. Here, we propose several ways in which this could be accomplished, and in a related set of papers offer specific implementations of these ideas.

Clustering of information within PWMs.

Transcription factors rarely contact a single base without interacting with adjacent bases. For example, many types of transcription factors insert an alpha-helix into the major groove of DNA and make base-specific contacts with 4 or 5 adjacent nucleotides, with the most contacts being made to the central 2 or 3 nucleotides [34]. It follows that the position of high information (and thus also low information) positions should be clustered within PWMs.

Such clustering is observed in transcription factor PWMs based on experimental data. Figure 2 shows that, in PWMs from the transcription factor database TRANSFAC [35], there is a strong correlation between the information at adjacent position (the information content of all pairs of adjacent positions shows a Pearson correlation of 0.57, as compared to an average Pearson correlation of 0.14 for 100 trials where the positions within each matrix were randomly permuted).

As will be discussed below, this common feature of PWMs that represent *bona fide* transcription factor binding sites can be readily incorporated into motif detection algorithms and used to improve the specificity and sensitivity.

Shared information profiles for structurally related transcription factors.

An important corollary of the observation that there is a relationship between the structural footprint of a transcription factor bound to DNA and the information profile of its PWM, is that if we knew (or could predict) the footprint of a transcription factor on

DNA then we would expect the information profile of the PWM describing sequences bound by this factor to match this footprint.

Of course, it is not practical to experimentally determine the structural footprint of every factor in which we are interested. However, it should often be possible to infer the structural footprint – or equivalently the expected information content of the PWM – from those of structurally related transcription factors. An examination of transcription factor-DNA complexes for factors within the same broad structural class, suggests that the structural footprint of TFs on DNA is often reasonably well conserved, even when the amino acid sequence and binding specificity of the factor are not. Therefore, and we can hypothesize that the PWMs for homologous transcription factors should have similar information profiles. To the extent that this is true (a detailed examination of the PWMs in TRANSFAC loosely supports this hypothesis, although the quantity and quality of the data were insufficient to demonstrate it conclusively), this property could have a significant impact on methods to recognize transcription factor binding sites and on our ability to match identified motifs with specific transcription factors.

For example, PWMs describing the binding sites of homeodomain proteins (of the helix-turn-helix family of transcription factors) generally have a core of 4 highly conserved bases flanked on either side by 1 or 2 more partially conserved bases. This is consistent with the structures of homeodomain proteins complexed to DNA, in which an α-helix positioned in the DNA major groove makes extensive contacts with 4 or 5 bases and lesser contacts with a few bases flanking this core on either side. When attempting to construct a PWM describing sites that might be bound by an otherwise uncharacterized

homeodomain protein, it would make sense to begin by looking for motifs with similar information profiles to other homeodomain binding sites.

A more concrete example of where such a strategy could be used is the recent determination of sequences bound *in vivo* by 107 different transcriptional regulators (most of which are DNA binding proteins) of the yeast *Saccharomyces cerevisiae* [36]. The authors of this work attempted to use their data to discover or refine PWMs describing each factor's binding specificity by running the program MEME on each set of bound sequences. In some cases, this approach was successful. However, in a surprising number of cases the results were inaccurate or uninformative.

Ninety of these factors are members of well-characterized families of transcription factors or contain well-characterized DNA binding motifs [37]. We can use the expectation that transcription factors sharing a common DNA binding domain will have corresponding PWMs with similar information profiles to make predictions about the information profiles of the PWMs for most of these ninety factors. As is discussed in the four related papers, this expectation can be built into motif detection algorithms and used to search not simply for enriched motifs (as is done by MEME), but for enriched motifs that have the expected information profile. Our results in applying these methods to the data of [36] will be detailed in a forthcoming publication.

Use of Common Principles in Motif Detection Algorithms

Both the general and specific properties of transcription factor PWMs discussed above can be readily incorporated into standard motif detection strategies. From a statistical/algorithmic point of view, expectations about the information profile of PWMs

can be thought of in two complementary ways. First, they can be thought of as prior knowledge, and implemented as a statistical prior on the space of motifs representing the likelihood that a given motif is a transcription factor binding site or a binding site for a specific family of transcription factor. Most current motif detection algorithms (e.g. MEME, Gibbs sampler) assume a uniform prior - that all PWMs are equally likely to describe a transcription factor binding site regardless of how information is distributed within the PWM. Alternatively, these expectations can be thought of as a constraint on the motifs that are identified by the motif detection algorithm. For example, in searching for homeodomain binding sites we could search only for motifs with an appropriate information profile.

