BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i297–i304
doi:10.1093/bioinformatics/btm215

Inferring protein–DNA dependencies using motif alignments and mutual information

Shaun Mahony[1,*], Philip E. Auron[2,3] and Panayiotis V. Benos[1,4,5,*]

[1]Department of Computational Biology, [2]Department of Molecular Genetics and Biochemistry, School of Medicine, University of Pittsburgh, [3]Department of Biological Sciences, Duquesne University, [4]Department of Human Genetics, Graduate School of Public Health and [5]University of Pittsburgh Cancer Institute, School of Medicine, University of Pittsburgh, Pittsburgh, USA

ABSTRACT

Motivation: Mutual information can be used to explore covarying positions in biological sequences. In the past, it has been successfully used to infer RNA secondary structure conformations from multiple sequence alignments. In this study, we show that the same principles allow the discovery of transcription factor amino acids that are coevolving with nucleotides in their DNA-binding targets.

Results: Given an alignment of transcription factor binding domains, and a separate alignment of their DNA target motifs, we demonstrate that mutually covarying base-amino acid positions may indicate possible protein–DNA contacts. Examples explored in this study include C2H2 zinc finger, homeodomain and bHLH DNA-binding motif families, where a number of known base-amino acid contacting positions are identified. Mutual information analyses may aid the prediction of base-amino acid contacting pairs for particular transcription factor families, thereby yielding structural insights from sequence information alone. Such inference of protein–DNA contacting positions may guide future experimental studies of DNA recognition.

Contact: shaun.mahony@ccbb.pitt.edu or benos@pitt.edu

1 INTRODUCTION

Transcription factor (*TF*) proteins recognize their DNA targets via the formation of a network of specific and non-specific molecular interactions. TF DNA-binding preferences are usually modeled using frequency matrices derived from alignments of known sites. Typically, these *position-specific scoring matrices* (*PSSMs*) assume independence between the base positions (Stormo, 2000). Structurally related TFs often share similarities in their DNA-binding motifs. Generalized binding models or *familial binding profiles* (*FBPs*) constitute a measure of the 'average' binding specificity for a family of TFs (Sandelin and Wasserman, 2004). Structural information and protein sequence comparisons have been previously used to cluster TF binding profiles in order to build FBPs (Sandelin and Wasserman, 2004), and automatic methods have been recently introduced (Mahony *et al*., 2007). FBPs allow DNA pattern discovery algorithms to be *biased* towards a particular TF structural class (Mahony *et al*., 2005). In addition, FBPs can be used to infer the identity of the TF family for predicted novel motifs (Mahony *et al*., 2007; Sandelin and Wasserman, 2004), or to remove degeneracy between related motifs in the motif repositories (Cartharius *et al*., 2005).

In this study, we use FBP construction methods to define alignments of related DNA-binding motifs. Given an alignment of DNA-binding motifs from a family of related TFs, and a separate alignment of their corresponding DNA-binding domain sequences, we demonstrate that mutual information can be calculated for each pair of positions between the alignments. Positions of high covariance are shown to correspond to TF residues that have a critical effect on DNA recognition. We demonstrate the effectiveness of this method using C2H2 zinc finger, homeodomain and basic helix-loop-helix (bHLH) binding domain DNA motifs, where known protein–DNA contacting positions are recovered using sequence information alone. The prediction of nucleotide–amino acid contacting potential from sequence data alone is invaluable in directing mutagenic experimentation for elucidating mechanisms of TF-DNA recognition. As demonstrated in this article, mutual information analyses can certainly play a role in such predictions.

2 METHODS

2.1 Comparing PSSM columns

A PSSM model of length L is comprised of a set of $4 \times L$ weights (columns). Each column, $\boldsymbol{X}$, follows a probability distribution, $\{p_X(b)\}_{b\in\{A,C,G,T\}}$, with the base probability values reflecting the binding preference of the TF to the corresponding base in this position. The probability values can be estimated from the observed base counts, $\{n_X(b)\}_{b\in\{A,C,G,T\}}$. We denote the estimated values $f(X) = \{f_X(b)\}_{b\in\{A,C,G,T\}}$. In practice, p_X are estimated from n_X plus some pseudocounts to reduce small sample biases and to avoid zero probabilities. The assumption of independence between positions is not entirely accurate, but acts as a useful approximation (Benos *et al*., 2002a).

The *Pearson Correlation Coefficient* (*PCC*) has been previously used by us and others to compare DNA motif columns (Benos *et al*., 2002a; Hughes *et al*., 2000; Mahony *et al*., 2005), and gives a measure of agreement between two (unweighted) sets of observations by means of their covariance. PCC is defined by:

$$PCC(X,Y) = \frac{\sum_{b=A}^{T}\left(f_X(b)-\overline{f_X}\right)\cdot\left(f_Y(b)-\overline{f_Y}\right)}{\sqrt{\sum_{b=A}^{T}\left(f_X(b)-\overline{f_X}\right)^2\cdot\sum_{b=A}^{T}\left(f_Y(b)-\overline{f_Y}\right)^2}} \quad (1)$$

*To whom correspondence should be addressed.

