Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz889

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

Received August 15, 2019; Revised September 22, 2019; Editorial Decision September 28, 2019; Accepted October 01, 2019

ABSTRACT

DNAproDB (https://dnaprodb.usc.edu) is a web-based database and structural analysis tool that offers a combination of data visualization, data processing and search functionality that improves the speed and ease with which researchers can analyze, access and visualize structural data of DNA–protein complexes. In this paper, we report significant improvements made to DNAproDB since its initial release. DNAproDB now supports any DNA secondary structure from typical B-form DNA to single-stranded DNA to G-quadruplexes. We have updated the structure of our data files to support complex DNA conformations, multiple DNA–protein complexes within a DNAproDB entry and model indexing for analysis of ensemble data. Support for chemically modified residues and nucleotides has been significantly improved along with the addition of new structural features, improved structural moiety assignment and use of more sequence-based annotations. We have redesigned our report pages and search forms to support these enhancements, and the DNAproDB website has been improved to be more responsive and user-friendly. DNAproDB is now integrated with the Nucleic Acid Database, and we have increased our coverage of available Protein Data Bank entries. Our database now contains 95% of all available DNA–protein complexes, making our tools for analysis of these structures accessible to a broad community.

INTRODUCTION

Analyzing structures of DNA–protein complexes provides valuable insight into the physical mechanisms that drive fundamental biological processes such as chromatin structural organization and DNA transcription, replication and repair. Atomic resolution models of proteins bound to their cognate DNA-binding sites help to elucidate the relationships among sequence, structure and biological function, distinguish different mechanisms of DNA recognition (1), offer a deeper physical understanding of existing experimental results and give insights into the molecular machineries that drive living cells (2). The Protein Data Bank (PDB) (3) is an archival repository that currently contains ∼4800 structures of proteins bound to DNA (excluding those also containing RNA). These structures vary widely with respect to many features: molecular and biological function of the DNA-binding proteins, tertiary and secondary structure of the bound DNA, protein and DNA sequence and structure size.

A number of databases have been developed that provide data for structures of DNA–protein complexes from the PDB. PDIdb (4) is a database that provides information on effective atomic interactions for each DNA–protein interface in a complex and classifies proteins by function and structure. Users can search the database for entries based on features of the interface, DNA or protein. The 3D-footprint database (5) provides structure-based binding specificities for all DNA–protein complexes in the PDB and static figures that display DNA–protein interactions in the complexes. NPIDB (6) contains structural information on DNA–protein and RNA–protein complexes and computes hydrogen bonds, water bridges and hydrophobic interactions. PDBsum (7) is a database that summarizes the contents of each macromolecular structure deposited in the PDB, including DNA–protein complexes, and provides various analysis tools. The Nucleic Acid Database (NDB) (8) provides detailed structural annotations on nucleic acid structure and annotates protein function and can be searched for DNA–protein complexes.

DNAproDB (9) is a database, structure processing pipeline and web-based visualization tool designed to aide structural analysis of DNA–protein complexes, visualize features of DNA–protein interactions and generate struc-

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

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

tural data sets that meet specific criteria based on a variety of biophysical features and annotations. DNAproDB is unique in its combination of rich structural features and annotations, detailed structure, function and sequence-based search capabilities, and interactive, customizable visualizations. It is built on an automated end-to-end structure processing pipeline that is designed to handle the complexity inherent in structural data. The pipeline takes as input the atomic coordinates of a 3D DNA–protein complex, extracts from it a large variety of structural and biophysical features and combines these data in a hierarchical data format.

