SUPPLEMENTARY DATA

DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes

Jared M. Sagendorf[1], Nicholas Markarian[1], Helen M. Berman[2,3], and Remo Rohs[1,*]

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

[2] Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

[3] Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

*To whom correspondence should be addressed. Tel: +1 213 740 0552; Fax: +1 821 4257; Email: rohs@usc.edu

SUPPLEMENTARY METHODS

DNAproDB is available at https://dnaprodb.usc.edu

Detection of major and minor grooves for double-helical DNA

The major and minor grooves are important structural moieties for binding to double-stranded helical DNA. Many proteins recognize distinct biophysical signatures of the grooves such as hydrogen bond donor/acceptor patterns, DNA shape, or electrostatic potential. DNAproDB identifies the major groove and minor groove edges of base pairs which form double-stranded helices and distinguishes interactions that occur in either groove. For bases which form canonical Watson-Crick base pairs, the groove edges are known a priori. However, in general base-pairing geometry may substantially deviate from the Watson-Crick conformation and the glycosidic torsion angle and relative position and orientation of the base coordinate frames matters. Additionally, chemically modified nucleotides may have additional chemical groups that should be correctly identified if they protrude into one groove or the other.

We have developed a simple algorithm for distinguishing which atoms of each base in a base pair should be identified as being in the major groove or minor groove. Supplementary Figure S3 shows an illustration of our approach. First, the base-pair coordinate frame is determined for a given base pair using the program DSSR (1). The direction of the x-axis will either point towards the major or minor groove depending on the relative glycosidic bond angles of the two bases. The base pair is then treated as a graph with each atom in the pair being a node projected to the x-y plane of the base coordinate frame, and covalent bonds being edges. Knowing the direction of the minor groove, an edge joining the two bases is added to the graph which represents the hydrogen bond joining the two bases that is closest to the minor groove. The shortest path from the glycosidic nitrogen atoms of each base is then found. This path, in addition to the rays $\boldsymbol{N_1}$ and $\boldsymbol{N_2}$, bisects the plane. All atoms that lie along the path and are on the minor groove side are classified as minor groove atoms, and all atoms on the major groove side as major groove atoms. Note that this algorithm works for any nucleoside pair so long as a coordinate frame for the base pair can be defined.

DNA structural entity classification

DNAproDB classifies DNA structural entities based on the secondary structure of the entity. We assigned four classifications: *helical*, *single-stranded*, *helix/single-stranded* hybrid or *other*. The latter class is reserved for secondary structures which are either irregular, have no commonly used name, or are too infrequent to warrant their own classification. For every DNA structural entity, any helical segments and single-stranded segments present within the entity are first determined. Helical segments are identified using DSSR (1), which defines helices as one or more stems which stack on top of one another, allowing for some flexibility in terms of flipped out bases, backbone breaks etc. Single-stranded segments are defined as a segment of a DNA strand which does not belong to a helix, does not form any base pairs with other strands, does not have more than two consecutive intra-strand base pairs and is at least three nucleotides in length.

If a DNA entity has more than one helical segment, it is automatically classified as *other*. If a DNA entity has exactly one helical segment it is classified as *helical* if at least 60%

of nucleotides within the entity are part of the helical segment. Otherwise, if the number of helical nucleotides plus the number of single-stranded nucleotides is equal to or greater than 60%, the entity is classified as *helical/single-stranded*. If neither condition is met, it is classified as *other.*

Finally, if a structural entity contains no helical segments, then it is classified as *single-stranded* if at least 60% of nucleotides are in a single-stranded conformation, else it is classified as *other*. This is a heuristics-based classification scheme but it has been tested against a large number of structures and been found to work extremely well.

Within the *helical* assignment, three sub-classifications are provided; a *perfect helix*, an *imperfect helix* and an *irregular helix*. These classifications are based on a numerical feature DNAproDB computes for each helical segment called the *helix score*, which is the ratio of canonical base pairs to the total number of base pairs in the helix. A canonical base pair is a base pair in which both bases form stacking interactions to the neighboring bases on their respective strand. A *perfect helix* has a helix score above 0.9, an *imperfect helix* has a score between 0.9 and 0.6 and an *irregular helix* a helix score below 0.6.

Approximating parameters for chemically modified components

DNAproDB uses atomic van der Waals radii parameters for the calculation of solvent accessible and solvent excluded surface areas, and residue hydrophobicity scores for the calculation of spatial aggregation propensities (2) for protein surface residues. These parameters are generally not available for chemically modified components, so approximations are used in order to provide reasonable values for these features and still support as many chemical modifications as possible. DNAproDB supports chemical modifications of any of the standard 20 amino acids and 4 DNA nucleotides that have an entry in the PDB Chemical Component Dictionary (3) and do not significantly deviate from their parent component so as to make identification of structural moieties ambiguous (see main manuscript).

In the case of van der Waals radii, base atomic parameters for standard components are taken from the NACCESS (4) radii values. Radii for chemically modified components are then taken from any corresponding atoms in their standard parent and values for remaining atoms are taken from Table 9 of Batsanov S.S. (5) based on the element type of the atom.