We have recently found the PCC metric to have superior DNA motif alignment performance over alternatives (Mahony *et al.*, 2007).

2.2 Comparing motifs of different lengths: *P*-values

A dataset of 10 000 simulated PSSMs reflecting the properties of the PSSM models in the JASPAR database was constructed as described in the following web site: http://forkhead2.cgb.ki.se/jaspar/additional/index.htm. Sandelin and Wasserman's method (Sandelin and Wasserman, 2004) was then used for the calculation of empirical *P*-values that are independent of the length of the compared motifs. In this method, the alignment scores observed between all possible pairings of the simulated PSSMs are grouped according to the lengths of the paired matrices. Probability distributions specific to pairs of matrices of any given length are thus constructed and allow calculation of the probability that an observed similarity score is no better than that of a pair of random PSSMs of the same lengths.

2.3 Pairwise and multiple motif alignment and tree-building methods

An ungapped, extended Smith–Waterman local alignment strategy (Smith and Waterman, 1981) is used in this study, where the 'motif cores' of the PSSM models under comparison are aligned before extending the local alignment. The 'core' is defined as the longest of (a) the four most informative adjacent columns and (b) the 'trimmed' motif (starting and ending at a position with information content at least 0.3). Optimal alignment is sought in both forward/reverse motif directions.

Iterative refinement is used as the multiple alignment strategy, and aims to combat the common problem of local minima due to 'frozen' subalignments (Barton and Sternberg, 1987). Iterative refinement builds a rough multiple alignment by progressively adding to the current alignment the most similar input PSSM. Once the initial alignment is built, each PSSM is removed from the alignment in turn and realigned to a profile of the other aligned sequences. Iteration of the realignment continues a fixed number of times.

The trees constructed for the homeodomain and basic region examples are built using a UPGMA algorithm, where the distances between motifs are derived from the similarity *P*-values. All pairwise alignment, multiple alignment and tree-building algorithms employed in this study are accessible from the STAMP web-platform (http://www.benoslab.pitt.edu/stamp).

2.4 Mutual information

Mutual information (i.e. covariance dependency) has long been used as an aid to RNA secondary structure prediction, allowing the detection of pairs of codependant columns in an alignment of RNA sequences (Chiu and Kolodziejczak, 1991; Gutell *et al.*, 1992). In this study, we demonstrate that mutual information analysis of DNA motif multiple alignments may assist in the prediction of protein positions that affect DNA binding at particular base positions. The mutual information, M_{ij}, between a DNA motif multiple alignment column and a protein alignment column is defined as:

$$M_{ij} = \sum_{ib=A}^{T} \sum_{ja=A}^{Y} f_{ib,ja} \cdot \log_2 \frac{f_{ib,ja}}{f_{ib} \cdot f_{ja}}, \quad (2)$$

where f_{ib} is the observed frequency of base b ($b \in \{A,C,G,T\}$) in column i of the DNA alignment, f_{ja} is the frequency of amino acid a ($a \in \{A,C,D,\ldots,Y\}$) in column j of the protein alignment and $f_{ib,ja}$ is the joint (pairwise) frequency of this base-amino acid position combination. A multiple alignment of related DNA-binding motifs may be constructed using the methods described above. Given a multiple alignment of the corresponding DNA-contacting domain protein sequences, the mutual information between positions in the proteins and their DNA targets may be calculated. The protein positions that exhibit high mutual information for one or more base positions are more likely to be involved in the binding mechanism; either by directly contacting the corresponding bases or indirectly, e.g. by stabilizing a 'core' of contacting amino acids.

2.5 Limitations of mutual information analysis

Low mutual information values should be treated with caution. Low scores suggest that the corresponding base and amino acid positions show no codependence only if both positions are varying independently. Naturally, *covariance* cannot be used to measure anything useful if one or both positions are *invariant*. These cases should be treated as 'missing values' rather than 'no co-dependence'. On the other hand, high mutual information values may indicate covariance only if both positions have sufficient examples to provide statistical significance. For example, we may easily imagine the extreme scenario where four aligned protein sequences contain different amino acids in a particular position. This position will show 'high' mutual information value if the four amino acids happen to pair with different nucleotides. In such a case, however, the 'co-variance' would be entirely coincidental. We ideally want the number of observed pairs (x) to be high, as larger numbers of examples will allow us to distinguish between true and coincidental covariance.

We can use simulations to measure the extent to which coincidental covariance could occur. To do this, 10 000 sets of x random base/amino acid pairs were generated, and mutual information scores were calculated for each set. For varying x, the average proportion of the random sets that produce a mutual information score of less than 0.5 (an arbitrary low threshold) is displayed in Figure 1. As may be seen from the figure, 140 base/amino acid pairs are required before the chance of randomly receiving a mutual information score greater than 0.5 falls below 1%. In the EGR zinc finger example discussed below, $x = 3099$ for the coalesced set after separating each of the three zinc fingers and their DNA target, so this set obviously passes the significance threshold.