Features include information such as structural annotations (e.g. identifying a region of DNA as single-stranded or double-helical, and identifying protein secondary structure elements), sequence information (e.g. GC-content and identification of A-tracts) and statistical information about the DNA–protein interfaces in the complex (e.g. residue propensity and interface hydrophobicity). Information about individual nucleotide–residue interactions is also provided, such as hydrogen bonding, interaction geometry (based on SNAP (10)), buried solvent accessible surface areas and identification of the interacting residue and nucleotide moieties (see 'Identification of structural and interaction moieties'). DNAproDB includes annotations from other databases where available, such as UniProt (11), CATH (12) and the Gene Ontology knowledge base (13,14) that enhance the search capability of our database and integration with the biological community. The DNAproDB database provides data for over 4500 PDB entries that contain DNA–protein complexes. Our database can be searched for entries meeting specific criteria based on features of the DNA, protein or DNA–protein interactions. Data for the resulting entries can then easily be downloaded in JSON format (15) and parsed offline or explored and visualized with in-browser visualization tools provided through the DNAproDB website (https://dnaprodb.usc.edu). We also offer an option for users to upload and process individual structures using the same processing pipeline used to build our database and to analyze their structure with a privately generated report page (see 'Structure reports' below).

In this report, we describe major improvements and upgrades to DNAproDB that broadly improve its utility, scope and the raw amount of data available to users. The original version of DNAproDB, as described in (9), was designed to process structures of proteins bound to double-stranded helical DNA. Our newly updated processing pipeline and database can accommodate any DNA structure from double-stranded B-form DNA to single-stranded DNA to G-quadruplexes, and now support multiple DNA–protein complexes per DNAproDB entry. This and other improvements have increased our coverage of PDB structures containing DNA–protein complexes from 51% of available structures in the original version of DNAproDB to >95% in the current version. Accordingly, our visualization tools have been completely redesigned to handle a much wider variety of DNA–protein structures. In the process of implementing these changes, we have made many additional improvements both to our structure processing capabilities, data format and to the visualization tools and website interface that are described in the following sections.

STRUCTURE PROCESSING UPGRADES

The DNAproDB structure-processing pipeline (Figure 1) takes as input the atomic coordinates of DNA–protein complexes in PDB or mmCIF (16) format and extracts from them a variety of structural and biophysical features. These features describe aspects of the DNA, protein(s) and DNA–protein interactions observed in the structure. The pipeline is an end-to-end processing work flow, and the resulting data that are generated from an input structure are combined in a single document that we refer to as DNAproDB data file. One input structure produces one output data file, and if a structure contains multiple DNA–protein complexes then data for each complex are contained in the same data file. The data files are in JSON format (15), and a description of their content is given below. Our database is built from a collection of these data files and their structure is designed to be convenient for the user who wishes to use our data for their own analysis, enable elegant search capabilities of our database and allow for detailed reporting and visualization using our web-based analysis tools. Below we describe various notable aspects of DNAproDB that have been improved, added or modified in our latest revision.

Structure preprocessing

DNAproDB performs several preprocessing steps on structures before attempting to calculate features from them. These preprocessing steps may result in some minor differences in the resulting structure and those available for download from the PDB (or the original data in the case of an uploaded structure), so we note our preprocessing procedure here for clarity and completeness.

First, DNAproDB automatically generates biological assemblies from the asymmetric unit using the symmetry operations provided by each processed entry. Often a biological assembly may be identical to the asymmetric unit, although some may contain multiple copies of the asymmetric unit or only part of it. Therefore, multiple copies of a DNA or protein chain may occur in a biological assembly and DNAproDB assigns a unique identifier to each one. The identifier of the parent chain in the asymmetric unit is recorded in the DNAproDB data file. Currently, only one biological assembly per structure is used (users may upload any alternate biological assembly to our web server and generate a report for it if they wish). For structures that are uploaded to our web server, the provided coordinates are assumed to already be those of the biological assembly, and no symmetry operations are applied. Multiple models may exist within a biological assembly that represent different conformations of the assembly. This is common for NMR structures and is useful for analyzing snapshots of a simulation such as a molecular dynamics trajectory. Structures obtained from X-ray crystallography generally will only have one model.

