*Published online 4 June 2010* *Nucleic Acids Research, 2010, Vol. 38, No. 14* ***e150***
*doi:10.1093/nar/gkq478*

# A structural-alphabet-based strategy for finding structural motifs across protein families

**Chih Yuan Wu[1], Yao Chi Chen[2] and Carmay Lim[1,2,*]**

[1]Department of Chemistry, National Tsing Hua University, Hsinchu 300 and [2]Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan

Received February 12, 2010; Revised April 2, 2010; Accepted May 12, 2010

## ABSTRACT

**Proteins with insignificant sequence and overall structure similarity may still share locally conserved contiguous structural segments; i.e. structural/3D motifs. Most methods for finding 3D motifs require a known motif to search for other similar structures or functionally/structurally crucial residues. Here, without requiring a query motif or essential residues, a fully automated method for discovering 3D motifs of various sizes across protein families with different folds based on a 16-letter structural alphabet is presented. It was applied to structurally non-redundant proteins bound to DNA, RNA, obligate/non-obligate proteins as well as free DNA-binding proteins (DBPs) and proteins with known structures but unknown function. Its usefulness was illustrated by analyzing the 3D motifs found in DBPs. A non-specific motif was found with a 'corner' architecture that confers a stable scaffold and enables diverse interactions, making it suitable for binding not only DNA but also RNA and proteins. Furthermore, DNA-specific motifs present 'only' in DBPs were discovered. The motifs found can provide useful guidelines in detecting binding sites and computational protein redesign.**

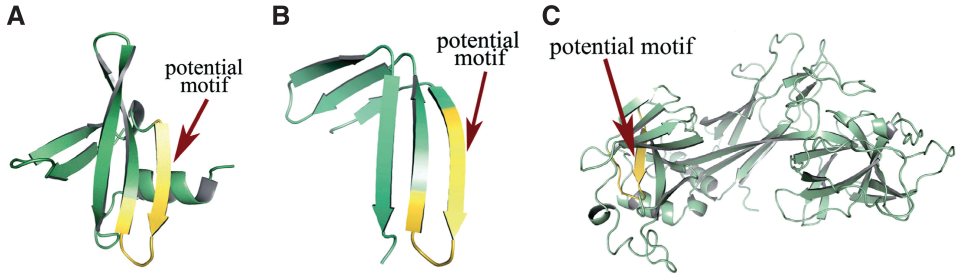
## INTRODUCTION

Sequence motifs can help to quickly relate a novel protein sequence to a known protein family (1) and to identify its plausible function. They usually include conserved essential residues involved in catalysis, in ligand binding or in maintaining a specific conformation. Hence, they can be detected by searching homologous protein sequences for the occurrence of invariant or highly conserved residues. Proteins with a sequence motif comprising functionally/structurally crucial residues share not only sequence similarity, but also structural similarity (2,3). In such cases, a structural motif, comprising of conserved 3D conformations of amino acid residues that are crucial for the protein's fold, stability and/or function, can be associated with the sequence motif.

For proteins that possess insignificant sequence and overall structural similarity, structural/3D motifs as opposed to sequence motifs may be present. 3D motifs have been used to suggest the function of proteins whose structures are known on the basis that similarity in the local structure implies similarity in biological function (4,5); e.g. the helix-turn-helix (HTH) motif has been used to predict proteins with DNA-binding function (6,7). In general, 3D motifs can be constructed manually or automatically using various methods. They have been constructed using conservation of sequence and structural features and compiled in the MegaMotifBase (8). Most methods; e.g. SPASM (9), HTHquery (10), PAR-3D (11), Superimpose (12) and RASMOT-3D PRO (13), require experimental data such as a known 3D motif to search for other similar structures or known active-site or binding-site residues (14).

Our key aim is to present a strategy for automatically discovering 3D motifs 'across' protein families based on a 16-letter structural alphabet (15) without requiring a known template motif or essential residues. A 3D motif is defined herein as a locally conserved contiguous structural segment recurring in $\geq$3 non-redundant proteins sharing <30% sequence identity (Figures 1 and 2). Another aim is to illustrate the usefulness of this strategy by applying it to discover 3D motifs in DNA-binding proteins (DBPs). DBPs were chosen for analysis because the HTH motif, which has been found in 16 DBP families, can be used for method validation and comparison with the motifs found. To evaluate the specificity of the 3D motifs found in DBPs, their occurrence frequencies in non-redundant non-DBPs were computed. Our strategy for finding 3D motifs can yield two types of functional motifs: (i) non-specific 3D motifs found in both DBPs and non-DBPs and (ii) DNA-specific motifs found

*To whom correspondence should be addressed. Tel: 886 2 2652 3031; Fax: 886 2 2788 7641; Email: carmay@gate.sinica.edu.tw



**KKVWRVGKMI**
**2.40.50.40**
Chromosomal protein Sso7d (1c8c-A)

**DKVTRTKARW**
**2.30.18.10**
Transcription initiation factor IIA (1nh2-C)

**NLKVWFGATP**
**2.60.40.10**
Universal transcriptional effector of Notch signaling, CSL (1ttu-A)

**Figure 1.** Recurring structural patterns in structurally non-redundant DBPs. The 3D motif *cdehja* (in yellow) and corresponding amino acid sequence and CATH code in (**A**) the chromosomal protein Sso7d (1c8c-A); (**B**) transcription initiation factor IIA (1nh2-C) and (**C**) the universal transcriptional effector of Notch signaling, CSL (1ttu-A).

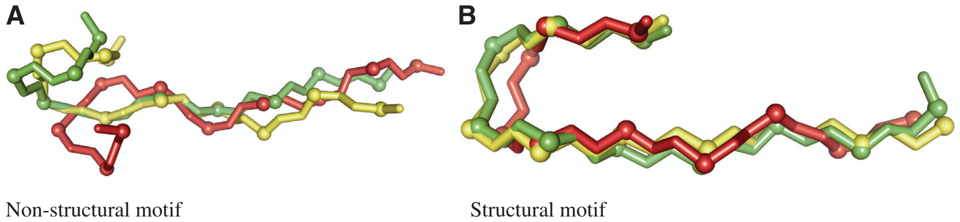


**Figure 2.** Defining 3D motifs. (**A**) The 6-mer structural pattern *ddehjl* is found in 1ecr-A:57–66 (green), 1xgn-A:763–772 (yellow) and 2bcq-A:410–419 (red); since their backbone structures do not superimpose, *ddehjl* is *not* considered to be a 3D motif. (**B**) The 6-mer structural pattern *cdehja* is found in 1c8c-A:21–30 (green), 1nh2-C:249–258 (yellow) and 1ttu-A:589–598 (red); as their backbone structures superimpose, *cdehja* is considered to be a 3D motif.

'only' in DBPs. The 3D motifs found can provide useful guidelines in detecting DNA-binding sites in proteins and redesigning DBPs to improve their DNA-binding affinity/specificity.

## MATERIALS AND METHODS

### Datasets of non-redundant DBPs and non-DBPs

Four datasets were created by searching the Protein Data Bank (PDB) (16) and the PPI-Pred server (17) for $\leq$3-Å X-ray structures of proteins bound to dsDNA/RNA and obligate/non-obligate protein including antigens (18). These DNA/RNA/obligate protein/non-obligate protein-binding chains were then grouped according to their CATH (19) codes, and the complex structure with the best resolution in each group was chosen. This yielded 76 DNA-binding, 72 RNA-binding, 88 obligate protein-binding and 77 non-obligate protein-binding non-redundant proteins (20) (Supplementary Table S1).