Note that the above simulations and associated significance threshold of 140 pairs are applicable only to those cases where single amino acids are paired with single bases. In the general case, where the target

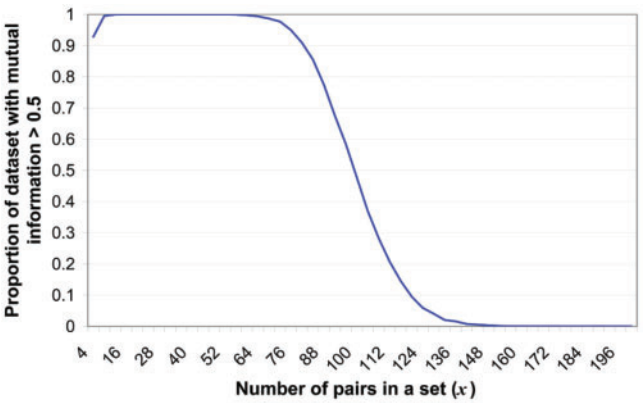

Fig. 1. Proportion of mutual information scores ≥0.5 for random base/amino acid pairs.

motif PSSM columns may represent many aligned bases, the minimum number of observed pairs required to reach statistical significance decreases. Simulations where the observed nucleotide–amino acid pairings are randomly shuffled may be used to assign *P*-values to observed mutual information scores. In the text below, all discussed peaks of mutual information are significant ($P \leq 0.0001$) according to such random shuffling simulations.

2.6 Structural analysis

Molecular structural inspection was carried out using RasMac version 2.7.3 (Sayle and Milner-White, 1995). Coordinate data was obtained from the RCSB Protein Data bank (www.rcsb.org). Analysis included examination of the C2H2 zinc fingers for *EGR1* (1AAY), and the homeodomains for *Engrailed* (1HDD), *Antennapedia* (9ANT) and *MATa1* (1YRN). For the basic helix-loop-helix (bHLH) example, protein–DNA complex structures were examined for *pho4* (1A0A) and *myoD* (1MDY).

3 RESULTS

3.1 Cys_2His_2 (C2H2) zinc finger proteins

The predictive efficiency of the mutual information measure was tested on the C2H2 TF family, the most abundant TF-DNA binding domain found in the *Pfam* motif database today [*Pfam* v.19.0 (Bateman *et al*., 2004)]. A previously published dataset of protein–DNA interaction examples for EGR1 (a member of the C2H2 family) and its mutants (Benos *et al*., 2002b) was used as the basis for this test. The dataset contains 1033 pairs of EGR1-derived protein sequence and corresponding bound DNA sequence as determined by phage display and SELEX experiments.

EGR1 proteins contain three zinc fingers, and co-crystal structures show that amino acids at positions −1 and either +3 or +6 of the helices contact one DNA base each (major groove) in an anti-parallel fashion (Elrod-Erickson *et al*., 1996; Pavletich and Pabo, 1991) (Fig. 2). In addition, the amino acid at position +2 can contact DNA position 4 in the opposite strand (overlapping base). In order to assess the general binding properties of an individual zinc-finger α-helix, the fingers in each of the protein–DNA binding examples are coalesced. Comparison of the aligned set of zinc-finger sequences against their preferred DNA target motifs showed strong peaks of mutual information in the expected protein–DNA 'contacting' positions as well as some additional ones (Fig. 3). For example, amino acid position +2 covaries with the third base position but not with the expected fourth base position. It is known, however, that aspartic acid at position +2 in EGR1 zinc fingers interacts with and helps orient the arginine at position −1, which contacts the third base (Elrod-Erickson *et al*., 1996; Pavletich and Pabo, 1991). Thus, replacing aspartic acid will influence the binding on the third base. Importantly, the glutamate at position +3 covaries with the second base position. This is in strong agreement with the crystal structure, in which helical position +3 makes an H-bond contact to the second base (Elrod-Erickson *et al*., 1996). Interestingly, position +3 also appears to covary somewhat with the first base position, possibly reflecting van der Waals contacts that are also observed in the crystal structure.

3.2 Homeodomain proteins

The coevolution of homeodomain proteins with their DNA-binding motifs was also analyzed. DNA-binding motifs of 25 homeodomain proteins [representing over 729 documented TF binding sites taken from TRANSFAC (Matys *et al*., 2003) and JASPAR (Sandelin *et al*., 2004)] were aligned using ungapped Smith–Waterman alignment with the PCC metric and iterative refinement multiple alignment. The choice of motifs was dependent on their displaying some similarity to the classical 'ATTA' homeobox target motif. Figure 4 presents the alignment of the motifs and their corresponding phylogenetic tree. Note that in this figure the tree branch lengths are based on the distances of the DNA motifs, not the proteins. The protein alignment of the homeodomain members was obtained directly from the *Pfam* database [*Pfam* v.19.0 (Bateman *et al*., 2004), Accession: PF00046].