For the calculation of spatial aggregation propensities (SAP), chemically modified residues are ignored. The SAP values of standard residues in close proximity to chemically modified residues may be slightly affected, and the SAP values for chemically modified residues is not reported.

Interacting moieties identification

Nucleotide–residue interactions are defined based on the minimum distance between a nucleotide and residue (excluding hydrogen atoms). The current release of DNAproDB uses a value of 4.5 Å, and any nucleotide–residue pairs with a distance beyond that cutoff value are not considered to be interacting. Given an interaction pair, DNAproDB identifies which structural moieties (see main manuscript) within the pair are interacting based on the values of interaction features (which are broken down by structural moiety). For example, if the NZ atom of a lysine (using PDB atom naming conventions) forms a hydrogen bond with the O6 atom of

a guanine in a helical conformation, then this is a side chain–major groove hydrogen bond. The total number of hydrogen bonds, van der Waals interactions (which are defined as heavy atom pairs within 3.9 Å but not forming a hydrogen bond) and values of buried solvent accessible surface area components (see Supplementary Data of Sagendorf et. al. (6) for a description) are used to determine which structural moieties are interacting. To determine cut-off values, the distributions of interaction feature values among a large sample of nucleotide–residue interaction pairs (464,303) was compiled using DNAproDB data generated from PDB structures. We note that not all nucleotide–residue pair types occur in equal number. Arginine interactions, for example, are much more numerous than glutamic acid interactions. A large bias towards zero values are present in these features because most interactions do not involve all possible structural moieties – for example, residue interactions in the DNA minor groove will have all zero values for any major groove or base moiety features. In order to avoid this bias, feature values were clipped at 0.5 before computing percentiles.

Feature values are broken down for each nucleotide–residue pair type, and each structural moiety interaction type. The 20th percentiles of these distributions are then used as lower bounds for determining when to consider a structural moiety interaction. For each structural moiety interaction type (e.g. side chain–sugar) there are three cut-off values to consider – the number of hydrogen bonds, number of van der Waals interactions, and the buried solvent accessible surface area (BASA) component. If any feature among the three meets the threshold for a moiety interaction, then the nucleotide–residue interaction pair is assigned that moiety interaction. Supplementary Table S1 shows 20th percentiles for interaction features of arginine interactions with the standard four nucleotides. If a nucleotide–residue interaction fails to meet the cut-offs for any moiety interaction pair, then the feature value with the largest ratio to its cut-off value is chosen to assign a moiety interaction, and the nucleotide–residue interaction is classified as a “weak” interaction.

Tyrosine interaction motif analysis

To examine the minor groove interactions of the residue tyrosine (as shown in Figure 5A), we compiled all tyrosine–nucleotide interactions which occur in the minor groove of a helical region of DNA. Using the DNAproDB data, we found all instances of a tyrosine–nucleotide interaction via the `interface.nucleotide-residue_interaction` feature arrays which list all nucleotide–residue interactions in a given interface, and filtered for those with a `nuc_secondary_structure` feature value of “helical” and which included a minor groove interaction moiety (see *Identification of structural and interaction moieties)*. For each tyrosine, we first noted how many nucleotides it interacted with simultaneously. To simplify the analysis, we kept only the subset of tyrosine which interact with exactly three nucleotides in the minor groove. The remaining list of interactions were then filtered for sequence redundancy of the tyrosine parent chain using the PDB-derived `protein.chain.sequence_clusters` features (see Supplementary Figure S2), keeping up to 15 tyrosine examples per 90% sequence cluster, resulting in 99 total examples. For each example the nucleotides were sorted by the center of mass distance from the tyrosine and placed into bins 1, 2 and 3 of a two-dimensional histogram, with the second index indicating the nucleotide type (A,C,G,T).

PCA analysis of interface features

A Principal Component Analysis (PCA) was performed on 134 DNAproDB features (or features derived from them) describing the DNA–protein interface for 4,758 different interfaces as shown in Figure 5B. Each interface used was that of a single protein chain interacting with a single DNA entity; DNAproDB breaks down every DNA–protein interface by protein chain under the `interface.interface_features` feature array (see Supplementary Figure S2). The features used describe characteristics such as geometry of the protein surface, BASA and hydrogen bond statistics, nucleotide–residue interaction geometries and residue propensities, and were concatenated to create a dense feature vector. Every vector was then projected along the eigenvectors corresponding to the first and second principal component axes and the projected components were plotted. Each interface was then grouped according to the GO annotations of the protein chain (available under `protein.chains` features) into one of four broad classes – transcription factor activity, DNA repair, DNA recombination, and single-stranded DNA binding.