We note that MEME and several other motif detection algorithms already implement one type of structural constraint imposed by specific structural characteristics of a class of transcription factors, namely those that bind DNA as homodimers. In most cases, these factors recognize motifs with an internal 2-fold axis of symmetry (e.g. CGTACG). If it is known that a factor is – or could be – a homodimer, it makes sense to only consider 2-fold symmetric motifs as possible examples of binding sites. MEME, for example, implements this "palindrome" constraint by averaging motifs across a 2-fold, reverse complemented axis of symmetry following the M-step of the EM algorithm.

It is important to note that in no case do the constraints we are discussing place any constraints on the sequence specificity at any position – the constraints only exist at the level of the information profile of the motif. Thus, these methods can be thought of as complementing methods that use amino acid preference rules to predict the base specificity of a factor [15, 17, 29, 38].

We have evaluated several of these methods, pursuing a number of complementary approaches described in four separate papers. Two of these methods [39, 40] use prior distributions to describe a dependence structure between the positions of the PWM, and two others [41, 42] use constraints on the entropy structure of the PWMs.

The approach described in [42] employs a motif model that allows specific ordering of the information of the individual motif positions (e.g. the information in position *i* is greater than that of position *j*, or, more generally, that the information in the motif has one or two peaks) and uses the EM-algorithm to maximize the likelihood of the sequence and model under this constraint. Under the simplifying assumption that at each position *j* of a motif there is one (unknown) preferred residue with (unknown) probability $p_j > ¼$ while the remaining nucleotides have a probability $\frac{1-p_j}{3}$ each, it becomes straightforward to compute the maximum likelihood estimator of a PWM under order restrictions on the information content of its columns and a global maximum can be obtained easily.

[41]employ a general constraint model in which the information profile of a motif is constrained to belong to a user-specified family of information profiles, e.g. motifs with maximal information in a central base and linearly declining information for positions flanking this central base. The method is fairly flexible in allowing for arbitrary parametric families of information profiles. The maximum likelihood PWM fitting the specified constraint is identified using the EM algorithm, where the M-step employs a constrained nonlinear maximization method.

[39]implement the concept of strong, moderate and weak “conservedness” (corresponding to high, intermediate and low information) at given positions of the PWM

through specific priors that do not constrain which specific nucleotides are conserved. Positions in the motif are partitioned into one of the three regimes (strong, moderate, or weak conservation), based on prior knowledge or assumptions about the information profile sought. The likelihood of PWMs deviating from the specified prior are penalized based on the extent of their deviation, with the strength of the penalty under user control. The penalized likelihood is maximized with the EM algorithm in which the M-step is closed form (and thus the optimization is more efficient) for most of the regime types.

[40]considers a more complex prior distribution on the motif PWM. The multinomial probabilities at each position are drawn from a mixture of Dirichlet distributions (each position of the PWM is indexed by a hidden class variable and the prior distribution on the multinomial nucleotide probabilities are drawn according to this class). To enforce the dependence structure among the positions of the motif, the hidden class variables are drawn from a first order markov chain identified by a K by K transition matrix (K is the number of components in the Dirichlet mixture) and marginal distribution of the class variable at the first position. The number of prior classes (the number of components in the Dirichlet mixture) are chosen by the user. The corresponding parameters of the Dirichlet prior distributions, and the transition matrix for the first order markov chain are supplied by the user, with parameters optimally obtained from a set of training motifs.

Future Directions

Here, and in a series of related papers, we have discussed how structural characteristics of transcription factor–DNA interactions constrain the families of

sequences bound by transcription factors, and how these constraints can be used in motif detection. We believe these methods are the basis for a more expansive and productive fields of structure based *de novo* motif detection.

There are clearly many challenges for fully realizing this idea. In particular, there is a need for far more high-quality data on the binding specificities of transcription factors. In attempting to analyze available binding matrixes in TRANSFAC [35], we were struck by how few examples there were of factors whose binding specificities were reasonably comparable, owing largely to extreme heterogeneity in the methods used to experimentally and computationally characterize these affinities. We believe that continued progress in this field is dependent upon the consistent application of high-throughput, high-accuracy measurements of *in vitro* binding specificities [2] of large numbers of transcription factors.

