

SUPPLEMENTARY DATA

DNAProDB: an interactive tool for structural analysis of DNA-protein complexes

Jared M. Sagendorf¹, Helen M. Berman^{2,*}, and Remo Rohs^{1,*}

¹Molecular and Computational Biology Program, Departments of Biological Sciences, Chemistry, Physics & Astronomy, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA

²RCSB Protein Data Bank, Center for Integrative Proteomics Research, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

*To whom correspondence should be addressed. Tel: +1 213 740 0552; Fax: +1 821 4257; Email: rohs@usc.edu (R.R.) or Tel: +1 848 445 4667; Email: berman@rcsb.rutgers.edu (H.M.B.)

SUPPLEMENTARY METHODS – FEATURE EXTRACTION

Protein secondary structure

DNAProDB assigns a three-state secondary structure to each residue in the DNA-protein interface using the program DSSP (1). DSSP assigns an eight-state secondary structure; H – α helix, G – 3_{10} helix, I – π helix, E – extended strand, B – isolated β -bridge, S – bend, T – hydrogen bonded turn, and blank - loop/irregular. To map from eight-state to three-state secondary structure, H, G and I are assigned as helix (H); E and B are assigned as strand (S); and S, T, and blank are assigned as loop (L). Individual strands are treated independently without regard to the formation of β -sheets.

Buried solvent-accessible surface area

The buried solvent-accessible surface area (BASA) is the difference between the solvent-accessible surface area (SASA) (2) of some unit in the structure when it is in the free state and when it is in the bound state. For example, for every residue in the protein, SASA is calculated with all DNA removed from the structure, which is defined as the free state of the protein. All residues with a free SASA ($SASA_F$) > 0 constitute the protein surface, and may potentially

interact with the DNA. SASA values are re-calculated with the DNA present to determine the complex SASA ($SASA_C$). The $BASA$ of each residue is defined as

$$BASA = SASA_F - SASA_C,$$

which will always be greater than or equal to zero. Residues with $BASA > 0$ are considered to be in contact with the DNA, and the $BASA$ value describes the extent of the contact. The same calculation is performed for each nucleotide, with the free state of the DNA corresponding to the structure with all protein residues removed.

Different contributions of the $BASA$ values are determined. For each residue, it is possible to specify how much of the total $BASA$ is due to contact with the DNA major groove, minor groove, or backbone. To calculate these contributions, instead of removing the entire DNA structure for the free-state calculation, only part of the DNA is removed. To calculate the contribution of the DNA major groove to the $BASA$ value of each residue, only the DNA major groove atoms are removed from the structure in the free state. Therefore,

$$BASA_{wg} = SASA_F^{(wg)} - SASA_C,$$

where $BASA_{wg}$ is the surface area in \AA^2 of lost SASA for each residue due to contact with the DNA major groove (wg for 'wide groove'); and $SASA_F^{(wg)}$ is the SASA of each residue with only DNA major groove atoms removed.

In this way, $BASA_{wg}$ (major groove; wg), $BASA_{sg}$ (minor groove; sg for 'small groove') and $BASA_{bb}$ (backbone; bb for 'backbone') contributions are determined for each residue. Similarly, for each DNA nucleotide, the $BASA$ contributions due to helices, strands, and loops are determined. This analysis gives a more detailed description of the interface in terms of how much different parts of the structure are contacting each other. However, the sum of each contribution of the $BASA$ may be less than the total $BASA$ (but never greater) because calculating the $BASA$ contributions in this way excludes overlapping surface areas.

