Pacific Symposium on Biocomputing 13:477-488(2008)

AN ANALYSIS OF INFORMATION CONTENT PRESENT IN PROTEIN-DNA INTERACTIONS

CHRIS KAUFFMAN AND GEORGE KARYPIS*

Department of Computer Science, University of Minnesota
117 Pleasant St. SE
Minneapolis, MN 55455, USA
E-mail: {kauffman,karypis}@cs.umn.edu

Understanding the role proteins play in regulating DNA replication is essential to forming a complete picture of how the genome manifests itself. In this work, we examine the feasibility of predicting the residues of a protein essential to binding by analyzing protein-DNA interactions from an information theoretic perspective. Through the lens of mutual information, we explore which properties of protein sequence and structure are most useful in determining binding residues with a particular focus on sequence features. We find that the quantity of information carried in most features is small with respect to DNA-contacting residues, the bulk being provided by sequence features along with a select few structural features. Supplemental information for this article is available at `http://www.cs.umn.edu/~kauffman/supplements/psb2008`

1. Introduction

Complex behaviors of the genome are now beginning to be understood in terms of feedback network models in which regulatory elements promote or inhibit transcription of genes and are themselves affected by the transcription of other elements. Key to this system are interactions between DNA, the main storage unit for genetic information, and proteins, which are both products and managers of transcription. To that end, a plethora of computational methods have been presented to predict which proteins will bind to DNA[1,15], what parts of a protein will bind to DNA[2,11,17,18], and which segments of DNA a protein will favor for binding. These methods have yet to reach a performance plateau and researchers continue to apply machine learning and statistical techniques in an attempt reach the

*Work supported by the NIH Training for Future Biotechnology Development grant, NIH T32GM008347

Pacific Symposium on Biocomputing 13:477-488(2008)

highest accuracy and sensitivity supported by available information.

We endeavor in this study to provide some insight into the inherent difficulty of predicting protein-DNA interactions. From a thermodynamic perspective, the interactions have been found to be quite sensitive. Binding is marginally favored when considering the whole complex[7]. This leaves very little in the way of individual contributions for each residue requiring methods that predict binding residues to make shrewd use of any available features to achieve accuracy. Predicting binding residues would benefit genome studies as mutating them to less favorable analogues gives a mechanism to affect a protein's role in the system. In particular, prediction of binding residues from sequence alone is desirable as it would open the door to a wide variety of experiments involving transcription regulatory elements which have not been co-crystallized with DNA and for which CHIP-Chip experiments[10] are not feasible.

In this paper we focus on sequence and structure features of single protein residues and how they may describe a residue's contributions to the DNA-binding event. We lay out an information theoretic framework in which to conduct the study, illustrate the features of interest, and report the most likely candidates for use in prediction methods.

2. Methods and Materials

2.1. *Mutual Information (MI)*

The main tool we employ for analysis is mutual information (MI)[5,14]. The MI between two random variables is a measure of how easily the value of one may be predicted given the other's value. That is, mutual information measures how much information two variables carry about one another. In the discrete case, it is defined for random variables X and Y as

$$I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where x and y are the discrete values or classes which random variables X and Y can take on and $p(x,y)$ is the probability of x and y occurring together. Due to the base-two logarithm, mutual information in this paper is reported in bits.

2.2. *Features*

In our setting, each residue of a protein has associated with it features that are represented by random variables. The first feature considered is always whether the residue is DNA-contacting or not, a binary feature, while the

Pacific Symposium on Biocomputing 13:477-488(2008)

second feature is varied. The MI between the DNA-contacting feature and other features gives us an idea of how informative these other features will be for predicting binding residues.