Next, DNAproDB removes components of the structure that are not part of the protein, DNA, solvent or a known coordination center such as a zinc ion. Any small ligands or other chemical entities that are not chemically derived from an amino acid or nucleic acid are removed and ignored. Additionally, components that are missing too many atoms,

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

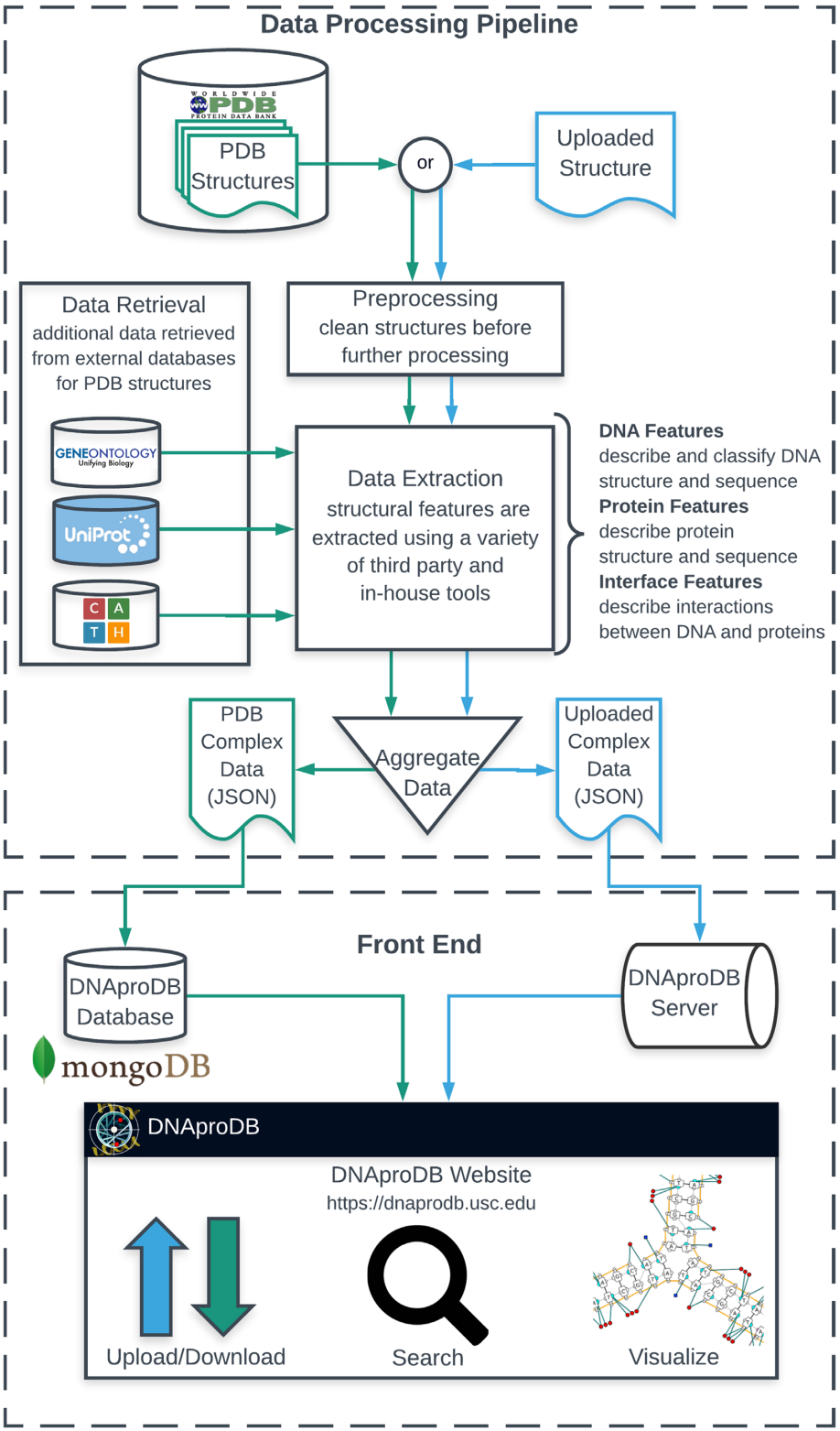

Figure 1. A schematic overview of the DNAproDB structure processing pipeline and front-end interface. The main stages are structure preprocessing where structures are prepared for processing, feature extraction where various biophysical features are calculated, data retrieval that pulls additional annotations from existing databases for protein chains, and data aggregation where features are combined in a standard data format. The DNAproDB database stores processed structural data for >4500 PDB entries containing DNA–protein complexes. Users can search the database on features of the DNA, protein or DNA–protein interactions, can generate reports for the returned results and can upload their own structures for private analysis. The report page contains functionalities for downloading extracted features as a JSON file (15) and for visualizing data using interactive visualization tools that can export static figures.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

clash with other components or do not appear across all models are removed. Any removed component and the reason why it was removed are recorded in the DNAproDB data file.

Hydrogen atoms are added to all structures using the program Reduce (17). For structures that are uploaded to our web server by a user, we provide an option to repair any missing heavy atoms using the program PDB2PQR (18,19); however, we do not add missing heavy atoms to structures used to build our database.

DNAproDB data overview

The features DNAproDB extracts from structures are organized in a hierarchical manner with three main feature categories being DNA features, protein features and DNA–protein interface features (Figure 2). We have significantly improved and updated the way we organize DNAproDB data files, and we now include many more fields and support model-level features, meaning features that can vary from one model in the structure to another. Figure 2 shows an overview of the conceptual hierarchy that defines how features are organized in DNAproDB.

Two important concepts new to DNAproDB are that of a DNA structural entity and a protein structural entity. DNA structural entities are a collection of DNA strands that are connected via base pairing or base stacking. We define a DNA strand as a collection of nucleotides that are connected via a continuous set of phosphodiester sugar-phosphate bonds with no backbone breaks. A DNA structural entity thus forms a discrete structural component of the overall structure. Base pairing and base stacking between nucleotides are identified using the program DSSR (20). DNA structural entities can be thought of as undirected connected subgraphs, where every