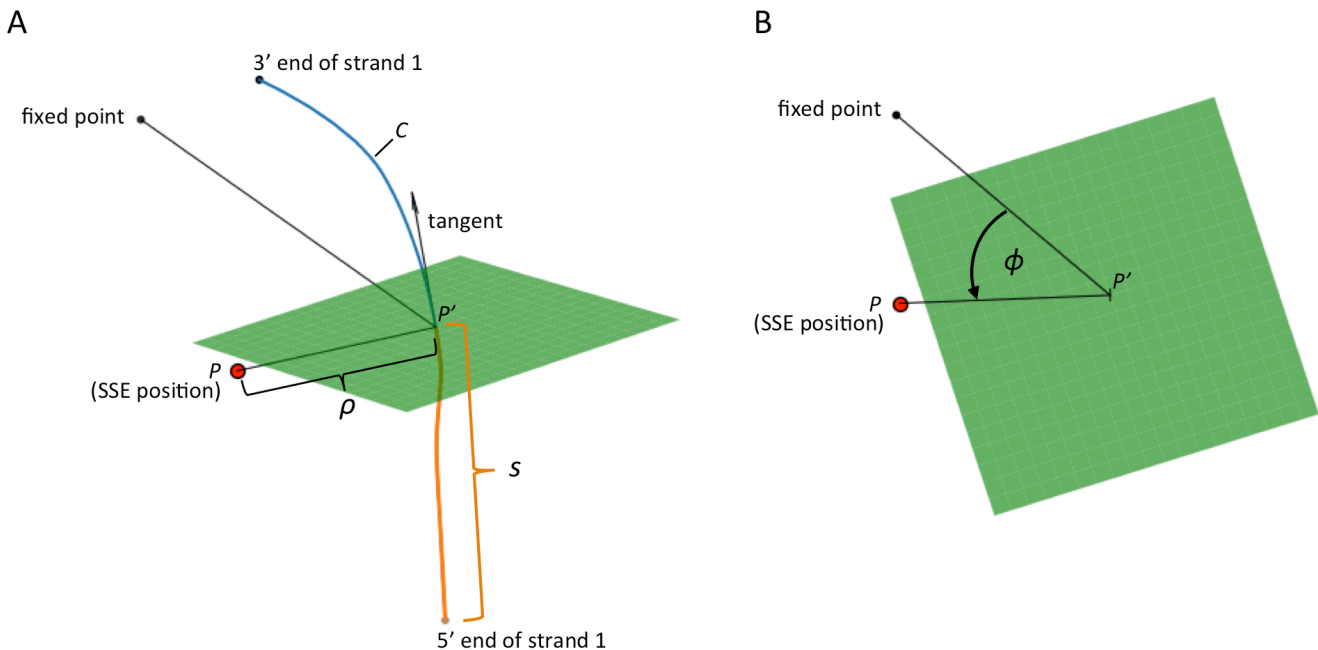
$BASA$ values are also determined for each residue-nucleotide pair. Here, the entire structure is removed except for the pair, and the free and complex states are calculated as described. Individual contributions of the $BASA$ are not determined for residue-nucleotide pairs.