The features we consider are described in Table 1 and include sequence and structure properties. Only a few of them have a natural discrete definition (such as the 20 amino acids). Solvent accessible surface area (SASA) and information per position (IPP), both single continuous values, were discretized by choosing boundaries to divide the values into bins. These boundaries were chosen by a grid search so that the resulting class definitions maximized mutual information with the DNA contacting classes. Residues were assigned as either DNA contacting or non-contacting based on distance cutoffs which were varied by 0.25 angstroms. The SASA and IPP class boundaries were varied in increments of 0.01 and boundaries that achieved high MI across several DNA-contacting cutoffs were further considered. The values selected for these boundaries are shown in the rightmost column of Table 1.

In order to discretize the remaining vector-valued features we employed clustering techniques. The toolkit CLUTO[9], version 2.1.2, was used with default options to create various numbers of clusters. Each cluster is then one of the discrete values this feature takes on when calculating mutual information. Some experimentation was done using similarity measures other than the default cosine measure, but none yielded a significant change.

A sensible prediction method will employ a variety of features to decide whether a residue contacts DNA. To partially address this, we explore joint features, combinations of two single features, whose values represent every possible combination of the values of the single features. The size of the joint feature is the product of the sizes of the two single features, e.g., amino acids may take on 20 values, secondary structure 3 values, and their joint feature may take on 60 values.

As it is central to the whole study, the definition of DNA binding and non-binding residues is treated with special attention. Distances are calculated between each atom of a residue in a protein and each atom in the DNA structures of each data file. The minimum distance of these is taken as the residue-DNA distance. When computing mutual information, the cutoff distance is varied in increments of 0.2 Å which defines the DNA contacting and non-contacting residues. This allows us to plot a curve for each feature showing characteristics of the signal separating contacting and non-contacting residues. If any combination of feature values does not occur, mutual information becomes undefined. This frequently happens at low

Pacific Symposium on Biocomputing 13:477-488(2008)

Table 1. Residue Features Considered for Mutual Information with DNA-contacting classes.

Feature	**Description**	**Discrete Values**
Amino Acid	Amino acid type of the residue	20 values
Positive, Negative, Neutral Amino Acids	The 20 amino acids divided into 3 classes for their charge. Divisions taken from Cline et al.[4]	Pos: Arg, Lys His Neg: Asp, Glu Neu: All others
Profiles	Combination of the position specific scoring matrix (PSSM) and position specific frequency matrix (PSFM) generated from 3-iterations of PSI-BLAST[3] against the NCBI NR sequence database.	5, 10, and 20 clusters
Concatenated Profiles	A sliding window of size 5 around each residue was used to concatenate the full profiles of adjacent residues. End residues without enough sequence neighbors were assigned 0 in each column of the profile for a missing residue.	5, 10, and 20 clusters
PSSMs	Only the PSSM from the PSI-BLAST profile.	5, 10, 20 clusters
Concatenated PSSMs	Only the PSSMs of residues within a sliding window of size 5 concatenated together.	5, 10, and 20 clusters
Information Per Position (IPP)	The second to last column in PSI-BLAST profiles, gives an account of the sequence diversity in a column of the profile. Low values indicate a strong preference for certain amino acids in that column.	2-value: 0.0-0.62, >0.62 3-value: 0.0-0.48, 0.48-1.0, >1.0 4-value: 0.0-0.48, 0.48-0.81, 0.81-1.27, >1.27
Solvent Accessible Surface Area (SASA)	Surface area of a residue accessible to solvent (water) molecules, normalized based on the maximum SASA of a residue in Gly-X-Gly. Calculated using DSSP[8] and normalized using the values of Miller et al.[13].	2-value: 0.0-0.09, >0.09 3-value: 0.0-0.09, 0.09-0.20, >0.20 4-value: 0.0-0.01, 0.01-0.07, 0.07-0.20, >0.20
Structural Neighbors	Sum of amino acid types within a 14 Å sphere and with sequence distance $\geq$3; distance is between alpha carbons.	5, 10, and 20 clusters
Structural Neighbor PSSMs	Sum of the PSSMs of structural neighbors.	5, 10, and 20 clusters
Secondary Structure	The secondary structure assigned to a residue, by DSSP and mapped into 3 values for helix, strand, and coil	3 values, DSSP letters H,G,I are helix, E is strand, and all others are coil
Physical Quantities	Features of Wang and Brown[17] which are pK_a, a measure of the acidity of side-chains (7 for neutral side-chains), hydropathy according to the scale of Kyte and Doolittle[12], and molecular mass. A sliding window of size 11 around each residue was used to create features which were then used in clustering.	5, 10, and 20 clusters

and high distance cutoff values, especially for features which take on many values. In the plots shown subsequently, undefined MI is set artificially to 0.