node corresponds to a nucleotide and edges between nodes indicate either a base-pairing or a base-stacking interaction or phosphodiester bond. Many structures will contain only a single DNA entity, but some may contain several. We make a distinction between a DNA 'strand' and a DNA 'chain' (as defined by the input file in PDB/mmCIF format) because, while ideally a DNA strand and a DNA chain have a one-to-one correspondence, missing nucleotides or backbone breaks may produce several DNA strands within a chain. Similar to DNA structural entities, a protein structural entity is a collection of protein chain segments that interact with one another to form a closed molecular surface, where a protein chain segment is a continuous segment of a protein chain with no backbone breaks. Thus, each discrete closed protein molecular surface (as defined by the solvent excluded surface) in the structure corresponds to a protein structural entity. See Supplementary Figure S1 for an example of a structure with two DNA structural entities and one protein structural entity.

A DNA–protein complex is then defined as a DNA structural entity and a protein structural entity that are in sufficient proximity and share at least one nucleotide–residue interaction. Each DNA–protein interface described in a DNAproDB data file corresponds to one interacting DNA entity–protein entity pair. Descriptions of the interface are grouped into two sets of features: individual nucleotide–residue interaction features and global features of the interface, which include aggregations of some nucleotide–residue interaction features, geometrical features of the protein surface and other statistical descriptions. Note that a nucleotide–residue pair is considered an 'interaction' based on the minimum distance between them—DNAproDB uses a cutoff of 4.5 Å. For a detailed description of the DNAproDB data hierarchy and a list of features currently provided, see Supplementary Figure S2.

DNA secondary structure

The original release of DNAproDB included only DNA–protein complexes where the DNA was in a helical, double-stranded conformation. This is the most common DNA structural conformation that accounts for roughly 74% of DNA–protein complexes currently available in our database (see Figure 3). Additionally, the original release could not support structures with more than one double-stranded helix. We have redesigned the way DNAproDB processes DNA structures in order to accommodate virtually any DNA secondary structure, to classify DNA structural entities (see 'DNAproDB data overview' for a description of this term) based on their secondary structure, and to support an arbitrary number of both DNA structural entities and protein structural entities within a structure. These improvements have not only vastly increased our total coverage of currently available PDB entries containing DNA–protein complexes (currently 95% coverage), but also greatly enriched the diversity of DNA-binding proteins and DNA structural conformations that we now provide data for, and we have developed many new features to support this increased diversity.