Mutual information was calculated from the DNA motifs and protein multiple alignments, and is presented (color coded) in Figure 5 for 12 amino acids contained within the homeodomain recognition helix. The amino acids and bases are numbered in Figure 5 to correspond with the convention for the *Engrailed* homeodomain TF-DNA complex in the Protein Data Bank (PDB:1HDD). One of the highly mutually informative sites is the pairing of position 50 in the protein domain with position 8 in the motif multiple alignment. The importance of this pair of positions has been long recognized and experimentally confirmed (Treisman *et al*., 1989), and provides a textbook example of how changing a single amino acid can drastically affect the DNA-binding specificity of the homeodomain proteins (Latchman, 2004). There are two other known

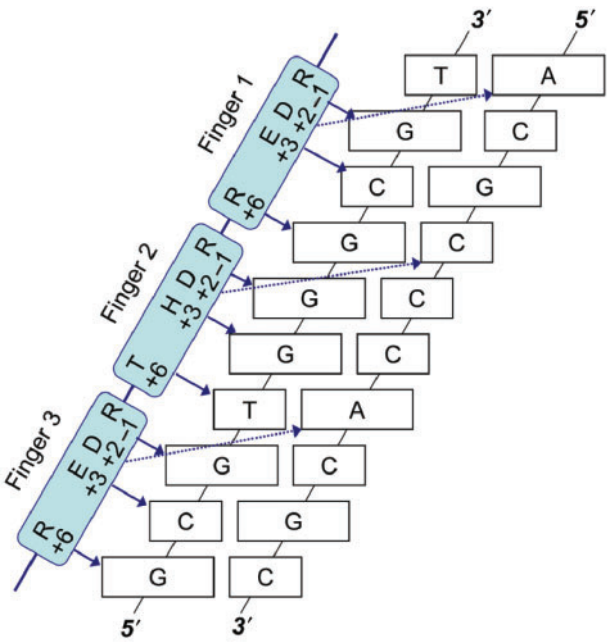

Fig. 2. A model of the mode of DNA recognition for EGR-family zinc fingers.

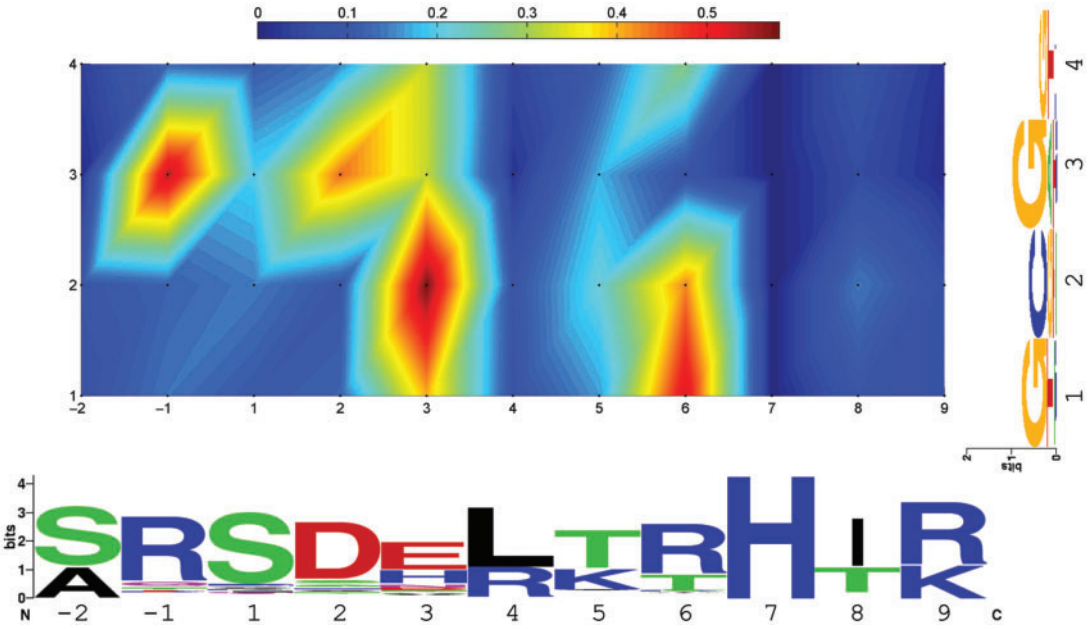

Fig. 3. Mutual information between EGR1-derived mutant finger domains and their targets as derived from *in vitro* selection experiments.