SUPPLEMENTARY TABLE

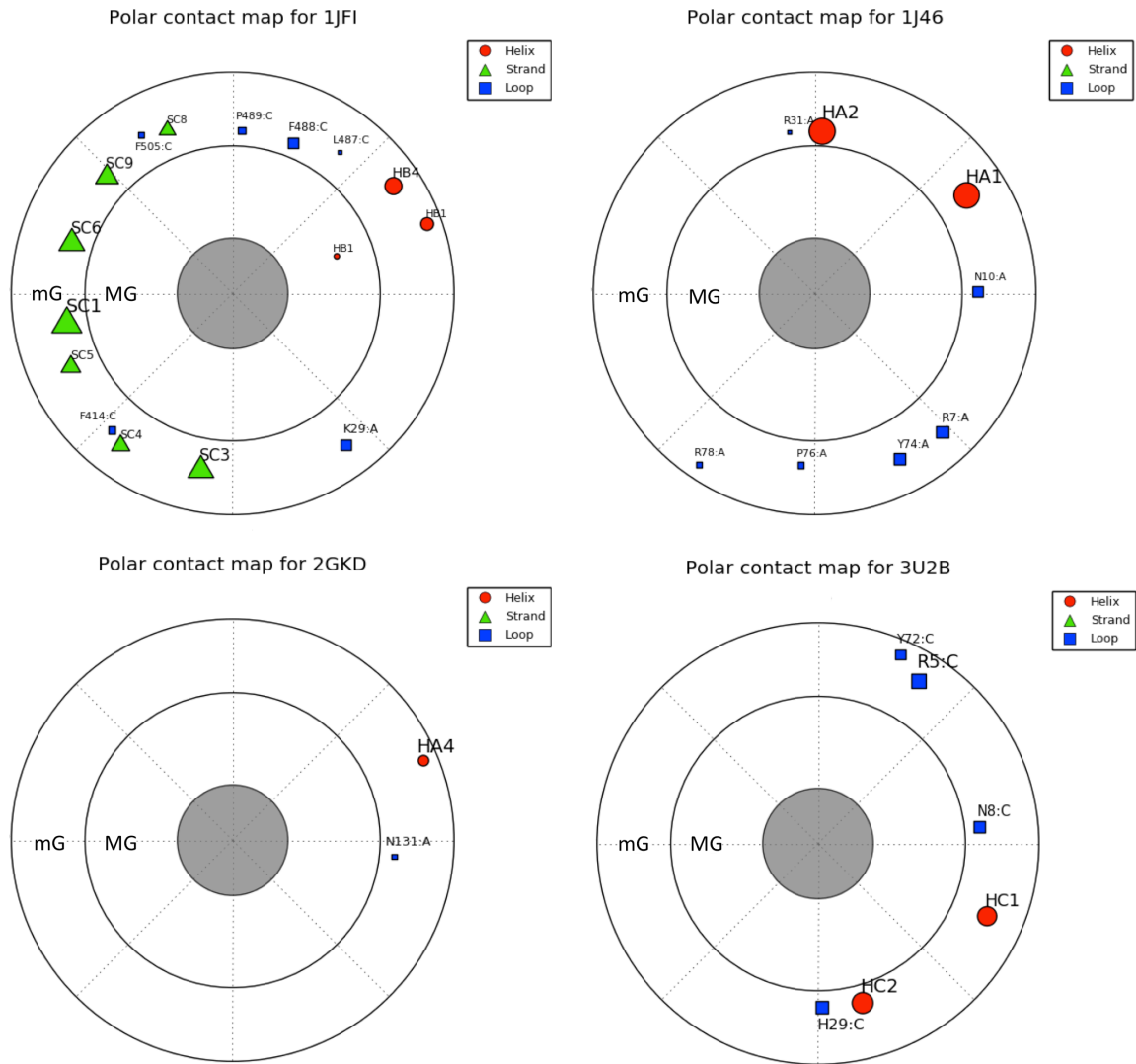
Modified Nucleotides		Modified Residues	
PDB ID	CHEMICAL NAME	PDB ID	CHEMICAL NAME
5CM	5-METHYL-2'-DEOXY-CYTIDINE-5'-MONOPHOSPHATE	MSE	SELENOMETHIONINE
DI	2'-DEOXYINOSINE-5'-MONOPHOSPHATE	SEP	PHOSPHOSERINE
6OG	6-O-METHYL GUANOSINE-5'-MONOPHOSPHATE	TPO	PHOSPHOTHREONINE
DOC	2',3'-DIDEOXYCYTIDINE-5'-MONOPHOSPHATE	PTR	O-PHOSPHOTYROSINE
8OG	8-OXO-2'-DEOXY-GUANOSINE-5'-MONOPHOSPHATE	CME	S,S-(2-HYDROXYETHYL)THIOCYSTEINE
DDG	2',3'-DIDEOXY-GUANOSINE-5'-MONOPHOSPHATE	OCS	CYSTEINESULFONIC ACID
2DA	2',3'-DIDEOXYADENOSINE-5'-MONOPHOSPHATE	ALY	N(6)-ACETYLLYSINE
2DT	3'-DEOXYTHYMIDINE-5'-MONOPHOSPHATE		
2PR	2-AMINO-9-[2-DEOXYRIBOFURANOSYL]-9H-PURINE-5'-MONOPHOSPHATE		
MRG	N2-(3-MERCAPTOPROPYL)-2'-DEOXYGUANOSINE-5'-MONOPHOSPHATE MERCAPTOPROPYL)-2'-DEOXYGUANOSINE-5'-MONOPHOSPHATE		
CTG	(5R,6S)-5,6-DIHYDRO-5,6-DIHYDROXYTHYMIDINE-5'-MONOPHOSPHATE		
1CC	5-CARBOXY-2'-DEOXYCYTIDINE MONOPHOSPHATE		

Supplementary Table 1. Currently supported chemically modified nucleotides and residues. Any non-standard residues or nucleotide not in this table will be removed from the structure during processing.