Acknowledgements

This paper is dedicated to my graduate advisor Don C. Wiley (1944-2001), who continues to inspire my work.

I wish to acknowledge the members of my lab, as well as regular attendees of the monthly meetings of the Berkeley gene expression analysis group, where these ideas were refined and developed, including Mark van der Laan, Peter Bickel, Dick Karp, Sandrine Dudoit, Sunduz Keles, Katherina Kechris, Erik van Zwet, Eric Xing, Biao Xing, Katie Pollard, John Storey and Roded Sharan. I thank Derek Chiang and Audrey Gasch for useful comments on the manuscript.

Figure 1. Representations of transcription factor binding sites. A) A hypothetical collection of sequences bound by a transcription factor. **B**) Consensus sequence model of sequences from *A*. The base at each position in the consensus sequence is the base most frequently observed at that position. Where two or three bases are observed at roughly equal frequencies, a redundant IUPAC base is used. **C**) Position-weight matrix (PWM) model of the sequences from *A*. **D**) Information content of PWM from **C**.

Figure 2. Clustering of information in transcription factor binding sites in TRANSFAC. The information content of each position in all transcription factor binding site matrixes in release 5 of TRANSFAC [35] were computed using the standard information equation $I = 2 - \sum_{B=\{A,C,G,T\}} f_B \log_2 f_B$. Positions were binned (n=20) based on their information content, and for all positions in each bin the average information content of adjacent positions was computed and plotted here (red line). The analysis was repeated on randomized data in which the information content of positions within a matrix were randomly permuted. The blue line shows the averaged results of 100 random trials.

REFERENCES

1. Pollock R, Treisman R: A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Res* 1990, 18(21):6197-6204.
2. Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, Bucher P: High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* 2002, 20(8):831-835.
3. Liu XS, Brutlag DL, Liu JS: An algorithm for finding protein DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* 2002, 20(8):835-839.
4. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 2001, 409(6819):533-538.
5. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al*: Genome-wide location and function of DNA binding proteins. *Science* 2000, 290(5500):2306-2309.
6. Benos PV, Lapedes AS, Stormo GD: Is there a code for protein-DNA recognition? Probab(ilistical)ly. *Bioessays* 2002, 24(5):466-475.
7. Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000, 16(1):16-23.
8. Pabo CO, Sauer RT: Protein-DNA recognition. *Annu Rev Biochem* 1984, 53:293-321.
9. Matthews BW: Protein-DNA interaction. No code for recognition. *Nature* 1988, 335(6188):294-295.
10. Pabo CO, Nekludova L: Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J Mol Biol* 2000, 301(3):597-624.
11. 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(23):11168-11172.
12. 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(23):11163-11167.
13. Greisman HA, Pabo CO: A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* 1997, 275(5300):657-661.
14. Wolfe SA, Greisman HA, Ramm EI, Pabo CO: Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J Mol Biol* 1999, 285(5):1917-1934.
15. Suzuki M, Yagi N: DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc Natl Acad Sci U S A* 1994, 91(26):12357-12361.
16. 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.

17. **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(2):370-382.**
18. **Stormo GD, Schneider TD, Gold L, Ehrenfeucht A: Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* 1982, 10(9):2997-3011.**
19. **Stormo GD, Fields DS: Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* 1998, 23(3):109-113.**
20. **Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994, 2:28-36.**
21. **Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993, 262(5131):208-214.**
22. **Pevzner PA, Sze SH: Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol* 2000, 8:269-278.**
23. **van Helden J, Andre B, Collado-Vides J: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 1998, 281(5):827-842.**
24. **Vilo J, Brazma A, Jonassen I, Robinson A, Ukkonen E: Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc Int Conf Intell Syst Mol Biol* 2000, 8:384-394.**
25. **Bussemaker HJ, Li H, Siggia ED: Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A* 2000, 97(18):10096-10100.**
26. **Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using correlation with expression. *Nat Genet* 2001, 27(2):167-171.**
27. **Keles S, van der Laan M, Eisen MB: Identification of regulatory elements using a feature selection method. *Bioinformatics* 2002, 18(9):1167-1175.**
28. **Berg OG, von Hippel PH: Selection of DNA binding sites by regulatory proteins. Statistical- mechanical theory and application to operators and promoters. *J Mol Biol* 1987, 193(4):723-750.**
29. **Benos PV, Lapedes AS, Fields DS, Stormo GD: SAMIE: statistical algorithm for modeling interaction energies. *Pac Symp Biocomput* 2001:115-126.**
30. **Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986, 188(3):415-431.**
31. **Shannon CE: A Mathematical Theory of Communication. *Bell Syst Tech J* 1948, 27:379-423,623-656.**
32. **Mirny LA, Gelfand MS: Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res* 2002, 30(7):1704-1711.**
33. **Suzuki M, Brenner SE, Gerstein M, Yagi N: DNA recognition code of transcription factors. *Protein Eng* 1995, 8(4):319-328.**
34. **Luscombe NM, Austin SE, Berman HM, Thornton JM: An overview of the structures of protein-DNA complexes. *Genome Biol* 2000, 1(1):REVIEWS001.**