Nucleotide–residue stacking probability analysis

To generate the stacking probabilities shown in Figure 5C, each residue with a planar side chain (arginine, phenylalanine, tyrosine, tryptophan, histidine, asparagine, aspartic acid, glutamine and glutamic acid), all instances of a base interaction (determined via the interaction moiety feature under the `interface.nucleotide-residue_interactions` feature array; see *Identification of structural and interaction moieties*) with a nucleotide in a single-stranded DNA conformation were gathered from the DNAproDB dataset (4,509 entries), and the conditional probability for that residue–nucleotide interaction pair to form a base-stacking geometry was computed in the following way

$$P(\text{stack}|N,R) = \frac{P(\text{stack},N,R)}{P(N,R)}$$

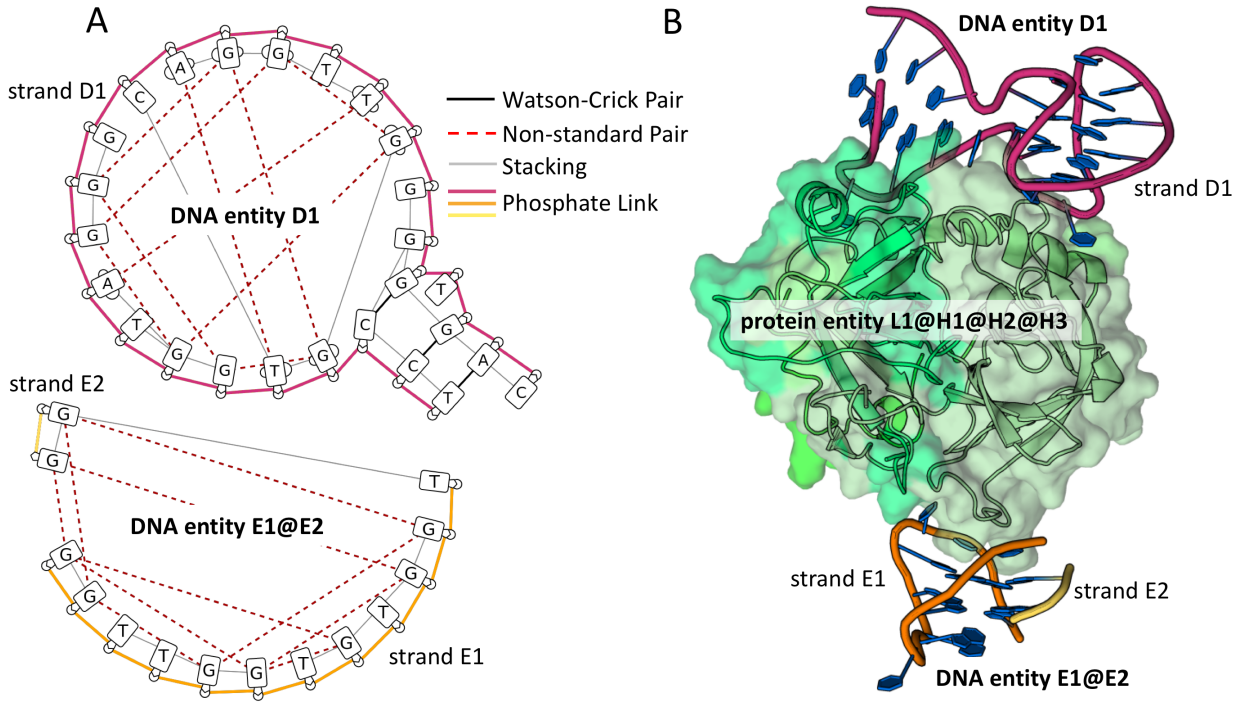
where

$$P(\text{stack},N,R) = \frac{n(g=\text{stack},N,R)}{\sum_{g,N,R} n(g,N,R)}$$

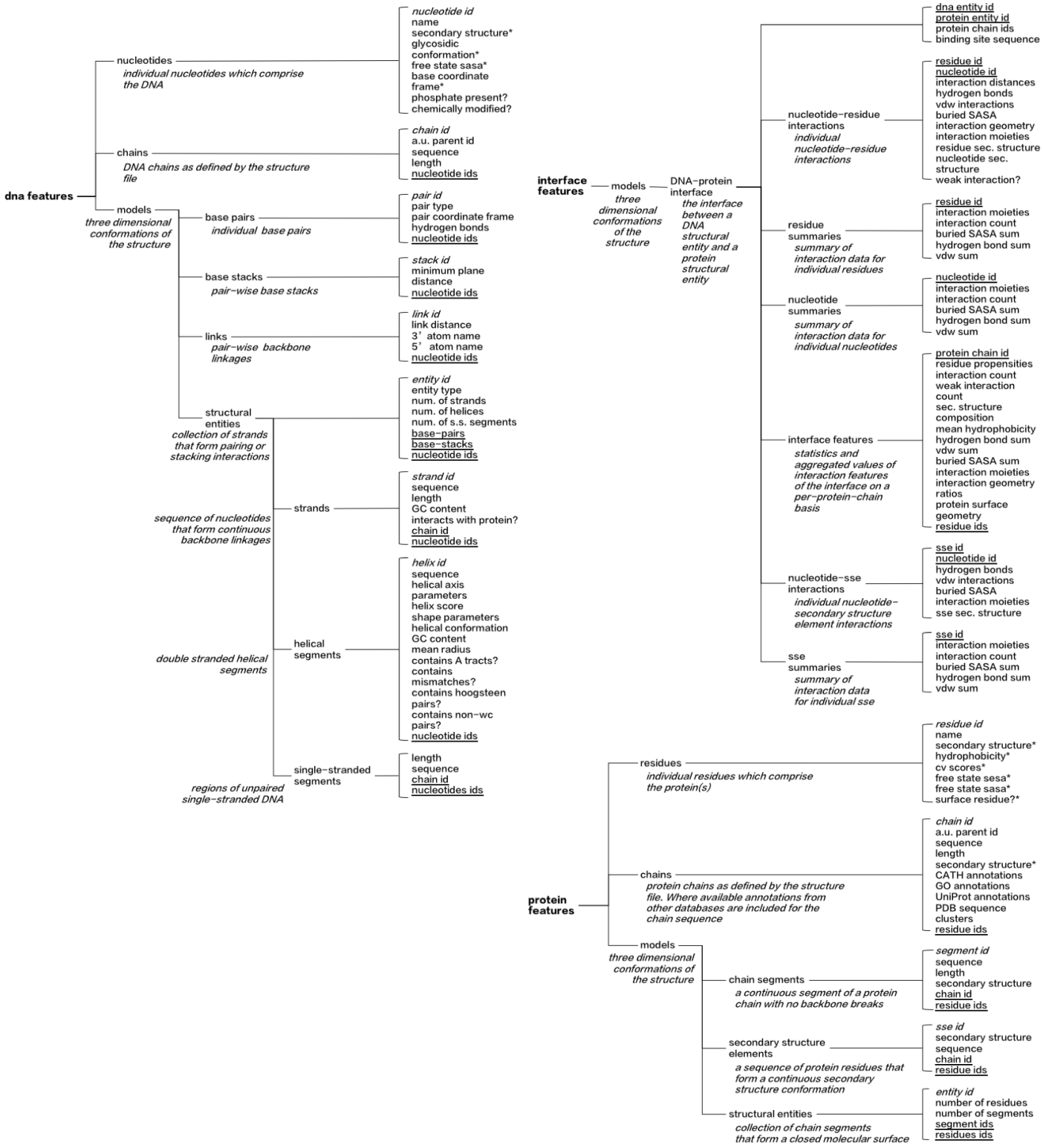
is the joint probability for an interaction between a residue R and a nucleoide N to be in a geometry $g \in [\text{stack}, \text{pseudo-pair}, \text{other}]$ (interaction geometry was determined by the program SNAP) (7,8) with $g = \text{stack}$, and $n(g,N,R)$ is the number of residue–nucleotide interactions between N and R in the dataset with a stacking geometry g. The prior probability is given by