### Definition of DNA/RNA/protein-binding residues

For each complex structure, the HBPLUS (21) program was used to compute all possible protein–DNA/RNA/protein van der Waals (vdW) contacts and H bonds, which are defined, respectively, by a distance of 4.0 and 3.5 Å between a donor atom and an acceptor atom. An amino acid residue was assigned as binding if its atoms are in vdW contact or are H-bonded directly/indirectly via water molecules with DNA/RNA/protein atoms, and its solvent accessible surface area (SASA) in the free protein is non-zero.

### Converting a 3D protein structure to a 1D structural letter sequence

Each protein structure was encoded into its 1D structural sequence according to the structural alphabet (15), which was derived as follows: the backbone of each protein from a non-redundant protein structure database was represented by consecutive 5-residue segments, each described by a vector of 8 backbone dihedral angles $\boldsymbol{V}(\psi_{n-2},\phi_{n-1},\psi_{n-1},\phi_n,\psi_n,\phi_{n+1},\psi_{n+1},\phi_{n+2})$. The dissimilarity between two vectors $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ of dihedral angles was measured by the root-mean-square deviation (RMSD) of the dihedral angle values, defined as:

$$\mathrm{RMSD}a(\mathrm{V}_1,\mathrm{V}_2) = \sqrt{\frac{\sum_{i=1}^{4}[\psi_i(V_1)-\psi_i(V_2)]^2+[\phi_{i+1}(V_1)-\phi_{i+1}(V_2)]^2}{8}} \qquad \mathbf{(1)}$$

Using an unsupervised cluster analyzer based on the above RMSDa of the segments, 16 protein blocks/letters were identified and illustrated in previous work (15). These 16 letters comprise the structural alphabet.

The 3D protein structures were converted into strings of structural letters using the PDB reader program (15), as follows. For a given *l*-residue protein, $l-4$ letter assignments were obtained by scanning the sequence using a 5-residue sliding window. The structure of each 5-residue segment was compared with that of each of the 16 letters and the letter that had the closest structure (as measured by the RMSDa) to the 5-residue segment was assigned to the middle residue of that segment (14). The terminal four residues of each protein, which cannot be treated as center residues of the 5-residue segments, were assigned the letter Z.

### Definition of a recurring structural pattern

To discover locally conserved structural segments in the protein, the structural letter sequence of each protein was scanned using an *l*-letter sliding window yielding various *l*-mer structural patterns. A recurring *l*-mer structural pattern was defined when the latter was found in ≥3 non-redundant proteins (14); e.g. the *cdehja* pattern was found in three DBPs with different overall structures (Figure 1), and is thus a potential 3D motif.

### Definition of a structural/3D motif

To verify that the backbone structures pertaining to each recurring structural pattern are indeed conserved, they were compared using the MultiProt program (22). MultiProt derives multiple structural alignments from simultaneous superposition of the input protein structures and detects common geometric cores. Each input molecule was treated as a pivot in turn, and the RMSD of the geometric core $C^{\alpha}$ atoms of a non-pivot input molecule from those of the pivot molecule was computed. The resulting RMSD values were then averaged. Since a continuous *l*-mer structural pattern encompasses $l+4$ residues (the *l* center residues and two residues each at the N- and C-terminal sides), the structures of the respective $l+4$ residue segments of a given recurring structural pattern were compared using MultiProt. As MultiProt does not require all input molecules to participate in the alignment, a 3D motif was defined when the recurring *l*-mer structural pattern has a geometric core composed of ≥*l* residues common to 'all' the input structures; e.g. although the 6-mer structural patterns, *cdehja* and *ddehjl*, are each found in three DBPs, only *cdehja* was considered to be a 3D motif, as the backbone structures are truly conserved (Figure 2).

### Definition of a DNA-binding motif and a DNA-specific motif

The occurrence frequencies of a given 3D motif in non-redundant proteins that bind DNA, RNA, obligate proteins and non-obligate proteins were computed. A DNA-binding motif is defined as an *l*-mer motif containing ≥*l*/6 DNA-binding residues whose frequency in DBPs relative to that in all non-redundant proteins is ≥1.5. If the DNA-binding motif is absent in non-redundant proteins that bind RNA, obligate proteins and non-obligate proteins, it was considered to be DNA-specific.

## RESULTS

### 3D motifs for DBPs

To illustrate our motif discovery strategy, each DNA-bound protein structure was encoded into its 1D structural sequence and scanned using an *l*-letter sliding window ($l = 6, 7,\ldots, 27$). The *l*-mer structural patterns found in ≥3 non-redundant DBPs with conserved backbone segments (Figures 1 and 2), which in turn dictate the rotameric state of the respective side chains, are regarded as 3D motifs. Since each letter consists of five residues, a *l*-mer motif comprises $l+4$ residues: the central residues denoting each letter are labeled $P_1$, $P_2,\ldots,P_l$, whereas the two residues N-terminal to $P_1$ are labeled $P_{-2}$ and $P_{-1}$, while the two residues C-terminal to $P_l$ are labeled $P_{+1}$ and $P_{+2}$. To discover 3D motifs that may be biologically important but may not be specific to DBPs, we analyzed the 'most common' 6-mer motif that persists with increasing motif size. This yielded a novel *minimalist* functional scaffold, as described below.

### A novel 10-residue scaffold: the 'corner' motif and its structural definition

The most popular 6-mer 3D motif is *afklmm*, which is found 75 times in 40 of the 76 non-redundant DBPs (Table 1). It remains conserved as the motif size increases from 6 to 26; Supplementary Table S2 lists all *l*-mer motifs ($l = 6,\ 7,\ldots, 26$) containing the *afklmm* segment. This *afklmm* motif comprises part of a turn connected to a helix (Figure 3A, left panel) and appears at a corner, hence we will refer to it as the 'corner' motif. It can be characterized by the following three features (Figure 3B): (i) a helix starting at $P_4$ characterized by two conserved H bonds between the $P_3$ and $P_4$ amide N atoms and the $P_{+1}$ and $P_{+2}$ carbonyl O atoms, respectively; (ii) a solvent exposed surface formed by the $P_2$, $P_3$ and $P_4$ side chains; and (iii) H bonding or vdW interactions between the $P_1$ side chain, which points toward the protein interior, and the $P_5$ and/or $P_6$ side chains, resulting in a mean $P_1+P_5+P_6$ SASA (12 Å$^2$) that is generally smaller than the mean $P_2+P_3+P_4$ SASA (71 Å$^2$). In cases where $P_1$ lacks a side chain to interact with $P_5$ or $P_6$ (e.g. $P_1$ is Gly in the Flp recombinase structure, 1flo-A), its role is played by its neighbor $P_2$ whose side chain interacts with $P_5$. In many cases, a third conserved H bond is formed between the $P_2$ amide N and the $P_6$ carbonyl O, in addition to the $P_{+1}\rightarrow P_3$ and $P_{+2}\rightarrow P_4$ backbone−backbone H bonds. Despite these conserved structural features, the *afklmm* motif exhibits little sequence conservation except that $P_{-1}$ is often Gly, whereas $P_6$ is generally an aliphatic residue, Val, Ile, Ala or Leu (Figure 3C).