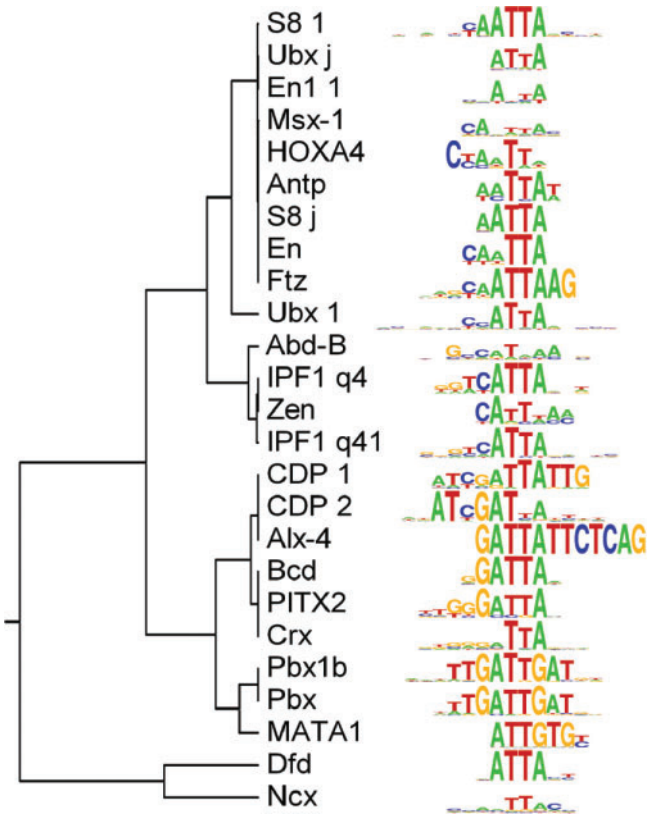

Fig. 4. The aligned DNA motif targets for the homeodomain proteins. The phylogenetic tree represents relative distances between DNA motifs.

DNA-contacting residues (positions 47 and 51) that the mutual information plot misses. For position 51, the ultraconserved (invariant) amino acid residue prohibits any covariance analysis.

The mutual information plot also suggests that other positions in the protein domain can influence the binding specificity in and around the core 'ATTA' homeobox target motif. These include amino acid positions 46 and 54, which, like position 50, usually project away from the homeodomain protein core and toward the DNA major groove. Position 46 is, for some homeodomains, in contact with position 50. Examples include the polar interaction that occurs in the *Engrailed* homeodomain (PDB:1HDD) (Kissinger *et al*., 1990) and the close contact for these residues in the *Antennapedia* homeodomain (PDB:9ANT) (Fraenkel and Pabo, 1998). The potential for position 46 to affect the conformation of position 50 may explain the similar observed covariance patterns. Also, position 54 reveals a related, but somewhat distinct, covariance pattern. This residue can be found either in van der Waals contact or close proximity to backbone sugar atoms at positions 7 and 8 in the motif multiple alignment for the *Antennapedia* complex and at position 13 for *MATa1* (PDB:1YRN) (Li *et al*., 1995), consistent with the observed covariance pattern (Fig. 6).

3.3 bHLH and bHLH-ZIP proteins

Basic helix-loop-helix regions are used by various TF families to mediate DNA binding. Most of these TFs form homo- or hetero-dimers, and many recognize target sites of the form 5′-CANNTG-3′, also known as the E-box. For example, many bHLH and bHLH-ZIP TFs (such as *pho4*, *myc* and *max*)

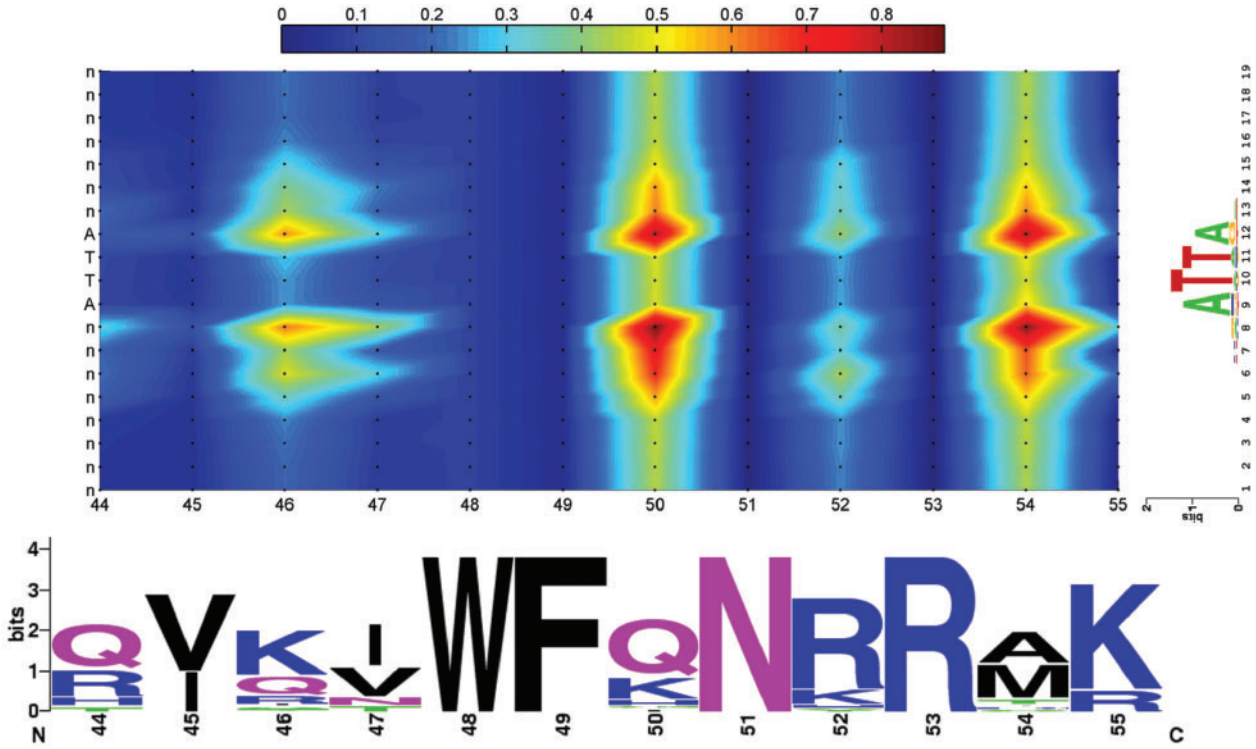

Fig. 5. Mutual information between recognition helices from the homeodomain protein family and their corresponding DNA-binding motifs.