$$P(N,R) = \frac{\sum_g n(g,N,R)}{\sum_{g,N,R} n(g,N,R)}$$

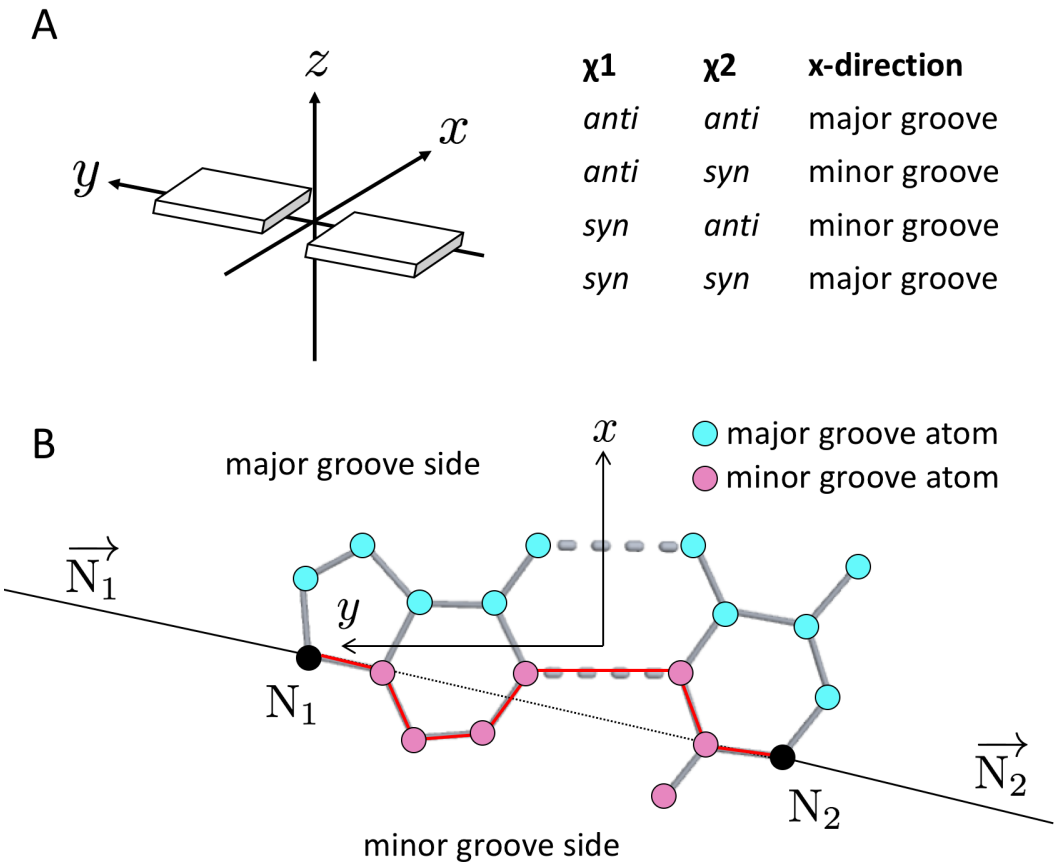
SUPPLEMENTARY FIGURES



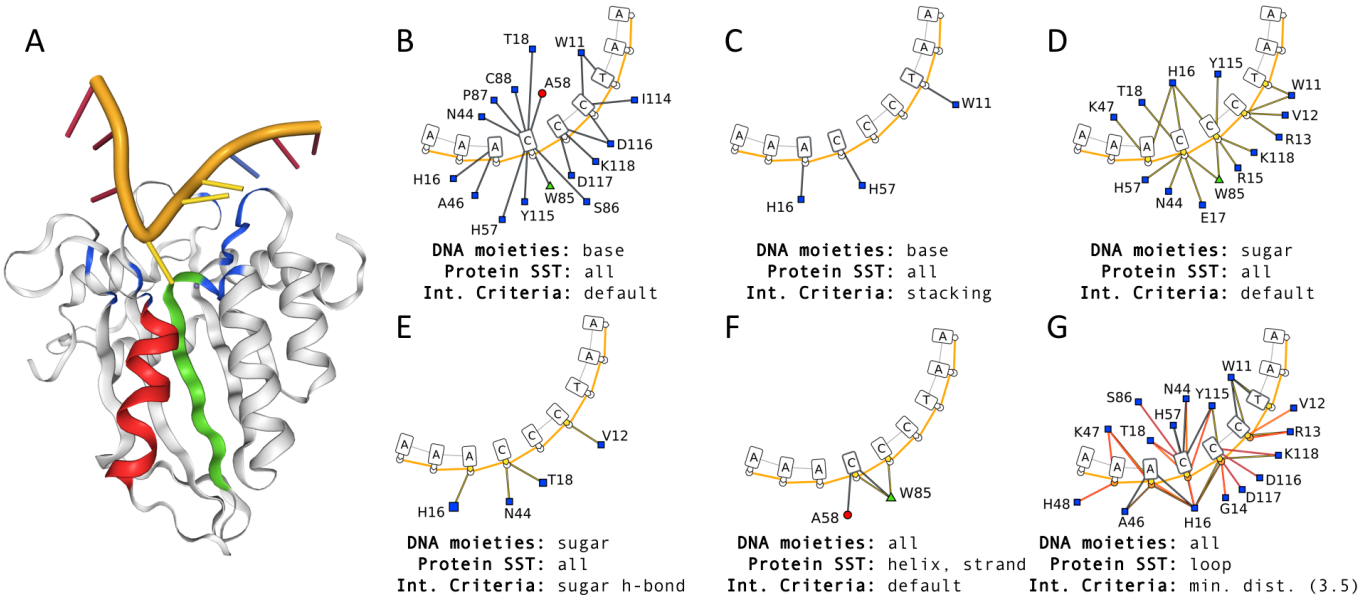
Supplementary Figure S1. An example of a DNA–protein complex which contains two DNA structural entities and one protein structural entity. **A.** A graph showing nucleotide base pairing, base stacking and sugar-phosphate linkages which has been stylized using DNAproDB visualization tools. Two disparate sub-graphs can be seen which represent the two discrete DNA structural entities seen in the second panel. DNA structural entities are named based on the DNA strands they are composed of – their identifier is simply a concatenation of their component strand identifiers. The DNA strand identifiers are based on the DNA chain which the strand belongs to. The first strand in chain E is named “E1”, the second strand “E2” etc. In this structure, the DNA chain E has a break (due to a missing nucleotide) thus forming two strands. **B.** The three-dimensional structure which shows the two discrete DNA structural entities and the single discrete protein structural entity. This protein entity has two chains – chain L and H which form a heterodimer, however chain H consists of three chain segments, H1, H2 and H3 which are due to backbone breaks. The four chain segments form a single closed molecular surface, and hence are considered a single structural entity. Protein structural entities are named in the same way that DNA entities are.