### Relationship between the 'corner' motif and the HTH motif

To determine if the 'corner' motif is part of the HTH motif found in several DBPs, our motif discovery strategy was applied to the HTH group of 16 DBP families reported in previous work (23). These proteins were grouped according to their CATH codes, and the

**Table 1.** The *afklmm* segments found in 40 structurally non-redundant dsDNA-binding proteins

| PDB code | Protein | *afklmm* segments[a] | CATH code[b] | DNA-contact residues[c] | Ligand-contact residues[d] |
|---|---|---|---|---|---|
| 1a31−A | DNA Topoisomerase 1 | 582−591 | – | ---**TAK**V--- | ---------- |
| 1ais−B | TFIIB | 1123−1132 | 1.10.472.10 | ---------- | ---------- |
| | | *1179−1188* | | ---------- | ----K----- |
| | | 1243−1252 | | --**KS**--G-- | --KSP----- |
| | | 1275−1284 | | --**VT**—**T**--- | ---------- |
| 1bdt−A | Transcriptional repressor arc | 29−38 | 1.10.1220.10 | ---**SVNS**--- | ----VN---- |
| 1bf5−A | Signal transducer & activator of transcription 1-α/β | 669−678 | 3.30.505.10 | ---------- | ---------- |
| 1d02−A | R.MunI | *107−116* | 3.40.580.10 | ---------- | ---------- |
| | | 182−191 | | ---------- | --E------- |
| 1dc1−A | R.BsoBI | 76−85 | 1.10.238.90 | ---**SDK**A--- | --IS-KA--- |
| 1diz−A | DNA-3-methyladenine glycosylase 2 | 124−133 | 1.10.340.30 | --**VSV-M**--- | ---------- |
| | | 172−181 | | --**MP**--**R**--- | ---------- |
| | | 213−222 | | --**IGRWT**--- | ---------- |
| | | *246−255* | 1.10.1670.10 | ---------- | ---------- |
| 1dnk−A | Deoxyribonuclease-1 | 9−18 | 3.60.10.10 | ---**GETK**--- | ---------- |
| | | *239−248* | | ---------- | ---------- |
| 1e3o−C | POU domain, class 2, transcription factor 1 | *23−32* | 1.10.260.40 | ---------- | ---------- |
| | | 138−147 | 1.10.10.60 | ------**VI**-- | ---------- |
| 1ecr−A | Ter-binding protein | 132−141 | 3.50.14.10 | ----**TA**-**R**-- | ---------- |
| | | 148−157 | | ----**TL**N--- | ---------- |
| 1efa−A | Lactose operon repressor | 13−22 | 1.10.260.40 | --**VSYQT**--- | ---------- |
| 1flo−A | FLP | 188−197 | 1.10.443.10 | ---**RFS**D--- | ---------- |
| | | 299−308 | | --**GPKSH**--- | ------H--- |
| 1fok−A | R.FokI | 321−330 | 1.10.10.10 | ---------- | ---------- |
| 1gdt−A | Transposon γ-δ resolvase | 157−166 | 1.10.10.60 | ---**GASH**--- | ---------- |
| | | 168−177 | | --**IARST**--- | ---------- |
| 1i3j−A | IRF protein | 222−231 | | --**ISSGL**--- | ---------- |
| 1ic8−A | HNF-1A | 126−135 | 1.10.260.40 | --**IPQR**---- | ---------- |
| | | 137−146 | | --**NQSH**--- | ---------- |
| 1ig9−A | DNA polymerase | 176−185 | – | ---------- | ---------- |
| | | *351−360* | – | ---------- | ---------- |
| | | 799−808 | – | --CPF**HIR**-- | ---------- |
| 1j1v−A | Glutathione transferase GST1-3 | 387−396 | 1.10.1750.10 | ---------- | ---------- |
| | | 430−439 | | --**RDHTT**--- | ---------- |
| 1j3e−A | Protein seqA | 102−111 | 1.20.1380.10 | ---------- | ---------- |
| 1jey−A | ATP-dependent DNA helicase 2 subunit 1 | 209−218 | 3.40.50.410 | ---------- | ---------- |
| 1jt0−A | HTH-type transcriptional regulator qacR | 32−41 | 1.10.10.60 | --**SSKGN**--- | ---------- |
| 1jx4−A | Pol IV | 57−66 | – | ---**P**-**V**E--- | ---------- |
| | | 184−193 | 1.10.150.20 | --**IGNITA**-- | ---------- |
| 1l3l−A | Transcriptional activator protein traR | 198−207 | 1.10.10.10 | ---**KYNS**--- | ---------- |
| 1m3q−A | *N*-glycosylase/DNA lyase | 148−157 | 1.10.340.30 | --**NNI-RI**-- | ---------- |
| | | 244−253 | | --**VGTKV**--- | ---------- |
| 1mjo−A | Met repressor | 49−58 | 3.30.310.40 | ---**TNS**---- | ----NS-L-- |
| 1mus−A | Transposase for transposon Tn5 | *220−229* | 3.90.350.10 | ---------- | ---------- |
| 1orn−A | Endonuclease III | 41−50 | 1.10.340.30 | --**CTD-L**--- | ---------- |
| | | 116−125 | | --**VGRKT**--- | ---------- |
| 1p71−A | DNA-binding protein HU | 14−23 | 4.10.520.10 | ---------- | --V---QA-- |
| 1qrv−A | HMG-D | 28−37 | 1.10.30.10 | ----**VT**---- | ---------- |
| 1r8d−A | HTH-type transcriptional activator mta | 12−21 | 1.10.1660.10 | ---S**IRT**--- | ---------- |
| | | 58−67 | | --F**RLD**---- | ---------- |
| 1rrq−A | A/G-specific adenine glycosylase | 47−56 | 1.10.340.30 | --**TRV-T**--- | ---------- |
| | | 121−130 | | --**VGPYT**V-- | ---------- |
| 1tau−A | DNA polymerase I, thermostable | 37−46 | 3.40.50.1010 | ---------- | ---------- |
| | | 114−123 | | ---------- | ---------- |
| | | 172−181 | – | ---------- | ---------- |
| | | *634−643* | 1.10.150.20 | ---------- | ---------- |
| 1tro−A | Trp operon repressor | 75−84 | 1.10.1270.10 | --**AGIAT**--- | ---------- |
| 1u8r−A | Iron-dependent repressor ideR | 34−43 | 1.10.10.10 | --**QS-PT**--- | ---------- |
| | | 90−99 | 1.10.60.10 | ---------- | --LP--E--- |
| 1uut−A | DNA binding trs helicase | 20−29 | 3.40.1310.20 | -----**SF**--- | ---------- |
| | | 97−106 | – | ---------- | ---------- |
| 1x9n−A | DNA ligase 1 | 271−280 | 1.10.3260.10 | ---------- | ---------- |
| | | 347−356 | – | --**VGDG**---- | ---------- |

(continued)