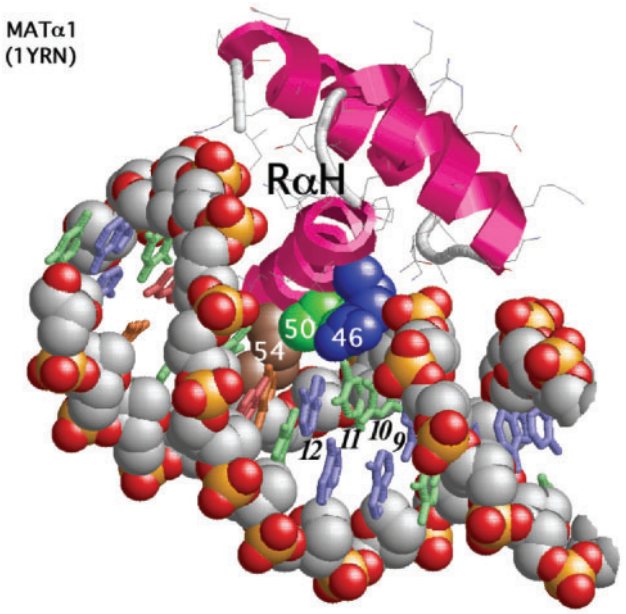

Fig. 6. Representative homeodomain structure reveals the relative location of amino acid positions 46, 50 and 54 with respect to the ATTA target motif. In this example, the *MATa1* homeodomain protein (ribbons and thin sticks) is shown with the DNA recognition alpha helix (RαH) positioned within the DNA major groove. Amino acid side chains 46, 50 and 54 are displayed as space-filling atoms (numbered), as are the DNA backbone atoms (gray sugar and red/orange phosphates). The DNA bases are displayed as thick sticks and numbered (italics) as presented in Figure 5.

recognize target sites of the form 5′-CACGTG-3′. In contrast, some bHLH TFs (including myogenic bHLH TFs, such as *myoD* and *myogenin*) preferably bind to sites of the form 5′-CAGGTG-3′.

A number of published protein–DNA complex structures have illustrated the binding mechanism for various representative bHLH TFs. For example, positions 3L, 2L, 2R′ and 3R′ at the edges of the CANNTG target are contacted using direct recognition by His5 and Glu9 in *pho4* (Shimizu *et al*., 1997) and *myoD* (Ma *et al*., 1994). The position corresponding to Glu9 is ultra conserved throughout all bHLH proteins. Subclasses of bHLH TFs differ in their recognition mechanism of the central 2 bp. Arg13 was shown to directly contact the central 2 bp in *pho4* (Shimizu *et al*., 1997), but in *myoD* the contact is water mediated (Ma *et al*., 1994). In addition, an asymmetrical contacting pattern in the central 2 bp was observed in *E47*, a CAGGTG-binding TF (Ellenberger *et al*., 1994).

Mutual information is used here to explore other potential protein sequence variations that underlie the distinct binding preferences in the central binding position (1R′). The bHLH protein domain alignment was downloaded from Pfam (Accession: PF00010). A set of 24 DNA-binding motifs (representing over 528 documented TF binding sites) that bind to CACGTG or CAGGTG motifs were extracted from JASPAR and TRANSFAC and aligned as before. Only homodimer binding motifs were included. The motifs are aligned as shown in Figure 7. Note how the tree distinguishes between the two subclasses of DNA-binding motif. The mutual information plot based on this motif multiple alignment and

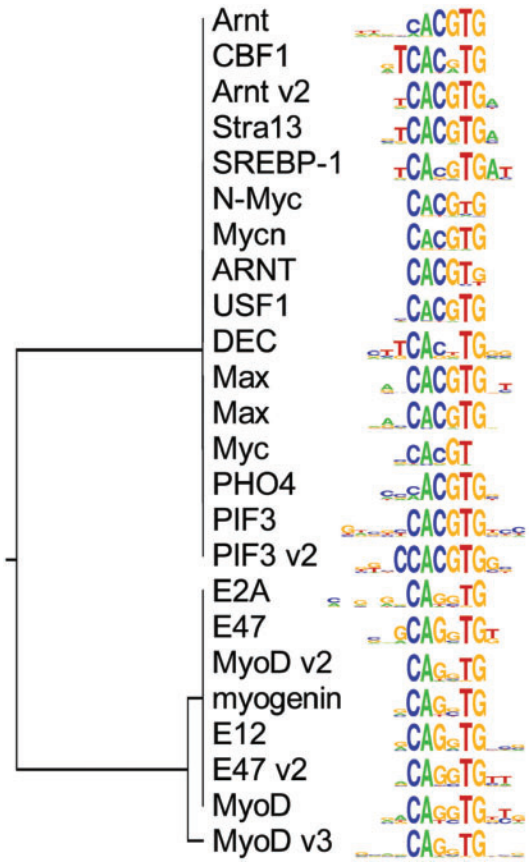

Fig. 7. Iterative refinement multiple alignment and UPGMA tree constructed from the bHLH DNA-binding motifs.