Pacific Symposium on Biocomputing 13:477-488(2008)

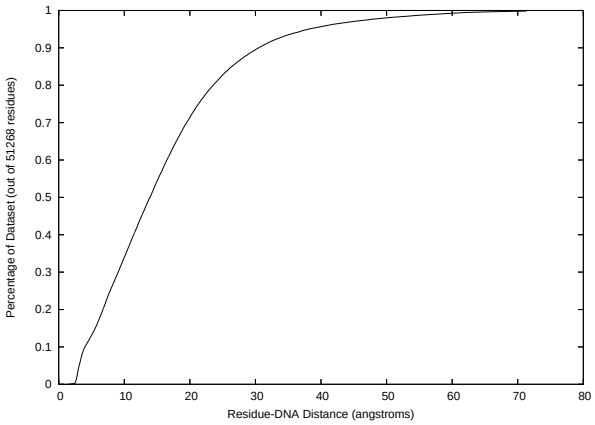


Figure 1. Percentage of Contacting Residues vs. Distance Cutoff.

2.3. *Data Sets*

The data that we employ is derived from that used by Tjong and Zhou[15] with further culling. Beginning with their 264 PDB files, we separated each into protein chains according to the PDB chain identifier. Within protein-DNA co-crystal PDB files, there may exist several chains with identical sequence. This type of duplication may cause an unfair bias in calculating mutual information so the chains were submitted to the PISCES server[16] to be culled to less than 30% sequence identity. The remaining data set comprises 246 chains from 218 different PDB files and includes 51268 residues. Figure 2.3 illustrates the percentage of residues classified as DNA-contacting according to a sliding distance cutoff. The full list of PDB chains used and their associated data is available in the online supplement.

2.4. *Corrections for Small Sample Size*

Calculations of mutual information must be done with care as they may yield an artificially high estimate particularly with small sample sizes. Two approaches taken in the literature to overcome this have been to use bootstrap sampling[6] and to calculate the excess mutual information over a random shuffling of the data[4]. We employ the latter method on single features by leaving the DNA-contacting classes fixed and randomly permuting the values of the second feature. This shuffling preserves the background probabilities of each value of the feature. Calculating mutual information with

Pacific Symposium on Biocomputing 13:477-488(2008)

these shuffled values gives an idea of what MI we can expect to get at random for the background probabilities and number of values for the feature. We compute the average MI over 200 permutations of each feature. Subtracting this quantity led to only a slight drop in MI, about 1% for single features in the worst case. Based on this, we report raw MIs for the rest of the paper.

Joint features pose a problem as they are likely to be more inflated due to the large number of values they take on. We find this difficult to correct as random permutation of class values often leads to zero probability of some combinations and an undefined MI. We report raw values for joint classes here and will attempt to estimate the bias in future works through sampling methods.

3. Results

3.1. *Single Features*

None of the features we explore yield a large magnitude of mutual information with the DNA-binding feature. The most informative features are on the order of hundredths of bits for both single and joint features. This is the same order of magnitude at which previous works have shown contact potentials[4] and aspects of sequence-structure correlations[6] to reside.

For features discretized via clustering, an increased number of clusters leads to an increase in mutual information. In order to give a basis of comparison to the largest natural set of values, amino acids with 20 discrete values, we consider 5, 10, and 20 clusters per feature.