**Table 1.** Continued

| PDB code | Protein | *afklmm* segments[a] | CATH code[b] | DNA-contact residues[c] | Ligand-contact residues[d] |
|---|---|---|---|---|---|
| | | 362–371 | – | ---------- | ---------- |
| | | 452–461 | – | --**LAEQS**V-- | ---------- |
| | | *485–494* | – | ---------- | ---------- |
| | | 798–807 | 2.40.50.140 | --**FSD**E---- | ---------- |
| 1xo0–A | Recombase cre | 81–90 | 1.10.150.130 | --**LA**-**KT**--- | ---------- |
| | | 170–179 | 1.10.443.10 | ---**K**------ | ---------- |
| | | 302–311 | | ---------- | --V---E--- |
| 2bcq–A | Pol Lambda | 303–312 | 1.10.150.110 | --**IGKRM**--- | ---------- |
| | | 319–328 | – | -----**K**---- | ---------- |
| | | 342–351 | 1.10.150.20 | --**AGTKT**--- | ---------- |
| 2cgp–A | Catabolite gene activator | 176–185 | 1.10.10.10 | --**CSRET**--- | ---------- |
| 3orc–A | Regulatory protein cro | 23–32 | 3.30.240.10 | --**VY**--**A**--- | ---------- |

[a] Segments in *italics* might have ≥1 atoms close to a DNA atom if the protein had been complexed with a longer DNA, whereas segments underlined do not contain any atoms close to DNA, but have at least one atom within 5 Å of an atom in another protein chain.
[b] CATH code of the domain containing the *afklmm* segment—a dash means no CATH code has been assigned for that domain.
[c] Residues whose atoms are in vdW contact or are H bonded directly or indirectly via water molecules with DNA atoms are in bold, whereas the other residues have atoms within 5 Å of a DNA atom.
[d] Residues whose atoms are within 5 Å of an atom in another protein chain.

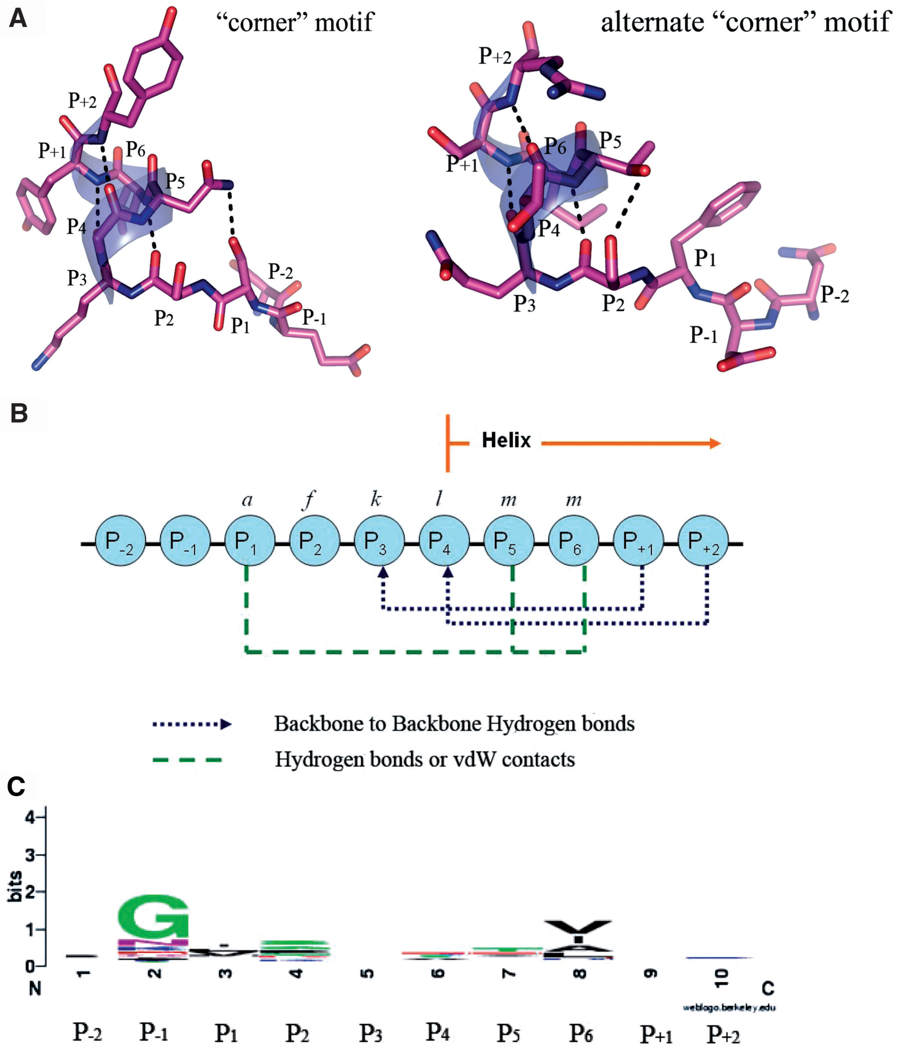


**Figure 3.** The 'corner' motif. (**A**) Representative 3D structures of the *afklmm* 'corner' motif (left) in the *Staphylococcus aureus* multidrug-binding protein QacR (1jt0-A) consisting of residues $^{32}$SESSKGNLYY$^{41}$ and an alternative *dfklmm* 'corner' architecture (right) in POU domain, class 2, transcription factor 1 (1e3o-C) consisting of residues $^{40}$NDFSQTTISR$^{49}$. The dotted lines denote H bonds. (**B**) Structural definition of the 'corner' motif (see text). (**C**) Sequence logo of the 'corner' motif.

best resolution complex structure in each group was chosen, analogous to our dataset construction. The 11 non-redundant HTH DBPs are listed in Table 2 along with the HTH motif and the corresponding structural letter sequence for each protein. The *afklmm* motif is found to be part of the HTH motif identified in these proteins, except the heat shock transcription factor, 3hts. The largest common structure that is shared by 9 of the 11 HTH DBPs is *mmmmnopafklmmmm* or *m*(4)*nopafklm*(4) composed of 19 amino acid residues. It is also found in other regions of four proteins (1d3u-B:1116−1134, 1d3u-B:1212−1231, 1ddn-A:83−101, 1gdt-A:150−168, 1lmb-3:22−40 and 1lmb-3:62−80).

## Relationship between the 'corner' motif and the α–α corner motif

The α−α corner is structurally defined by two consecutive, crosswise-packed helices connected by ≥2 residues (31). To determine its relationship with the HTH or 'corner' motif, our motif discovery strategy was applied to the proteins reported to contain α–α corners by Efimov (31). Since the latter listed only the protein names and the α−α corner amino acid sequences, the PDB was searched for those proteins containing the reported α−α corner amino acid sequences and structure. This resulted in nine proteins (Supplementary Table S3). In two proteins (1run-A, 1lmb-3), the 4 α−α corner amino acid sequences overlap with the respective HTH amino acid sequences in Table 2. For the other proteins (1d1l-A, 1crn-A, 1rqu-A, 1grt-A, 2mb5-A, 1eca-A, 1ibe-A/B), the α−α corner structural sequences all encompass *m*(4)*nopafklm*(3), which is characteristic of the HTH motif.