13 amino acids within the basic binding region is shown in Figure 8. The amino acids and bases are numbered in Figure 8 according to the convention for the *pho4* bHLH TF-DNA complex in the Protein Data Bank (PDB:1A0A) (Shimizu *et al.*, 1997).

Since the DNA-binding motifs are relatively similar in all positions other than 1R′, we should not expect significant mutual information peaks outside of this position. Peaks of mutual information with the variant base (1R′) appear at amino acid positions 8, 13 and 14. As mentioned above, position 13 is known to contact 1R′ for those TFs that bind the CACGTG motif. Blackwell *et al.* demonstrated that mutating *myoD*'s positions 8 and 13 (Arg8 & Leu13) to their corresponding *c-myc* residues (Leu8, Arg13) was sufficient to switch *myoD*'s CAGGTG-binding preference to CACGTG (Blackwell *et al.*, 1993). The role of residue 8 is not well known; no solved structure shows contacts between position 8 and the central 2 bp. However, all CAGGTG-binding TFs share an arginine at position 8, while other bHLH TFs typically possess hydrophobic residues (e.g. *max*, a CACGTG-binding TF, has a leucine at this position). Mutating *myoD*'s position 8 arginine to a leucine knocks out DNA binding (Van Antwerp *et al.*, 1992). Therefore, while the arginine at position 8 may not be involved in directly contacting 1R′, the residue is clearly a distinguishing feature of CAGGTG-binding TFs, and is correctly identified as such by the mutual information analysis.

4 DISCUSSION

While mutual information is widely used to find covariant pairs of positions within single molecules (e.g. in alignments of RNA sequences), this study demonstrates that mutual

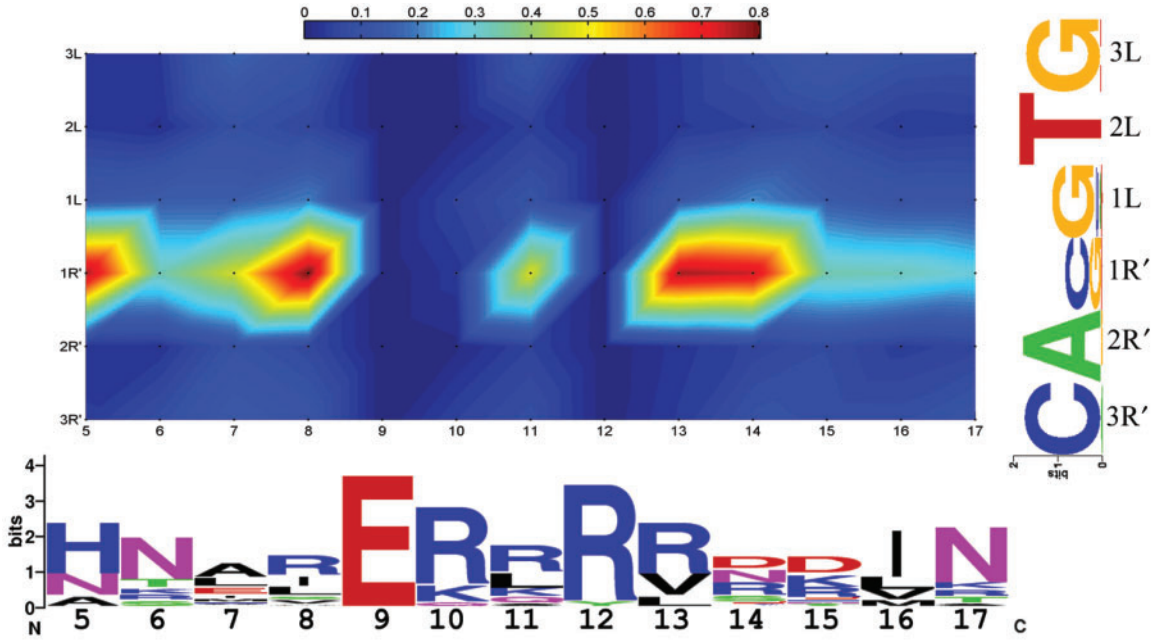

Fig. 8. Mutual information between the recognition helices from selected basic region-containing TFs and their corresponding DNA-binding motifs.

information may also find dependence between two interacting biomolecules. We have shown that appropriately aligned DNA motifs of related TFs can be used to predict the amino acid positions that critically affect DNA-binding preference (directly or indirectly). This is expected to enhance the field of protein engineering. Currently, in the absence of a protein structure, a number of time-consuming protein–DNA binding experiments (such as SELEX on wild type proteins and their mutants), are required for the identification of DNA-contacting positions and their subsequent mutation towards a desired DNA specificity. Mutual information analysis may help guide such experiments. The above examples on three TF families demonstrate that key protein positions for DNA binding can be identified using mutual information plots. In all three examples, known DNA-contacting residues were shown to share a high degree of mutual information with their contacted base. Changes in other non-contacting, but mutually informative, residues may induce structural conformations and may therefore have an indirect effect on DNA-binding preference.