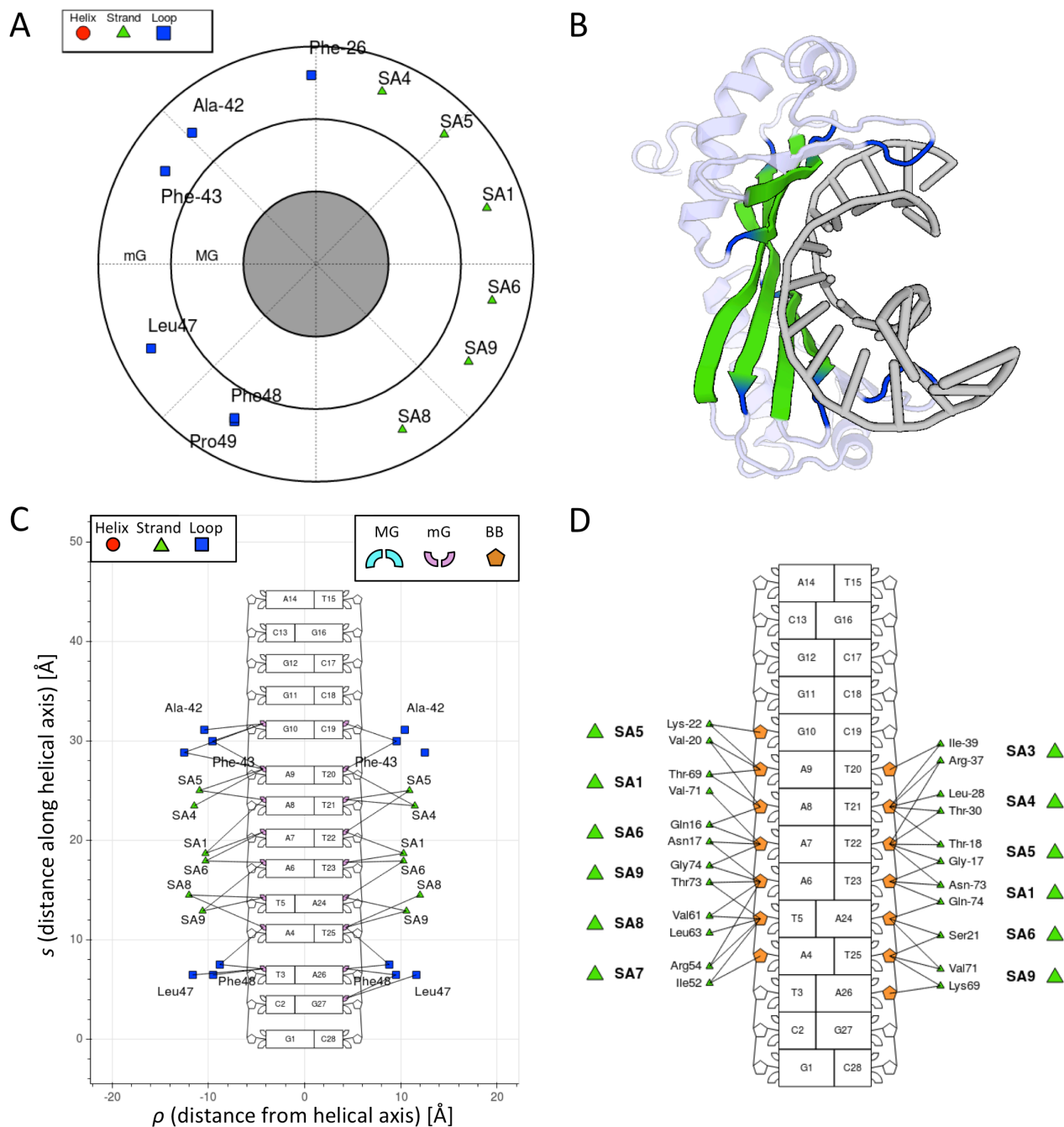
SUPPLEMENTARY FIGURES



Supplementary Figure S1. Illustration of the axial coordinate system used to specify the position of each SSE in the DNA-protein interface. Let P be a point in space that indicates the position of an SSE, C be the curve defining the DNA helical axis, and P' be a point on C where a perpendicular line can be drawn from P to C . The axial coordinates that define P are (ϕ, ρ, s) . ρ is the distance from P to P' ; s is the distance along the curve from a specified point on the curve to P' (here, chosen as the end of the helix axis, corresponding to the 5' end of the first DNA strand in the structure); and ϕ is the angle made in a plane defined by the tangent vector of the curve at P' between the line joining P and P' which lies in this plane, and the projection of a line joining P' and a fixed point in space to the plane. This fixed point is arbitrary; changing the fixed point simply amounts to adding some constant offset to ϕ relative to the original choice. In practice, the center of mass of the protein is chosen, which allows similar structures to be compared consistently. If this center of mass lies too near the DNA helix axis (which is the case for structures with perfect two-fold symmetry), the point is chosen to lie along a principal axis of the protein. **(A)** The curve in the figure represents the DNA helix axis (C), which, in this case, is not linear. The red circle is the position of an SSE (P). The length of the line joining P and P' is the coordinate ρ . The distance along C from the 5' end to P' is the coordinate s . The green plane contains the line joining P and P' , and is defined by the tangent vector of C at P' . **(B)** The angle between the projection of the line that joins P' and the fixed point to the plane, and the line that joins P and P' is the coordinate ϕ .



Supplementary Figure S2. Polar contact maps of selected structures meeting the search criteria described in Table 1. The complex with PDB ID 1JFI (3) contains a ternary complex of a TATA-binding protein (TBP) with Negative Cofactor 2 (NC2) bound to a TATA-box. The polar contact map shows a characteristic TBP interaction, with a series of strands in the minor groove (mG) in addition to two helices that contact the minor groove belonging to an NC2 domain. The complex with PDB ID 2GKD (4) shows the bacterial resistance protein Calicheamicin gene Cluster (CaIC), which binds with a single helix in the minor groove and few other contacts. The complexes with PDB IDs 1J46 (5) and 3U2B (6) contain proteins that predominantly bind with two helices and several loop contacts in the minor groove.