We provide coarse classifications of DNA structural entities based on their secondary structure. An entity can be classified as 'helical' (meaning double-stranded), 'single-stranded', 'helical/single-stranded' (referring to a helical conformation with at least one single-stranded overhang) or 'other' that encompasses a wide variety of conformations that are either irregular, unnamed or are not abundant enough to warrant their own class. For more details of how we classify structural entities, see Supplementary Methods.

Chemically modified components

DNAproDB now supports a much wider range of amino acids or nucleic acids that are chemical modifications of the standard 20 amino acids or 4 DNA nucleotides. A chemical modification can be a substitution of a chemical group such as the replacement of the terminal nitrogen with an oxygen atom in arginine citrullination or addition of a chemical group, such as the addition of the methyl group to the C5 atom of cytosine, forming 5-methylcytosine. The modified component must have an entry in the PDB Chemical Component Dictionary (21), and must not significantly deviate from its parent component so as to make identification of structural moieties (see below) ambiguous. DNAproDB requires a small amount of parameterization for the calculation of solvent accessible and excluded surface areas (van der Waals radii) and hydrophobicity calculations (residue hydrophobicity). Parameters for chemical modifications are

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

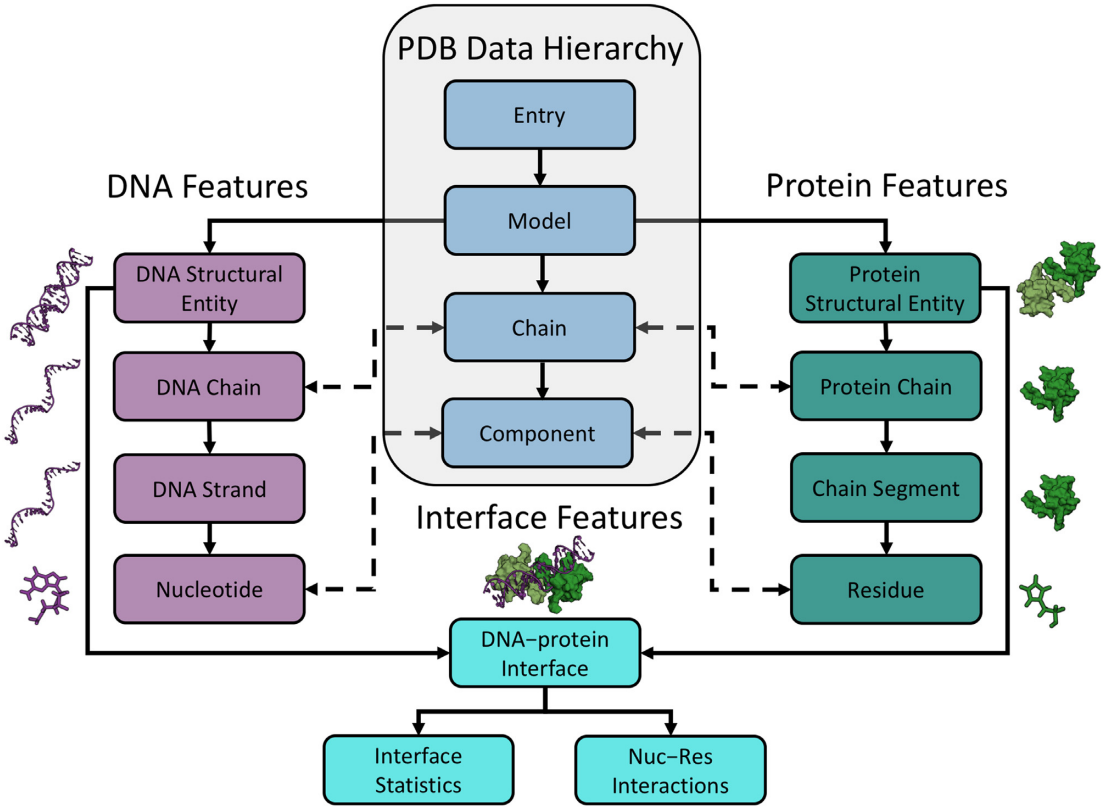

Figure 2. The data hierarchy used by DNAproDB, which inherits part of its overall structure from that used by the Protein Data Bank. Lines with a single arrow indicate one-to-many relationships, and lines with two arrows indicate one-to-one relationships. The PDB data hierarchy begins at 'entry', which is often referred to as simply 'the structure' or 'a structure' and contains all information about a macromolecule. An entry may have multiple 'models' that are different conformations of the structure (many entries contain only a single model). Models contain 'chains', which are generally polymerized chains of amino acids or nucleic acids and may also contain ligands, solvent or other small molecules. Finally a chain contains 'components', which are the molecular units that make up the chain—protein residues, DNA nucleotides or other small molecules. DNAproDB directly inherits the entry-model levels of the PDB hierarchy. From there, models contain structural entities (see 'DNAproDB data overview') that are distinct structural components within a particular model. DNA (protein) structural entities contain DNA strands (protein chain segments), which are derived from a DNA (protein) chain, but