35. **Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F: TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 2000, 28(1):316-319.**
36. **Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al*: Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002, 298(5594):799-804.**
37. **Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al*: The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* 2001, 29(1):37-40.**
38. **Benos PV, Lapedes AS, Stormo GD: Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol* 2002, 323(4):701-727.**
39. **Kechris KJ, van Zwet E, Bickel PJ, Eisen MB: Detecting DNA regulatory motifs by incorporating positional trends in information content. *Genome Biol* 2004, 5(7):R50.**
40. **Xing EP, Wu W, Jordan MI, Karp RM: LOGOS: A modular Bayesian model for de novo motif detection. In: *IEEE Computer Society Bioinformatics Conference, CSB2003: 2003*; 2003.**
41. **Keles S, van der Laan M, Dudoit S, Xing B, Eisen MB: Supervised Detection of Regulatory Motifs in DNA Sequences. *Statistical Applications in Genetics and Molecular Biology* 2002, 3(1):1-40.**
42. **van Zwet E, Kechris K, Bickel P, Eisen MB: Estimating motifs under order restriction. *Statistical Applications in Genetics and Molecular Biology* 2005, 4(1):1-18.**

A

ACGCATCACGAA
CAACATCATGAC
ATCGCTCATGCG
TAGGATCACTCT
GTCCATCTTGGG
AGCCATCATATA
CGAGATCACATC
GGAGATCACTGT
TCGCATCATTGG
TTGCCTCTTTAA
CAAGATCACATC
GCCGATCACACT

B

NNVSATCAKDNN

C

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.25	0.25	0.33	0.00	0.83	0.00	0.00	0.83	0.00	0.00	0.25	0.25
C	0.25	0.25	0.33	0.50	0.17	0.00	1.00	0.00	0.50	0.33	0.25	0.25
G	0.25	0.25	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.33	0.25	0.25
T	0.25	0.25	0.33	0.00	0.00	1.00	0.00	0.17	0.50	0.33	0.25	0.25