dna features
nucleotides
individual nucleotides which comprise the DNA
nucleotide id
name
secondary structure*
glycosidic conformation*
free state sasa*
base coordinate frame*
phosphate present?
chemically modified?
chains
DNA chains as defined by the structure file
chain id
a.u. parent id
sequence
length
nucleotide ids
models
three dimensional conformations of the structure
base pairs
individual base pairs
pair id
pair type
pair coordinate frame
hydrogen bonds
nucleotide ids
base stacks
pair-wise base stacks
stack id
minimum plane distance
nucleotide ids
links
pair-wise backbone linkages
link id
link distance
3' atom name
5' atom name
nucleotide ids
structural entities
collection of strands that form pairing or stacking interactions
entity id
entity type
num. of strands
num. of helices
num. of s.s. segments
base-pairs
base-stacks
nucleotide ids
strands
sequence of nucleotides that form continuous backbone linkages
strand id
sequence
length
GC content
interacts with protein?
chain id
nucleotide ids
helical segments
double stranded helical segments
helix id
sequence
helical axis parameters
helix score
shape parameters
helical conformation
GC content
mean radius
contains A tracts?
contains mismatches?
contains hoogsteen pairs?
contains non-wc pairs?
nucleotide ids
single-stranded segments
regions of unpaired single-stranded DNA
length
sequence
chain id
nucleotides ids
interface features
models
three dimensional conformations of the structure
DNA-protein interface
the interface between a DNA structural entity and a protein structural entity
dna entity id
protein entity id
protein chain ids
binding site sequence
nucleotide-residue interactions
individual nucleotide-residue interactions
residue id
nucleotide id
interaction distances
hydrogen bonds
vdw interactions
buried SASA
interaction geometry
interaction moieties
residue sec. structure
nucleotide sec. structure
weak interaction?
residue summaries
summary of interaction data for individual residues
residue id
interaction moieties
interaction count
buried SASA sum
hydrogen bond sum
vdw sum
nucleotide summaries
summary of interaction data for individual nucleotides
nucleotide id
interaction moieties
interaction count
buried SASA sum
hydrogen bond sum
vdw sum
interface features
statistics and aggregated values of interaction features of the interface on a per-protein-chain basis
protein chain id
residue propensities
interaction count
weak interaction count
sec. structure composition
mean hydrophobicity
hydrogen bond sum
vdw sum
buried SASA sum
interaction moieties
interaction geometry
ratios
protein surface geometry
residue ids
nucleotide-sse interactions
individual nucleotide-secondary structure element interactions
sse id
nucleotide id
hydrogen bonds
vdw interactions
buried SASA
interaction moieties
sse sec. structure
sse summaries
summary of interaction data for individual sse
sse id
interaction moieties
interaction count
buried SASA sum
hydrogen bond sum
vdw sum
protein features
residues
individual residues which comprise the protein(s)
residue id
name
secondary structure*
hydrophobicity*
cv scores*
free state sesa*
free state sasa*
surface residue?*
chains
protein chains as defined by the structure file. Where available annotations from other databases are included for the chain sequence
chain id
a.u. parent id
sequence
length
secondary structure*
CATH annotations
GO annotations
UniProt annotations
PDB sequence clusters
residue ids
models
three dimensional conformations of the structure
chain segments
a continuous segment of a protein chain with no backbone breaks
segment id
sequence
length
secondary structure
chain id
residue ids
secondary structure elements
a sequence of protein residues that form a continuous secondary structure conformation
sse id
secondary structure
sequence
chain id
residue ids
structural entities
collection of chain segments that form a closed molecular surface
entity id
number of residues
number of segments
segment ids
residues ids

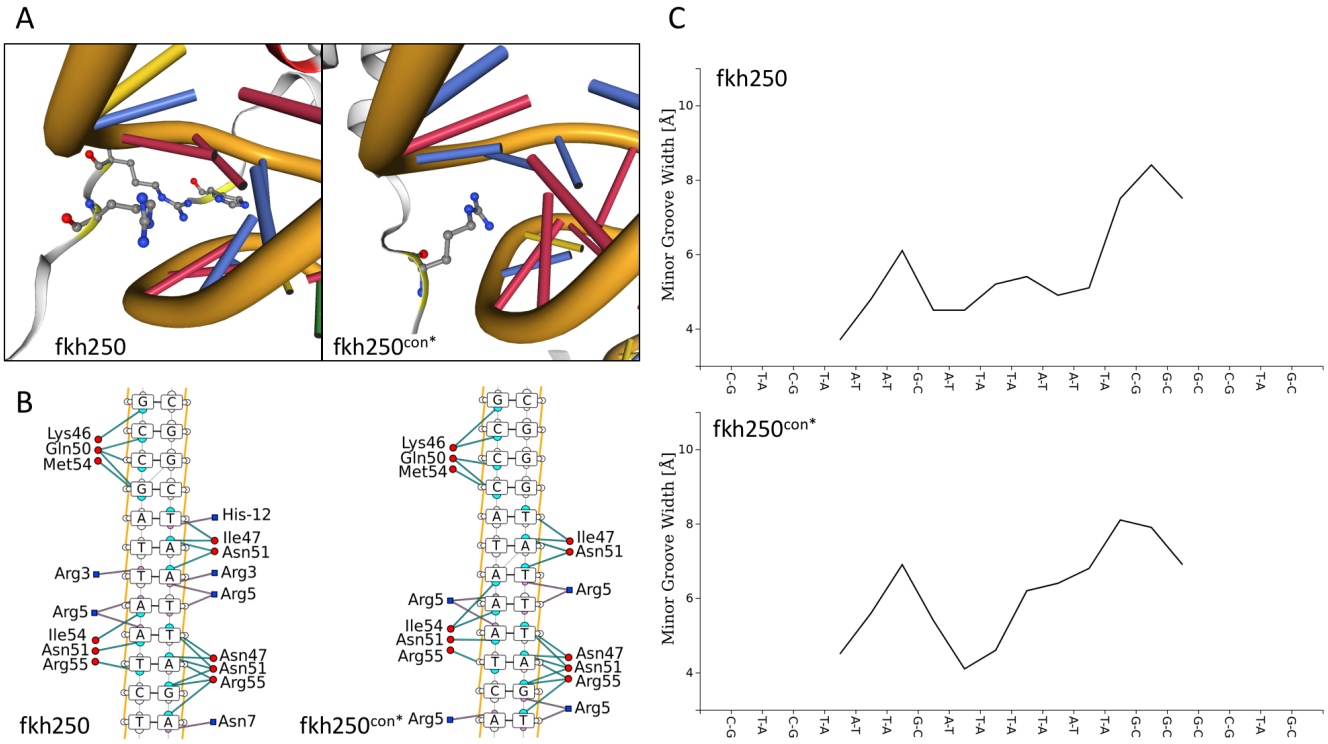
Supplementary Figure S2. An overview of the DNAproDB feature hierarchy. Features are grouped into three main categories – DNA specific features, protein specific features, and interface specific features. Within each category there are two levels of features – entry-level and model-level. Entry level features are those which can be described at the level of the entry (i.e. the structure), and do not vary across models. For instance under protein features, chains are an entry level feature because the number of chains and most of their properties (e.g. chain identifier, sequence, sequence-based annotations, length etc.) do not vary or depend on the coordinates assigned in a particular model. Any feature within an entry-level feature branch that does depend on the model will be stored as an array with one element per model. For example, the secondary structure feature of a protein chain can vary from model to model since secondary structure depends on the residue coordinates. Features under an entry-level branch which depend on the model index are noted with an asterisk in the figure. Model-level features are those which depend on and may change with the three-dimensional coordinates of the structure or represent a data object which may only exist for a particular model. These include items such as DNA base pairs, all interface properties, and protein secondary structure elements. Some features refer to other features which are used as identifiers. An identifier feature is analogous to a primary key in a relational database setting and are italicized in the figure. Underlined features refer to identifier features.