Table 2 summarizes the calculated values for single features while Figure 3.1 illustrates how mutual information for some of the features alters as the distance cutoff defining DNA-contacting residues is altered. The single features yielding the most information on contact vs. non-contact residues are entirely sequence based. Amino acid sequence alone yields a maximum of 0.029 bits at a distance cutoff of 3.37 Å. This is modestly exceeded by PSSMs with 20 clusters (0.032 bits at 4.97 Å cutoff) and profiles (0.032 bits at 4.97 Å cutoff) and is succeeded in information by 10 clusters of profiles (0.027 bits at 4.77 Å cutoff). Using a sliding window of PSSMs or profiles did not improve mutual information: 20 clusters generated using a sliding window of 5 full profiles gives a maximum of 0.020 bits at 5.77 Å while using only the PSSM in clustering yields 0.016 bits at 5.17 Å. Dividing the 20 amino acids into three classes for positive, negative, and neutral residues significantly reduces the information content to a maximum 0.016 bits at 3.57 Å.

Pacific Symposium on Biocomputing 13:477-488(2008)

Table 2. Mutual Information of Single Features. The mutual information is with the DNA-contacting/non-contacting class (binary) and the distance cutoff is at the maximum MI achieved by the feature. The table is sorted by MI. The column N_{val} is the number of discrete values the feature may take.

Feature	N_{val}	MI	Dist. Cutoff
PSSMs	20	3.1933e-02	4.97
Profiles	20	3.1856e-02	4.97
Amino Acids	20	2.9465e-02	3.37
Profiles	10	2.6765e-02	4.77
Struct. neighbor PSSMs	20	2.6379e-02	10.17
PSSMs	10	2.4402e-02	4.97
Struct neighbors	20	2.2810e-02	8.57
Concat. profiles	20	2.0252e-02	5.77
Struct. neighbor PSSMs	10	1.9237e-02	9.57
PSSMs	5	1.8971e-02	4.97
Struct neighbors	10	1.8597e-02	7.17
Concat. PSSMs	20	1.6257e-02	5.17
Pos/Neg/Neut Amino Acids	3	1.5879e-02	3.57
Solv. Acc. Surf. Area	20	1.5125e-02	3.97
Concat. PSSMs	10	1.4767e-02	4.97
Struct. neighbors	5	1.4166e-02	6.97
Solv. Acc. Surf. Area	4	1.4060e-02	3.77
Concat. profiles	10	1.3289e-02	5.17
Solv. Acc. Surf. Area	2	1.2471e-02	3.97
Info per position	4	1.1519e-02	9.57
Profiles	5	1.1500e-02	3.57
Concat. PSSMs	5	1.1399e-02	4.97
Struct. neighbor PSSMs	5	1.1114e-02	9.57
Info per position	3	1.0934e-02	9.57
Concat. profiles	5	1.0788e-02	5.17
Info per position	2	9.4190e-03	13.97
pK_a/hydropathy/mass	20	3.0624e-03	5.17
pK_a/hydropathy/mass	10	2.7191e-03	5.17
Secondary structure	3	2.4700e-03	5.77
pK_a/hydropathy/mass	5	2.1319e-03	7.17

The lowest information content for single features came from secondary structure assignment (max of 0.002 bits at 5.77 Å) and clusters formed from the combination of pK_a, hydropathy, and molecular mass in sliding window of 11 residues (20 clusters, max of 0.003 bits at 5.17 Å).