## Functional role of the 'corner' motif

To determine if the 'corner' motif plays a functional role, *afklmm* segments that contain ≥1 DNA-binding residues were deemed to be functional. The results indicate that the 'corner' motif is generally involved in binding dsDNA: 49 of the 75 *afklmm* segments found in 32 out of 40 non-redundant DBPs possess ≥1 DNA-binding residues. Another nine *afklmm* segments (highlighted in *italics* in Table 1) in nine proteins (1ais-B, 1d02-A, 1diz-A, 1dnk-A, 1e3o-C, 1ig9-A, 1mus-A, 1tau-A and 1x9n-A) might have ≥1 atoms close to a DNA atom if the protein had been complexed with a longer DNA, as illustrated in Figure 4A. Six of these nine proteins

**Table 2.** The *m*(4)*nopafklm*(4) motif (in bold italics) in structurally non-redundant HTH proteins

| PDB | Protein name | HTH motif amino acid and structural letter sequence | *m*(4)*nopafklm*(4) segment[a] | CATH code[b] |
|---|---|---|---|---|
| 1d3u−B | Transcription initiation factor IIB | 1268−1292[c]<br>***mmmmnopafklmmmm***mmmmmm | 1116−1134<br>1212−1231<br>1268−1286 | 1.10.472.10 |
| 1ddn−A | Diphtheria toxin repressor | 27−50[d]<br>***mmmmnopafklmmmm***mmmmm | 27−45<br>83−101 | 1.10.10.10<br>1.10.60.10 |
| 1fok−A | Type-2 restriction enzyme FokI | 325−353[e]<br>mmmmmmmnopacb*fklmmm*mmmmmm | 321−330* | 1.10.10.10 |
| 1gdt−A | Transposon γ-δ resolvase | 161−181[f]<br>***mmmmnopafklmmmm***mm | 150−168<br>161−179 | 1.10.10.60 |
| 1lmb−3 | Repressor protein CI | 33−51[f]<br>***mmmmnopafklmmmm*** | 22−40<br>33−51<br>62−80 | 1.10.260.40 |
| 1qpz−A | HTH-type transcriptional repressor purr | 4−23[f]<br>***mmmmnopafklmmmm***m | 4−22 | 1.10.260.40 |
| 1run−A | Catabolite gene activator | 169−189[g]<br>***mmmmnopafklmmmm***mm | 169−187 | 1.10.10.10 |
| 1trr−A | Trp operon repressor | 68−91[h]<br>***mmmmnopafklmmmm***mmmmm | 68−86 | 1.10.1270.10 |
| 2hdd−A | Segmentation polarity homeobox protein engrailed | 28−57[f]<br>mmm***mmmmnopafklmmmm***mmmmmmmm | 31−49 | 1.10.10.60 |
| 3hts−B | Heat shock factor protein | 228−254[i]<br>mmmmmmmmmpfbdc*fklmmm*mmm | – | 1.10.10.10 |
| 6cro−A | Regulatory protein cro | 16−36[j]<br>***mmmmnopafklmmmm***mm | 16−34 | 3.30.240.10 |

[a] An asterisk means absence of *mmmmnopafklmmmm* but presence of *afklmmm*; a 'dash' means no *mmmmnopafklmmmm* or *afklmmm* structural letter sequence in the protein structure.
[b] CATH code for the domain containing the *afklmm* segment.
[c] From Littlefield *et al.* (24).
[d] From White *et al.* (25).
[e] From Wah *et al.* (26).
[f] From Jones *et al.* (6).
[g] From Parkinson *et al.* (27).
[h] From Lawson and Carey (28).
[i] From Littlefield *et al.* (29).
[j] From Albright and Matthews (30).

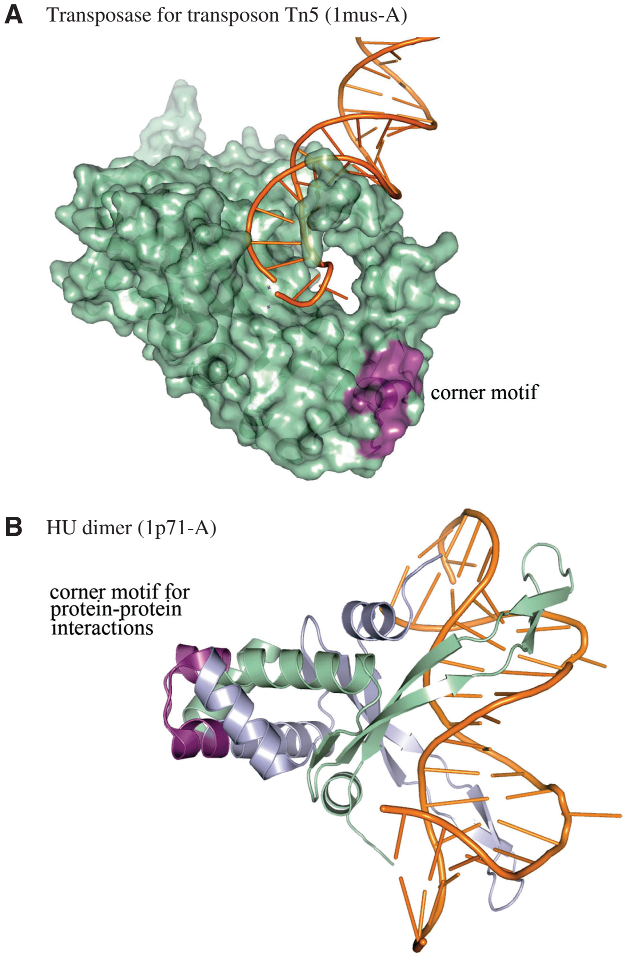

A
Transposase for transposon Tn5 (1mus-A)
corner motif
B
HU dimer (1p71-A)
corner motif for protein-protein interactions

**Figure 4.** Proteins with *afklmm* segments that might contact DNA or are involved in oligomerization. (**A**) The *afklmm* segment of the transposase for transposon Tn5 (1mus-A; 220–229) might contact DNA if the protein had been complexed with a longer DNA. (**B**) Residues 14–23 comprising the *afklmm* segment of chain A (green) contact residues in chain B (light blue) in the HU dimer (1p71-A), but they do not contact DNA. In (A) and (B), the *afklmm* segment is in magenta, while the DNA is in orange.

(except 1d02-A, 1mus-A and 1tau-A) already possess *afklmm* segments with ≥1 DNA-binding residues. The 'corner' motif does not appear to recognize specific DNA sequences, and binds in both the DNA major and minor grooves.

However, 17 *afklmm* segments appear to be located far from the DNA–protein interface even though the $P_2$, $P_3$ and $P_4$ side chains are solvent exposed. One plausible reason is that the residues in the *afklmm* segment may be involved in binding protein or RNA rather than DNA. Indeed, 4 of the 17 *afklmm* segments in *oligomeric* DBPs (1d02-A, 1p71-A, 1u8r-A, 1xo0-A) contain ≥1 atoms within 5 Å of an atom in another protein chain (Table 1 and Figure 4B). This indicates that the 'corner' motif may play a role in protein dimerization or oligomerization. The remaining 13 *afklmm* segments not involved in binding dsDNA or oligomerization await more new structures to verify if they participate in binding.

## Alternate forms of the 'corner' architecture

In the *afklmm* motif, the side chain of the center residue of the letter *a* in *afklmm* ($P_1$) always points toward the protein interior in order to interact with the $P_5$ and $P_6$ side chains, which are part of a helix. Since alternative interactions could stabilize this 'corner' architecture, those proteins lacking the *afklmm* motif may still employ similar 'corner' architecture to recognize DNA. For example, the ***d****fklmm* segment of POU domain class 2 transcription factor 1 (1e3o-C), consisting of amino acid 40–49, has a 'corner' architecture like the *afklmm* motif (see Figure 3A, right panel). Both $P_1$ (Phe) and $P_2$ (Ser) interact with $P_5$ (Thr) through vdW contacts and H bonds, respectively. This alternate 'corner' architecture also binds DNA via the $P_1$–$P_5$ residues.