Supplementary Figure S3. A graphical summary of our procedure for determining major and minor groove structural moieties. **A.** The base-pair coordinate frame used by DSSR, from which we gather base-pairing information. Depending on the glycosidic torsion angle conformation of the two bases, the x-direction will either point towards the minor groove or major groove. In Watson-Crick geometry, it will always point towards the major groove. **B.** An illustration of our procedure for defining major and minor groove atoms. An A/T Watson-Crick base pair is used for illustration purposes. In this case, both bases are in the *anti* conformation, so the x-direction of the base-pair coordinate frame is towards the major groove. The coordinates of the base atoms are projected to the x-y plane of the base-pairing frame, and the shortest path from the glycosidic nitrogen atom of both bases is found (shown in red) using the N1-N3 hydrogen bond as an edge joining the bases. This path in addition to the rays emanating from the glycosidic nitrogen atoms bisect the plane. All atoms lying on the path and on the minor groove side of the bisection are classified as minor groove, and atoms on the other side as major groove.



Supplementary Figure S4. An example of different ways to filter nucleotide–residue interactions in DNAproDB visualizations using a DNAproDB entry containing the catalytic domain of APOBEC3G bound to a single-stranded DNA oligomer (PDB ID 6BUX) (9). The visualizations in panels **B**–**G** are examples of the *residue contact map* visualization. **A.** The three-dimensional structure as depicted by NGL viewer (10,11). **B.** Default DNAproDB criteria (4.5 Å minimum distance, interactions not weak; see Supplementary Methods) with DNA base interactions shown and all protein secondary structures. Many interactions are shown involving the first cytosine nucleoside which is inserted into the active site's binding pocket of the APOBEC3G domain. **C.** The same criteria as in **B** but showing only residues which form stacking interactions. **D.** Interactions are shown using default criteria but only to the DNA sugar moieties indicated by the yellow interaction lines which connect to the sugar moiety symbol of each nucleotide. **E.** The same as in **D** except now involving only interactions with at least one sugar hydrogen bond. **F.** Here we are visualizing all DNA moiety interactions using default criteria but only for helix or strand residues. Only two residues are shown, and neither make any phosphate interactions. This is because these residues are in the active site, which is deeply buried in the protein and is mainly accessible by the inserted cytosine base with some sugar interactions. **G.** Here all DNA moiety interactions are shown but only for loop residues. The interaction distance cut-off has been lowered to a 3.5 Å minimum nucleotide–residue distance, and any interaction involving a hydrogen bond is highlighted with a red outline.