distinct in that a strand (chain segment) may not contain any backbone breaks. DNA strands (protein chain segments) then contain nucleotides (residues) at the lowest level of the hierarchy. Nucleotides and residues have a one-to-one correspondence with the component level of the PDB hierarchy, but structural entities have no direct correspondence. DNAproDB provides data on DNA–protein interfaces between every interacting DNA/protein structural entity pair, and these data are grouped into individual nucleotide–residue interaction features and global features of the interface. For a more detailed description of the DNAproDB data and features, see Supplementary Figure S2.

generally not available; however, we attempt to use approximate values where appropriate—therefore, some feature values may only be approximate for modified components. For a more detailed explanation of our approximation scheme, see Supplementary Methods.

Identification of structural and interaction moieties

A core and unique aspect of DNAproDB is the identification of DNA–protein interaction features broken down by both secondary structure and structural moiety. A structural moiety is a term we use to describe a chemical group or structural component of a nucleotide or residue that is distinguishable from the whole. Protein residues (amino acids) have two structural moieties—the main chain, which is the amine-carboyxl group that forms the protein backbone, and the side chain beginning at the α-carbon atom. With the addition of new DNA secondary structure conformations to DNAproDB, nucleotides now have either three or four structural moieties depending on their secondary structure. Helical (by which we mean double-stranded helices) DNA nucleotides have phosphate, sugar, and major groove and minor groove moieties. The groove moieties represent the edges of the paired bases that are exposed in the respective groove. Single-stranded and unclassified ('other') DNA nucleotides have phosphate, sugar and base moieties. Note that for helical DNA, the detection of the groove moieties has been improved to account for glycosidic bond angle and relative base orientation (see Supplementary Methods and Supplementary Figure S3).

For a given nucleotide–residue interaction pair, DNAproDB identifies interacting moieties within the pair. For example, given an adenine–arginine interaction, DNAproDB may identify that the interaction involves a sugar–side chain interaction and a minor groove–side chain interaction. Our procedure for identifying moiety interactions has been greatly improved in the newest release of DNAproDB. Moiety interactions are now determined by hydrogen bond, van der Waals interaction and buried solvent accessible surface area values. Cut-off values for these