## DNA-specific 3D motifs and their structural definition

To discover novel motifs characteristic of DBPs, DNA-binding motifs, as defined in the 'Materials and Methods' section, were identified and listed in Supplementary Table S4. 'Only' in DBPs, 76 DNA-binding motifs were found. These fall into two groups: 70 DNA-specific motifs with $l = 10–25$ were present in 'non-HTH' DBPs (1orn-A, 1rrq-A, 2bcq-A), while six with $l = 16–21$ were found in HTH DBPs (1jt0-A, 1r8d-A and 1tro-A). The longest motif in each of these two groups of DBPs was chosen as the representative DNA-specific motif. The 29-residue *cfbfklm*(4)*ghiafklm*(8) motif found in the non-HTH DBPs is structurally defined by (i) a 4–5-residue helix from $P_5$ and a second helix from $P_{16}$ containing ≥11 residues; (ii) conserved $P_5 \leftrightarrow P_{23}$ and $P_9 \leftrightarrow P_{19}$ vdW contacts between the two helices; and (iii) conserved backbone–backbone $P_8 \rightarrow P_{11}$ and $P_{11} \rightarrow P_{14}$ H bonds in the region connecting the two helices (Figure 5A). On the other hand, the 25-residue *cfklm*(4)*nopafklm*(6) motif found in the HTH DBPs is structurally defined by (i) a 8-residue helix from $P_3$ and a second helix from $P_{14}$ containing ≥7 residues, and (ii) conserved $P_3 \leftrightarrow P_{17}$ and $P_7 \leftrightarrow P_{14}$ vdW contacts between the two helices (Figure 5B). Notably, this motif encompasses the *m*(4)*nopafklm*(4) motif found in nine HTH protein families, validating it as a DNA-specific motif. Interestingly, although the *afklmmm* motif is not DNA-specific, it is common to both *cfbfklm*(4)*ghiafklm*(8) and *cfklm*(4)*nopafklm*(6) motifs.

## Effect of protein conformational change upon DNA binding on the 3D motifs

Thirty-six non-redundant DBPs have 3D structures solved with and without DNA and the $C^{\alpha}$ RMSDs of the DNA-bound protein structures from the respective free structures range from 0.4 to 5.9 Å. To assess how protein flexibility and conformational changes upon DNA binding affect the non-specific 'corner' motif and the two representative DNA-specific motifs found in the DNA-bound protein structures, the respective free structures were encoded into 1D structural sequences. The 'corner' motif or its alternate form was found in both

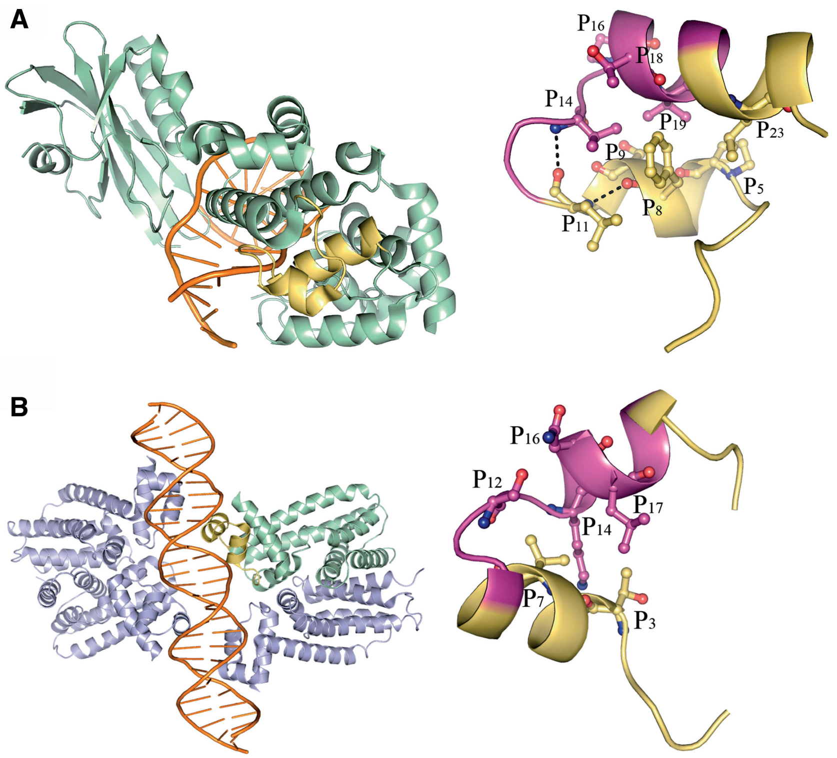


**Figure 5.** DNA-specific motifs. Representative 3D structure of (**A**) the *cfbfklmmmmghi**afklmmm**mmmmm* motif (yellow) in the MutY adenine DNA glycosylase (1rrq-A:108–136), and (**B**) the *cfklmmmmnop**afklmmm**mmm* motif (yellow) in the *S. aureus* multidrug-binding protein QacR (1jt0-A:21–45). The characteristic structural features of each motif are shown on the right. Note that $P_{14}$ in (A) and $P_{12}$ in (B) correspond to $P_1$ of the ***afklmm*** motif (in magenta).

the DNA-bound and free protein structures (Supplementary Table S5). The DNA-specific *cfklm*(4)*nopafklm*(6) motif found in the DNA-bound protein structures (1jt0-A, 1r8d-A) is nearly conserved in the respective free HTH proteins: the corresponding structural sequence is *cfklm*(4)*nopafklm*(4)*no* in 1rkw-A, and *Zfklm*(4)*nopafklm*(6) in 1jbg-A. The DNA-specific *cfbfklm*(4)*ghiafklm*(8) motif was 'not' found in any of the DBPs with both bound and free structures. Thus, conformational changes upon binding DNA, unless huge, do not seem to affect the 'corner' or representative DNA-specific motifs.

## DISCUSSION

### A novel strategy for discovering 3D motifs across protein families with different folds

We have presented a general and efficient strategy for discovering 3D motifs across protein families systematically on a large scale. The method requires as input the protein 3D structure, which is converted to a 1D structural letter sequence using a 16-letter structural alphabet. It yields as output a set of 3D motifs of various sizes shared by proteins with insignificant sequence or overall structural similarities. A non-specific motif (the 'corner' motif) was discovered in 40 non-redundant DBPs by analyzing the most common 6-mer motif that is conserved with increasing motif size. Furthermore, two representative DNA-specific motifs were found by choosing the largest DNA-binding motifs present 'only' in DBPs. One of these two DNA-specific motifs contain the HTH motif as a substructure, validating our strategy for discovering DNA-specific 3D motifs.

### Comparison with previous work

The new method of discovering 3D motifs across protein families with different folds complements previous motif discovery methods. A key advantage of our motif search strategy is that it does not require a query structure of a known motif for comparison against the PDB structures, or homologous sequences to identify conserved residues, or experimentally known functionally/structurally crucial residues. However, it is limited to detecting motifs composed of successive residues along the primary sequence. Thus, it complements previous methods (see 'Introduction' section), which require a known motif template or essential residues to create 3D templates, but yield 3D motifs composed of spatially interacting residues. A second advantage of our motif search strategy is that it can identify 3D motifs that are smaller than those defined by previous methods such as PROMOTIF (32), which detects motifs comprising of ~20–200 residues. For example, the HTH motif consists of ~20–30 residues and is found in 11 protein families (Table 2), whereas the 'corner' motif consists of only 10 residues and is found in 40 protein families (Table 1). A third advantage of our motif search strategy is that it provides a less ambiguous structural definition for 3D motifs by using two similarity measures, RMSDa [Equation (1)] and $C^{\alpha}$ RMSD (Figures 2, 3 and 5). For example, it provided a common structural *m*(4)*nopafklm*(4) sequence for the α–α corner connected by two residues and the HTH motif, except for the HTH motifs in type-2 restriction enzyme Fok I and the heat shock factor protein (Table 2).