Supplementary Figure S5. Visualizations from DNAproDB for a heterodimer of the Hox protein Sex combs reduced (Scr) and its cofactor Extradenticle (Exd) bound to two different DNA fragments (PDB IDs 2R5Z and 2R5Y) (12). Only major groove and minor groove contacts are shown. Joshi *et al.* (12) showed that for this protein complex Scr N-terminal linker residues Arg3 and His−12 are important for conferring sequence specificity via shape recognition of the minor groove. **A.** Three-dimensional views of the minor groove in the region of the Scr Arg5 linker residue. The left panel is an Scr *in vivo* binding site (PDB ID 2R5Z) in which the Arg3 and His−12 residues can be seen in the minor groove. On the right is a Hox consensus site which lack the Arg3 and His−12 contacts. **B**. Two residue contact maps showing major groove and minor groove contacts for the Scr−Exd heterodimer bound to the Scr *in vivo* binding site fkh250 on the left (PDB ID 2R5Z) and a Hox consensus site fkh250^{con*} on the right (PDB ID 2R5Y). The colored markers indicate residues and their secondary structure − helix residues are represented as red circles and linker residues are represented as blue squares. Residues are grouped into SSEs and markers on each nucleotide represent the major and minor groove contacts, respectively. The Scr residues Arg3 and His−12 are seen making contacts in the DNA minor groove of the *in vivo* binding site but cannot be seen contacting the Hox consensus site. **C.** Shape overlay plots of the minor groove width of the two binding sites, fkh250 and fkh250^{con*}. The differences in the intrinsic shape profile of these DNA sequences, which are described in (12), explain the preference for the Scr *in vivo* binding site.

SUPPLEMENTARY TABLES

	feature	pp/mc	pp/sc	sr/mc	sr/sc	wg/mc	wg/sc	sg/mc	sg/sc	bs/mc	bs/sc
A (23227)	h. bond	1	1	1	1	1	1	1	1	1	1
	vdw	1	1	1	1	1	1	1	1	1	1
	basa	1.0	4.4	1.0	5.8	0.9	3.2	0.8	3.4	1.3	9.7
C (19559)	h. bond	1	1	1	1	1	1	1	1	1	1
	vdw	1	1	1	1	1	1	1	1	1	1
	basa	0.9	5.6	1.2	4.5	1.0	3.7	0.9	1.9	2.0	4.8
G (27427)	h. bond	1	1	1	1	1	1	1	1	1	1
	vdw	1	1	1	1	1	2	1	1	1	1
	basa	0.8	3.9	1.4	4.7	0.8	5.3	0.9	3.5	0.9	7.7
T (24547)	h. bond	1	1	1	1	1	1	1	1	1	1
	vdw	1	1	1	1	1	1	1	1	1	1
	basa	1.0	4.9	1.2	3.8	0.8	7.0	0.8	2.3	1.3	7.5

Supplementary Table 1. Interaction feature cut-off values for arginine interactions with the standard four DNA nucleotides. The numbers in parentheses are the total number of interactions used to calculate 20th percentiles of the feature value distributions, which are then used as cut-off values. Note that these cut-off values are inclusive, and the feature values were clipped at 0.5 before computing percentiles to avoid a large bias towards zero. The first column indicates the feature type – h. bond is the total number of hydrogen bonds for a moiety interaction, vdw is the number of van der Waals interactions, and basa is the buried solvent accessible surface area. The remaining columns are the possible moiety interactions with the following abbreviations: pp – DNA phosphate; sr – DNA sugar; wg – DNA major groove; sg – DNA minor groove; bs – DNA base; mc – protein main chain; sc – protein side chain.

AUTHOR CONTRIBUTIONS

J.M.S., H.M.B., and R.R. conceived DNAproDB. J.M.S. designed DNAproDB conceptionally, with the help of H.M.B. and R.R., and developed all components of the processing pipeline, database and server functions. J.M.S. derived structural features and designed the report pages, and the web interface. N.M. implemented the 3D structure viewer functionality on report pages using the NGL web application. J.M.S. generated examples for DNAproDB analyses and further studies. J.M.S. wrote the manuscript with the help of H.M.B. and R.R.

SUPPLEMENTARY REFERENCES

1. Lu, X.J., Bussemaker, H.J. and Olson, W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142.
2. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. and Trout, B.L. (2010) Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, **114**, 6614-6624.
3. Westbrook, J.D., Shao, C., Feng, Z., Zhuravleva, M., Velankar, S. and Young, J. (2015) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics*, **31**, 1274-1278.
4. Hubbard, S.J. and Thornton, J.M. (1993) 'NACCESS', computer program. Department of Biochemistry and Molecular Biology, University College London.
5. Batsanov, S.S. (2001) Van der Waals Radii of Elements. *Inorganic Materials*, **37**, 871-885.
6. Sagendorf, J.M., Berman, H.M. and Rohs, R. (2017) DNAproDB: an interactive tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.*, **45**, W89-W97.
7. Lu, X.J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108-5121.
8. Lu, X.J. and Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213-1227.
9. Maiti, A., Myint, W., Kanai, T., Delviks-Frankenberry, K., Sierra Rodriguez, C., Pathak, V.K., Schiffer, C.A. and Matsuo, H. (2018) Crystal structure of the catalytic domain of HIV-1 restriction factor APOBEC3G in complex with ssDNA. *Nat. Commun.*, **9**, 2460.
10. Rose, A.S. and Hildebrand, P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576-W579.
11. Rose, A.S., Bradley, A.R., Valasatava, Y., Duarte, J.M., Prlic, A. and Rose, P.W. (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755-3758.
12. Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B. and Mann, R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, **131**, 530-543.