Perhaps more fundamentally, mutual information analysis can yield useful insights into the evolutionary history of a TF family's mode of DNA recognition. For example, by analyzing the representatives of a number of diverse bHLH TF subfamilies, we are able to find residues that distinguish the general binding preference of one subfamily from that of others.

As noted in the Methods section, the accuracy of mutual information plots is critically dependent on the number of available examples. If this number is low, high mutual information scores can occur by chance. Given that only a small fraction of known TFs have corresponding DNA-binding models stored in the databases, the issue of insufficient data may make mutual information analysis challenging for many TF families. One possible approach to partially alleviating the issue of high covariance scores occurring by chance is via reduced amino acid alphabets. For example, if amino acids are grouped according to their characteristics, the number of parameters that need to be estimated in mutual information analysis would also be reduced, thus reducing the number of training examples required to reach statistically significant conclusions. We note that the use of mutual information cannot yield insight into protein–DNA interactions at invariant protein or DNA positions (e.g. position 51 of the homeodomain proteins; Fig. 5), although this is a lesser problem for practical purposes, since invariant positions are generally the first targets in mutation experiments.

It should also be noted that C2H2 zinc-finger domains possess a greater flexibility for variation than homeodomain or bHLH recognition helices, resulting in an apparently higher potential for modification in the absence of structural perturbation (Pabo *et al*., 2001). In general, the flexible nature of C2H2 zinc-finger DNA binding makes this family more amenable to mutual information analysis than most other classes of TFs. TF families such as homeodomain and bHLH TFs possess structures that are specifically tuned to recognize a limited number of similar sequences. Since amino acid substitutions in such TFs may result in complete loss of DNA binding (rather than a change in specificity), invariant positions in the DNA-binding motif alignments are more likely, thus reducing the effectiveness of mutual information analysis. However, as demonstrated in this study, mutual information may still provide useful structural insights when subtle changes in DNA-binding preference are observed. For example, a high order of intradomain structural cooperativity for homeodomain recognition helices may contribute to the similar covariance profiles observed for positions 46, 50 and 54 in Figure 5.

The results presented here suggest that mutual information plots can become an important tool for guiding protein–DNA association studies as the databases of TF binding matrices become larger. In the interim, the structural significance of mutually informative residues will have to be further explored through examination of appropriate protein–DNA structures and by mutation experiments.

ACKNOWLEDGEMENTS

The authors would like to thank Carlos Camacho for useful discussion and help with the crystal structures, and an anonymous reviewer for a number of useful insights and suggestions. This work was supported by NSF grant MCB0316255 and by NIH-NIAID contract no. N01-AI50018. P.V.B. was also supported by NIH grants 1R01LM007994-01 and RR014214 and by TATRC/DoD USAMRAA Prime Award W81XWH-05-2-0066. P.E.A. was supported by NIH grant CA06668544.

Conflict of Interest: none declared.

REFERENCES

Barton,G.J. and Sternberg,M.J. (1987) A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, **198**, 327–3337.

Bateman,A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Benos,P.V. *et al.* (2002a) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.

Benos,P.V. *et al.* (2002b) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.

Blackwell,T.K. *et al.* (1993) Binding of myc proteins to canonical and noncanonical DNA sequences. *Mol. Cell Biol.*, **13**, 5216–5224.

Cartharius,K. *et al.* (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.

Chiu,D.K. and Kolodziejczak,T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.*, **7**, 347–352.

Ellenberger,T. *et al.* (1994) Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimmer. *Genes Dev.*, **8**, 970–980.

Elrod-Erickson,M. *et al.* (1996) Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure*, **4**, 1171–1180.

Fraenkel,E. and Pabo,C.O. (1998) Comparison of X-ray and NMR structures for the Antennapedia homeodomain-DNA complex. *Nat. Struct. Biol.*, **5**, 692–697.

Gutell,R.R. *et al.* (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–5795.

Hughes,J.D. *et al.* (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.

Kissinger,C.R. *et al.* (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. *Cell*, **63**, 579–590.

Latchman,D.S. (2004) *Eukaryotic Transcription Factors*. Elsevier Academic Press, London.

Li,T. *et al*. (1995) Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. *Science*, **270**, 262–269.

Ma,P.C. *et al*. (1994) Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell*, **77**, 451–459.

Mahony,S. *et al*. (2005) Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, **21** (Suppl. 1), i283–i291.

Mahony,S. *et al*. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Pabo,C.O. *et al*. (2001) Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.*, **70**, 313–340.

Pavletich,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809–817.

Sandelin,A. *et al*. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.

Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.

Shimizu,T. *et al*. (1997) Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J.*, **16**, 4689–4697.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Treisman,J. *et al*. (1989) A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell*, **59**, 553–562.

Van Antwerp,M.E. *et al*. (1992) A point mutation in the MyoD basic domain imparts c-Myc-like properties. *Proc. Natl Acad. Sci. USA*, **89**, 9010–9014.