### The 'corner' architecture: biological implications

The *afklmm* motif discovered herein (Figure 3A, left panel) has an architecture that confers a stable scaffold and enables diverse interactions, making it suitable for binding. Its 'corner' architecture enables the $P_1$, $P_5$ and $P_6$ side chains to interact, in addition to the $P_{+1} \rightarrow P_3$ and $P_{+2} \rightarrow P_4$ backbone−backbone H bonds of the helix, thus stabilizing the scaffold. The 'corner' architecture also exposes the $P_2$, $P_3$ and $P_4$ side chains, allowing for a wider variety of spatial arrangements than an architecture encompassing these side chains in a cavity or flat surface. This feature could help proteins employ the same architecture using different side chains to bind to different DNA targets; e.g. the *Staphylococcus aureus* multidrug-binding protein QacR (1jt0-A) employs $^{34}$SSKGN$^{38}$ in the *afklmm* motif (amino acid 32−41) to bind $^{14}$Ade-Cyt-Cyt-Gua$^{17}$ in the 1jt0-E DNA chain as well as $^{21}$Gua, $^{22}$Ade and $^{24}$Cyt in the 1jt0-F DNA chain, but the Catabolite gene activator (2cgp-A) employs a different set of residues $^{178}$CSRET$^{182}$ with the same conserved backbone (amino acid 176−185) to recognize two different DNA triplets, $^{503}$Gua-Thy-Cyt$^{505}$ and $^{539}$Thy-Gua-Thy$^{541}$.

### Applications of the 'corner' motif

The 'corner' architecture provides the following potential applications. It can provide a useful scaffold for computational redesign of DBPs for improved DNA-binding affinity and altered binding specificity: residues at $P_2$, $P_3$ and $P_4$ could be mutated without perturbing the scaffold (Figure 3C). Subsequently, the designed mutants can be computationally screened using free energy calculations (33,34) to predict if they exhibit enhanced DNA-binding affinity and altered binding specificity.

It can also be used in conjunction with DNA-binding residue prediction methods to suggest DNA-binding sites in proteins; e.g. the N-terminal fragment of topoisomerase I (1mw8-X) contains two *afklmm* segments, shown in magenta in Figure 6A, comprising residues 297−306 and 381−390, none of which contact the single-stranded DNA in the X-ray structure (35). To evaluate if these two 'corner' motifs can nevertheless bind DNA, DNA-binding residues were predicted using a method based on detecting a cluster of evolutionary conserved surface residues that are electrostatically stabilized upon mutation to negatively charged Asp/Glu, as described in previous work (36). This yielded two distinct DNA-binding sites (labeled S1 and S2) for topoisomerase I (Figure 6B): the DNA-binding residues in the S1 site (in red) are A282, I285, T288, L289, Q291, S294, T295, **M305**, D323, L393, Q397, A480, K484 and E487, while those in the S2 site (in yellow) are **K302, M305, M306**, R321, G492, R493, P494, S495, T496, A498, S499, I500, I501 and S502 (residues in bold comprise the *afklmm* segment). Notably, the S2 site contains R321, S495, T496 and S499, which are within H-bonding/vdW contact of the single-stranded DNA, and the R321 backbone N is only 3.9 Å away from the catalytic tyrosine Y319 $C^{\delta 2}$. Thus, the 297−306 'corner' motif in conjunction with predicted DNA-binding residues suggest that the S2 site is likely to be the DNA-binding site in topoisomerase I.

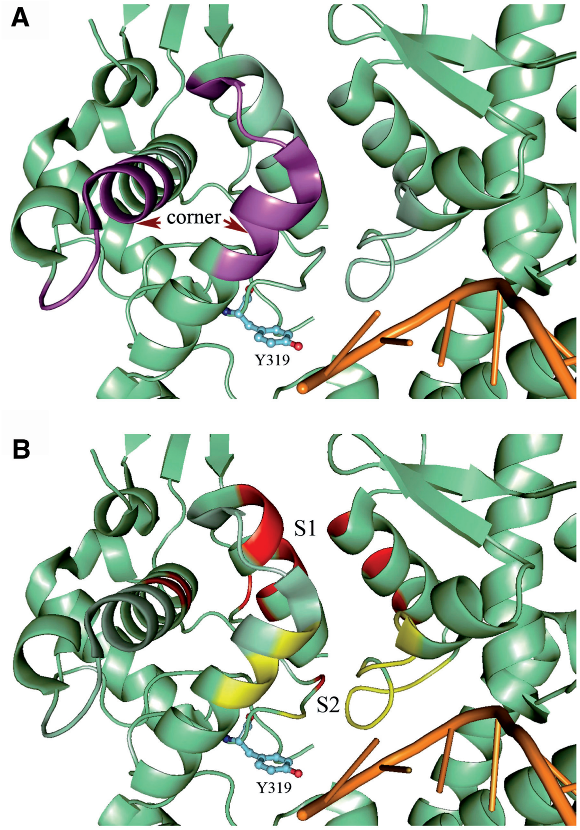


**Figure 6.** The predicted DNA-binding sites in the N-terminal fragment of topoisomerase I (1mw8-X). (**A**) The DNA-binding sites predicted by the two 'corner' motifs (magenta): 297−306 and 381−390, and (**B**) the DNA-binding sites, S1 (red) and S2 (yellow), predicted by the method described in our previous work (36). The catalytic residue, Y319, is shown as ball and stick.

### Applications of the DNA-specific motifs

The two DNA-specific motifs in Figure 5 may help to annotate proteins with known structures but unknown function. There are 2146 proteins in the Structural Genomics database with 'unknown function' in the title. For each of these proteins, the chain A structure was encoded into its 1D structural sequence and scanned using a 21- or 25-letter sliding window. None of the proteins with unknown function contain the 25-letter *cfbfklm*(4)*ghiafklm*(8) motif, but six contain the 21-letter *cfklm*(4)*nopafklm*(6) motif, out of which, only three possess the characteristic structural features depicted in Figure 5B (Table 3); viz., 2nx4 (amino acid 28−52), 2ia0 (amino acid 21−45) and 2g7u (amino acid 27−51). These

**Table 3.** Proteins with unknown function containing the *cfklm*(4)*nopafklm*(6) or HTH motif

| Proteins | Motif hits[a] | Motif features conserved[b] | HTHQuery hits[c] | HTH query prediction[d] |
|---|---|---|---|---|
| 3cym-A | 401−425 | No[e] | 403−424 | Unlikely (−9) |
| 2nx4-A | 28−52 | Yes | 30−51 | Possible (2) |
| 2ibd-A | 33−57 | No[f] | 35−56 | Possible (−2) |
| 2ia0-A | 21−45 | Yes | 23−44 | Likely (9) |
| 2g7u-A | 27−51 | Yes | 29−51 | Likely (9) |
| 2fi0-A | 48−72 | No[e] | 49−79 | Unlikely (−9) |

[a] Amino acid sequence corresponding to the *cfklm*(4)*nopafklm*(6) motif.
[b] The *cfklm*(4)*nopafklm*(6) motif features are conserved if the motif structure has a 8-residue helix from $P_3$ and a second helix from $P_{14}$ containing ≥7 residues, as well as $P_3 \leftrightarrow P_{17}$ and $P_7 \leftrightarrow P_{14}$ vdW contacts between the 2 helices (see Figure 5B).
[c] The amino acid sequence of the HTH motif according to HTHquery.
[d] The number in parentheses is an integer based score from the linear predictor of HTHquery. A score >3 is a likely hit (the protein is likely to have a DNA-binding HTH motif), a score between −3 and 3 is a possible hit, and a score less than −3 is an unlikely hit.
[e] Absence of $P_3 \leftrightarrow P_{17}$ and $P_7 \leftrightarrow P_{14}$ vdW contacts.
[f] The first helix contains only seven instead of eight residues.

three proteins are also predicted to be DBPs with HTH motifs according to HTHquery (10), suggesting that they are likely to bind DNA.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online

## ACKNOWLEDGEMENTS

We thank Dr Hanna Yuan and Ms Lauren Wang for help in preparing the figures.