Supplementary Figure S3. Different visualizations for a TATA-box binding DNA-protein complex (PDB ID 1VTO) (7). Green triangles represent strand SSEs; and blue squares represent loops. **(A)** *Polar contact map* for this structure. The DNA helix axis is curved, but the visualization shows the strands as contacting on a single side of the DNA (i.e. confined to one half of the figure). This visualization is intuitive when compared with the 3D structure in **(B)**. **(B)** 3D representation of the structure, with strand and loop residues that contact the minor groove colored according to SSE type in green and blue, respectively. **(C)** *Linear contact map* of this structure, showing minor groove (mG) and major groove (MG) contacts. **(D)** *Nucleotide-residue contact map*, showing only DNA backbone (BB) interactions with strand residues. Loop interactions have been turned off.

SUPPLEMENTARY REFERENCES

1. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.
2. Lee, B. and Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.
3. Kamada, K., Shu, F., Chen, H., Malik, S., Stelzer, G., Roeder, R.G., Meisterernst, M. and Burley, S.K. (2001) Crystal structure of negative cofactor 2 recognizing the TBP-DNA transcription complex. *Cell* **106**, 71-81.
4. Singh, S., Hager, M.H., Zhang, C., Griffith, B.R., Lee, M.S., Hallenga, K., Markley, J.L. and Thorson, J.S. (2006) Structural insight into the self-sacrifice mechanism of enediynes resistance. *ACS Chem. Biol.* **1**, 451-460.
5. Murphy, E.C., Zhurkin, V.B., Louis, J.M., Cornilescu, G. and Clore, G.M. (2001) Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation. *J. Mol. Biol.* **312**, 481-499.
6. Jauch, R., Ng, C.K., Narasimhan, K. and Kolatkar, P.R. (2012) The crystal structure of the Sox4 HMG domain-DNA complex suggests a mechanism for positional interdependence in DNA recognition. *Biochem. J.* **443**, 39-47.
7. Kim, J.L. and Burley, S.K. (1994) 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nat. Struct. Biol.* **1**, 638-653.