D

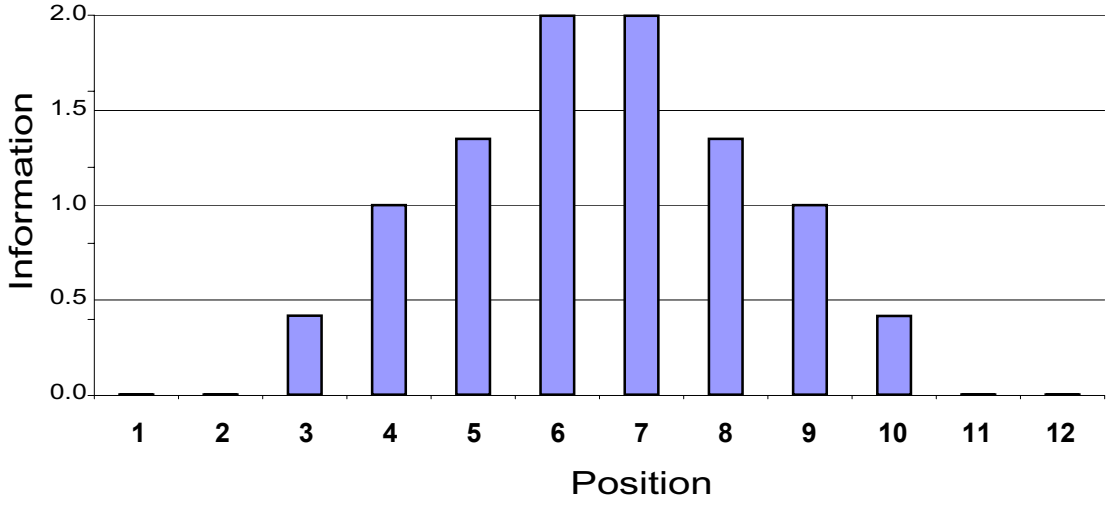

Figure 1

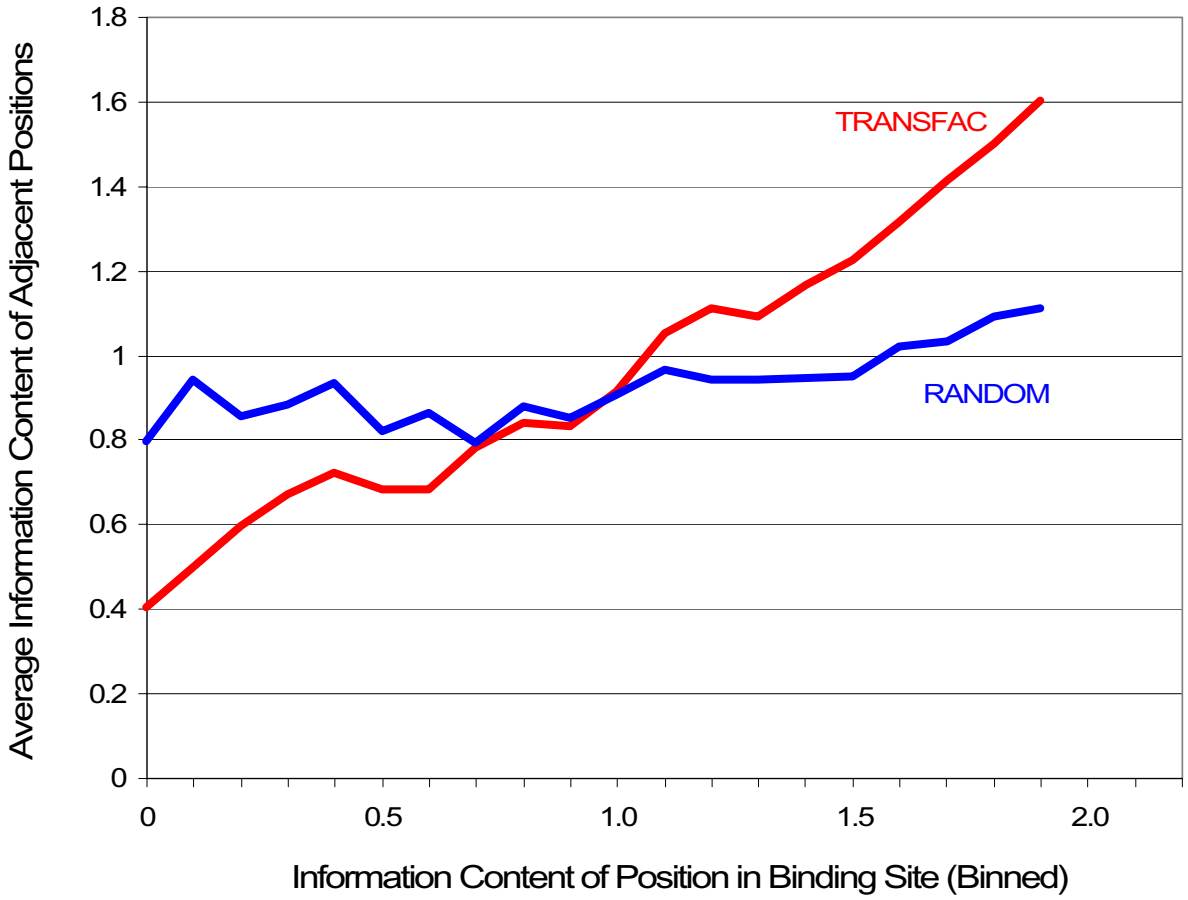
Average Information Content of Adjacent Positions
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
TRANSFAC
RANDOM
0
0.5
1.0
1.5
2.0
Information Content of Position in Binding Site (Binned)

Figure 2