## FUNDING

Funding for open access charge: National Science Council, Taiwan (grant NSC 95-2113-M-001-038-MY5).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) Identification and classification of protein fold families. *Prot. Engng.*, **6**, 485–500.
2. Kasuya,A. and Thornton,J.M. (1999) Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.*, **286**, 1673–1691.
3. Lin,K., Wright,J.D. and Lim,C. (2000) Long spacers in PROSITE patterns have a consensus backbone motif. *J. Mol. Biol.*, **299**, 539–548.
4. Watson,J.D., Laskowski,R.A. and Thornton,J.M. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.
5. Kristensen,D.M., Chen,B.Y., Fofanov,V.Y., Ward,R.M., Lisewski,A.M., Kimmel,M., Kavraki,L. and Lichtarge,O. (2006) Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity. *Prot. Sci.*, **15**, 1530–1536.
6. Jones,S., Barker,J.A., Nobeli,I. and Thornton,J.M. (2003) Using structural motif templates to identify proteins with DNA binding function. *Nucleic Acid Res.*, **31**, 2811–2823.
7. Shanahan,H.P., Garcia,M.A., Jones,S. and Thornton,J.M. (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.*, **32**, 4732–4741.
8. Pugalenthi,G., Suganthan,P.N., Sowdhamini,R. and Chakrabarti,S. (2008) MegaMotifBase: a database of structural motifs in protein families and superfamilies. *Nucleic Acids Res.*, **36**, D218–D221.
9. Madsen,D. and Kleywegt,G.T. (2002) Interactive motif and fold recognition in protein structures. *J. Appl. Cryst.*, **35**, 137–139.
10. Ferrer-Costa,C., Shanahan,H.P., Jones,S. and Thornton,J.M. (2005) HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics*, **21**, 3679–3680.
11. Goyal,K., Mohanty,D. and Mande,S.C. (2007) PAR-3D: a server to predict protein active site residues. *Nucleic Acids Res.*, **35**, W503–W505.
12. Bauer,R.A., Bourne,P.E., Formella,A., Frommel,C., Gille,C., Goede,A., Guerler,A., Hoppe,A., Knapp,E.W. and Pöschel,T.E.A. (2008) Superimpose: a 3D structural superposition server. *Nucleic Acids Res.*, **36**, W47–W54.
13. Debret,G., Martel,A. and Cuniasse,P. (2009) RASMOT-3D PRO: a 3D motif search webserver. *Nucleic Acids Res.*, **37**, W459–W464.
14. Dudev,M. and Lim,C. (2007) Discovering structural motifs using a structural alphabet: application to Mg-binding sites. *BMC Bioformatics*, **8**, 106–118.
15. de Brevern,A.G., Etchebest,C. and Hazout,S. (2000) Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Struct. Funct. Genet.*, **41**, 271–287.
16. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Iype,L., Jain,S., Fagan,P., Marvin,J. *et al.* (2002) The Protein Data Bank. *Acta. Crystallogr. D.*, **58**, 899–907.
17. Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.
18. Ponomarenko,J.V. and Bourne,P.E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct. Biol.*, **7**, 64–82.
19. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
20. Chen,Y.C. and Lim,C. (2008) Common physical basis of macromolecule-binding sites in proteins. *Nucleic Acids Res.*, **36**, 7078–7087.
21. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
22. Shatsky,M., Nussinov,R. and Wolfson,H.J. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.
23. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, 1–10.
24. Littlefield,O., Korkhin,Y. and Sigler,P.B. (1999) The structural basis for the oriented assembly of a TBP/TFB/promoter complex. *Proc. Natl Acad. Sci. USA*, **96**, 13668–13673.
25. White,A., Ding,X., vanderSpek,J.C., Murphy,J.R. and Ringe,D. (1998) Structure of the metal-ion-activated diphtheria toxin repressor/tox operator complex. *Nature*, **394**, 502–506.
26. Wah,D.A., Hirsch,J.A., Dorner,L.F., Schildkraut,I. and Aggarwal,A.K. (1997) Structure of the multimodular endonulcease FokI bound to DNA. *Nature*, **388**, 97–100.
27. Parkinson,G., Gunasekera,A., Vojtechovsky,J., Zhang,X., Kunkel,T.A., Berman,H. and Ebright,R.H. (1996) Aromatic hydrogen bond in sequence-specific protein DNA recognition. *Nat. Struct. Biol.*, **3**, 837–841.
28. Lawson,C.L. and Carey,J. (1993) Tandem binding in crystals of a trp repressor/operator half-site complex. *Nature*, **366**, 178–182.

29. Littlefield,O. and Nelson,H.C. (1999) A new use for the 'wing' of the 'winged' helix-turn-helix motif in the HSF-DNA cocrystal. *Nat. Struct. Biol.*, **6**, 464–470.
30. Albright,R.A. and Matthews,B.W. (1998) Crystal structure of lamda-Cro bound to a consensus operator at 3.0 Å resolution. *J. Mol. Biol.*, **280**, 137–151.
31. Efimov,A.V. (1984) A novel super-secondary structure of proteins and the relation between the structure and the amino acid sequence. *FEBS*, **166**, 33–38.
32. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF–a program to identify and analyze structural motifs in proteins. *Prot. Sci.*, **5**, 212–220.
33. Lo,C.-H., Chang,Y.-H., Wright,J.D., Chen,S.-H., Kan,D., Lim,C. and Liang,P.-H. (2009) A combined experimental and theoretical study of long-range interactions modulating dimerization and activity of yeast geranylgeranyl diphosphate synthase. *J. Am. Chem. Soc.*, **131**, 4051–4062.
34. Wang,Y.T., Wright,J.D., Doudeva,L.G., Jhang,H.C., Lim,C. and Yuan,H.S. (2009) Redesign of a non-specific endonuclease to yield better DNA-binding activity and altered DNA sequence preference in cleavage. *J. Am. Chem. Soc.*, **131**, 17345–17353.
35. Perry,K. and Mondragon,A. (2003) Structure of a complex between *E. coli* DNA topoisomerase I and single-stranded DNA. *Structure*, **11**, 1349–1358.
36. Chen,Y.C., Wu,C.Y. and Lim,C. (2007) Predicting DNA-binding sites on proteins from electrostatic stabilization upon mutation to Asp/Glu and evolutionary conservation. *Proteins: Struct. Funct. Bioinf.*, **67**, 671–680.