3.2. *Joint Features*

The large number of combinations prevents a full discussion of joint features. For brevity, we mention a few interesting cases and include the full numerical results in the online supplement. These cases are summarized in Table 3 and Figure 3. Unsurprisingly, combinations of the most informa-

Pacific Symposium on Biocomputing 13:477-488(2008)

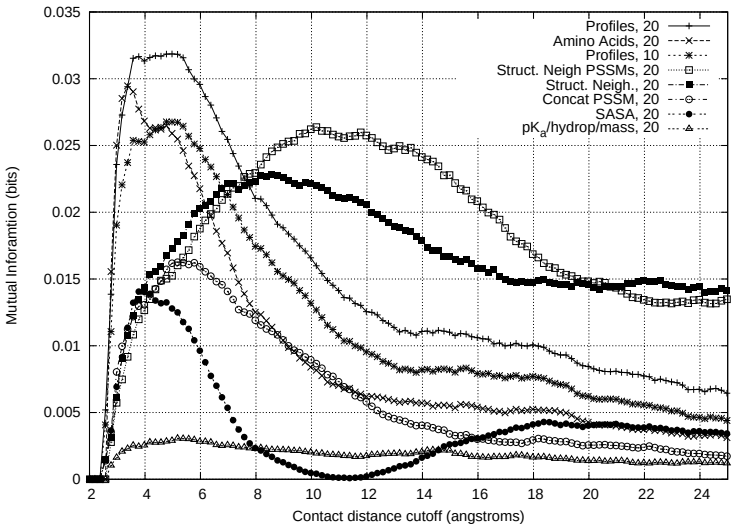
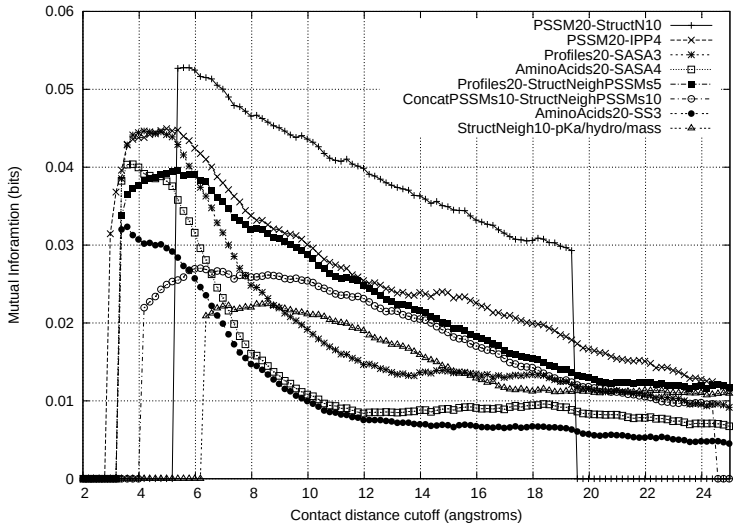


Figure 2. Single Features: Distance Cutoff for DNA-contacting residues versus Mutual Information. The cutoff distance which defines DNA-contacting versus non-contacting residues is varied by small increments to show the character of some single features and their mutual information with the DNA-contacting classes.

Table 3. Selected Mutual Information of Joint Features. The mutual information is with the DNA–contacting/not-contacting class (binary) and the distance cutoff is at the maximum MI achieved by the joint features. N_{val1} and N_{val2} are the number of discrete values features 1 and 2 may take on respectively while N_{tot} is their product, the number of discrete values the joint feature may take.

Feature 1	N_{val1}	Feature 2	N_{val2}	N_{tot}	MI	Dist. Cutoff
PSSMs	20	Struct neighbors	20	400	5.2781e-02	5.77
PSSMs	20	Struct neighbors	10	200	4.7563e-02	5.37
Profiles	20	Struct neighbors	10	200	4.6912e-02	6.57
Profiles	10	Struct neighbors	20	200	4.5558e-02	5.97
PSSMs	20	Info. per pos.	4	80	4.4948e-02	4.97
Profiles	20	SASA	3	60	4.4649e-02	4.17
Amino Acids	20	SASA	4	80	4.0379e-02	3.77
PSSMs	20	SASA	4	80	4.3894e-02	3.97
Amino Acids	20	Info. per position	4	80	4.2580e-02	3.57
Profiles	10	Info. per position	4	40	4.2397e-02	5.37
Profiles	20	Struct. neigh. PSSMs	5	100	3.9513e-02	5.37
Profiles	20	Second. Struct.	3	60	3.6432e-02	4.97
Concat. PSSMs	10	Struct. neigh. PSSMs	20	200	3.3650e-02	6.97
Amino Acids	20	Second. struct.	3	60	3.2341e-02	3.57
Struct neighbors	20	pK$_a$/hydropathy/mass	5	100	2.3224e-02	13.77

Figure 3. Joint Features: Distance Cutoff for DNA-contacting residues versus Mutual Information. The cutoff distance which defines DNA-contacting versus non-contacting residues is varied by small increments to show the character of some joint features and their mutual information with the DNA-contacting classes.



tive single features lead to the highest MIs, the best pairs being PSSMs or profiles with structural neighbors (first rows of Table 3). The next major combination that proved fruitful was between PSSMs, profiles, or sequence with SASA. Combining information per position with sequence or profiles provides the next highest mutual information followed by combinations of profiles or sequence with the PSSMs of structural neighbors. The lower quality single features result mostly in low joint MI, profiles with secondary structure being one exception.

4. Discussion

Most significant among the results are the contributions of sequence based features. Utilizing PSSMs, full profiles, or even simply sequence yields the most information about the differences between residues with high propensities for contacting DNA. It is well known that the negatively charged phosphate backbone of DNA prefers proximity to residues which have a positive charge such as arginine and lysine rather than neutral or positive alternatives. However, limiting the division of amino acids to simply positive, negative, and neutral types severely diminishes MI, giving only 0.016

Pacific Symposium on Biocomputing 13:477-488(2008)

bits versus 0.029 bits for all amino acid. Counter to intuition, the use of a sliding window with concatenated profiles does not increase MI over the single profile column. The reasons for this are unclear and are worth investigating further. Information per position, when combined with a PSSM, provides a surprisingly informative joint feature. The two together likely amplify the conservation signal present in many DNA contacting residues. With the majority of the information present coming from sequence sources, we can begin to understand why the performance of sequence-based methods such as Ahmad and Sarai[2] have produced prediction results that are nearly as good as those incorporating structure features.

The poor mutual information given by structural features such as SASA and secondary structure class may seem surprising as it is expected that most DNA-contacting residues at least have a high SASA and probably prefer a helix (a common binding motif is helix-turn-helix). However, considering that there are many surface residues with high SASA which do not contact DNA and that helices are a very common secondary structure element, these features are quite noisy. Combining profile information with SASA improves MI significantly, underscoring their reinforcement of one another.

Structural features which do carry information appear to come in the form of the local environment, i.e., descriptions of other residues proximal in space. This is evidenced by the relatively high MI of the structural neighbor feature. Information of this sort is used in a number of DNA-protein prediction methods[1,11,15] and seems to improve performance though not spectacularly. From the standpoint of sequence only predictions, these properties would need to be predicted in order to be used for DNA-contact predictions. Based on the fact they carry a moderate amount of information, there may be some hope that using predicted values would yield improvement.

The physical features of pK_a, hydropathy, and molecular mass did not yield much information and were uniformly lowest both on their own and in combinations. Wang and Brown report quite promising results using support vector machines with only these features[17] indicating that the clustering method used to discretize the feature may not be appropriate. We will explore alternatives in the future to verify that a signal is indeed present in these features as they are some of the easiest to utilize in the protein-DNA interaction prediction.

The literature pertaining to binding residue prediction has defined the binding class using cutoffs in the range of 3.5-5.0 Å. The ideal cutoff dis-

Pacific Symposium on Biocomputing 13:477-488(2008)

tances for both single and joint features seem to support this definition with preference towards the higher end.

5. Conclusion

Armed with the knowledge that signals pertaining to DNA proximity are weak but present, we can understand why prediction methods have enjoyed only marginal success thus far. Incorporating additional features that have not, as of yet, been explored may be the only way to boost performance. From the structure standpoint, this likely involves more complicated geometric information about residues or the consideration of multiple residues interacting with DNA simultaneously. This direction precludes DNA-binding protein with no available structure information. Including features of the DNA being contacted might be the only route as yet unexplored for sequence-only features. Training prediction methods with the knowledge that residues with specific characteristics favor a specific DNA sequence may lead to visible improvements. Approaching the problem from this side will also allow us to incorporate knowledge generated by DNA-binding motif studies.

As for an immediate extension of the present work, we plan to expand the study to account for several shortcomings. Previously mentioned is the issue of properly estimating bias in mutual information for the case of joint features with many values. Sampling techniques and additional compute time are likely to provide the remedy. Also, we have not yet incorporated truly non-contacting residues, only those that are in a DNA-binding protein but far from the interaction site. Adding proteins known not to bind to DNA, especially if they bind to something else such as a small molecule or another protein, will solve this problem and give a better assessment of those characteristics separating DNA-contacting residues from general interaction sites. Finally, the techniques applied here need not be limited to DNA but can also be applied to RNA interactions with proteins.

References

1. Shandar Ahmad, M. Michael Gromiha, and Akinori Sarai. Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 20(4):477–486, Mar 2004.
2. Shandar Ahmad and Akinori Sarai. Pssm-based prediction of dna binding sites in proteins. *BMC Bioinformatics*, 6:33, 2005.
3. SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.*, 25(17):3389–3402, 1997.

Pacific Symposium on Biocomputing 13:477-488(2008)

4. Melissa S Cline, Kevin Karplus, Richard H Lathrop, Temple F Smith, Robert G Rogers, and David Haussler. Information-theoretic dissection of pairwise contact potentials. *Proteins*, 49(1):7–14, Oct 2002.
5. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2006.
6. Gavin E. Crooks, Jason Wolfe, and Steven E. Brenner. Measurements of protein sequence-structure correlations. *Proteins: Structure, Function, and Bioinformatics*, 57:804–810, 2004.
7. B. Jayaram, K. McConnell, S. B. Dixit, A. Das, and D. L. Beveridge. Free-energy component analysis of 40 protein-dna complexes: a consensus view on the thermodynamics of binding at the molecular level. *J Comput Chem*, 23(1):1–14, Jan 2002.
8. Wolfgang Kabsch and Chris Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–637, 1983.
9. George Karypis. Cluto: A clustering toolkit. Online at `http://www.cs.umn.edu/~karypis/cluto`, 2007.
10. Tae Hoon Kim and Bing Ren. Genome-wide analysis of protein-dna interactions. *Annu Rev Genomics Hum Genet*, 7:81–102, 2006.
11. Igor B. Kuznetsov, Zhenkun Gou, Run Li, and Seungwoo Hwang. Using evolutionary and structural information to predict dna-binding sites on dna-binding proteins. *Proteins*, 64:19–27, 2006.
12. Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105–132, May 1982.
13. Susan Miller, Joel Janin, Arthur M. Lesk, and Cyrus Chothia. Interior and surface of monomeric proteins. *Journal of Molecular Biology*, 196:641–656, Aug 1987.
14. C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:pp. 379–423 and 623–656, 1948.
15. Harianto Tjong and Huan-Xiang Zhou. Displar: an accurate method for predicting dna-binding sites on protein surfaces. *Nucl. Acids Res.*, 35(5):1465–1477, 2007.
16. Guoli Wang and Jr Dunbrack, Roland L. Pisces: recent improvements to a pdb sequence culling server. *Nucl. Acids Res.*, 33:W94–98, 2005.
17. Liangjiang Wang and Susan J Brown. Bindn: a web-based tool for efficient prediction of dna and rna binding sites in amino acid sequences. *Nucleic Acids Res*, 34(Web Server issue):W243–W248, Jul 2006.
18. Changhui Yan, Michael Terribilini, Feihong Wu, Robert L Jernigan, Drena Dobbs, and Vasant Honavar. Predicting dna-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, 7:262, 2006.