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

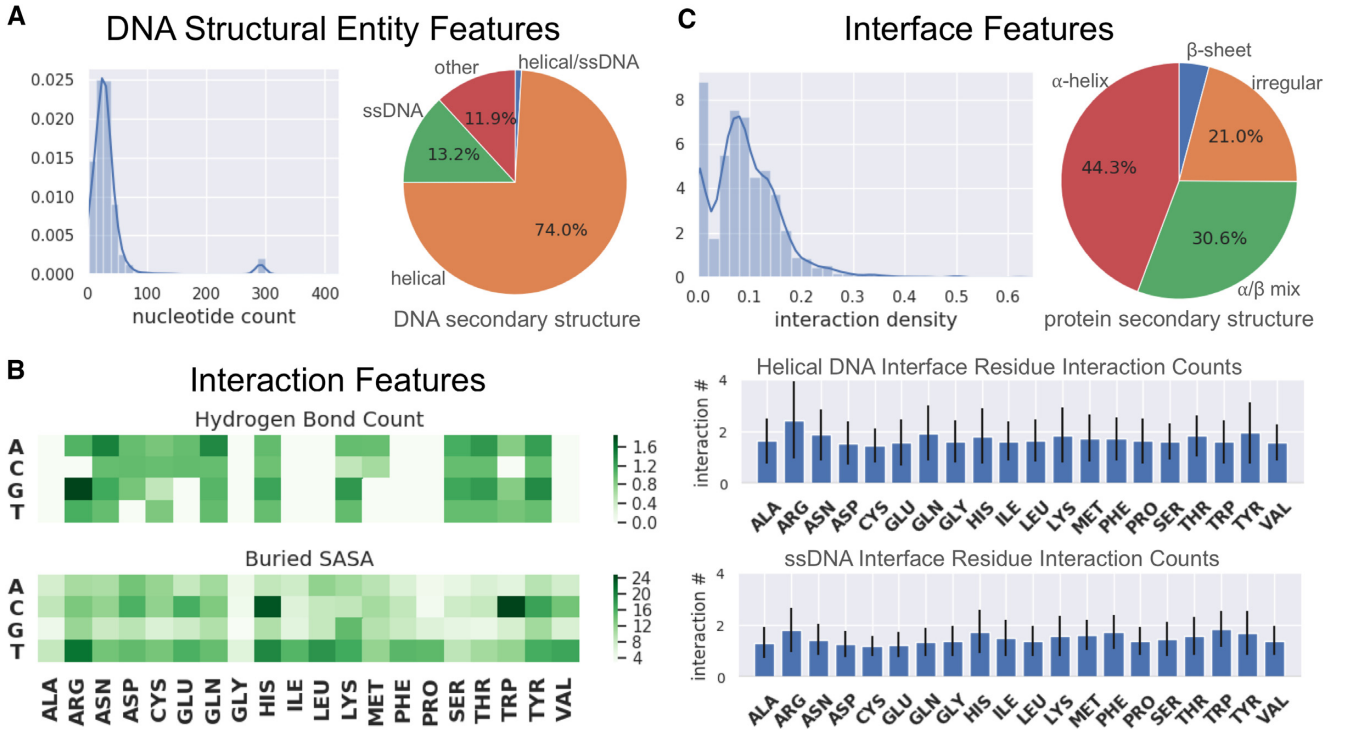

Figure 3. A variety of statistics and distributions for a set of selected DNAproDB features over the entire database. (**A**) Shown here are two features describing DNA structural entities: the number of nucleotides in the entity and the entity type. The mean nucleotide count is 34, with peaks at ∼35 and 300, the latter being mainly from nucleosome structures. The most common entity type is 'helical', making up 74% of all DNA structural entities, while single-stranded DNA ('ssDNA') is the next most common making up 13.2%. (**B**) This panel shows pair-wise nucleotide–residue interaction features. Both heat maps show values of interactions between the major groove moiety of nucleotides in a helical conformation with the side chains of each standard residue. The top heat map shows the mean number of major groove–side chain hydrogen bonds, and the bottom shows major groove–side chain buried solvent accessible surface area. Note that the hydrogen bond values for alanine, glycine, leucine, isoleucine, phenylalanine, proline and valine are all zero because these residues do not contain any hydrogen bond donors/acceptors in their side chains. The especially high value of guanine–arginine, major groove–side chain hydrogen bonds is due to the ability of arginine to form bidentate hydrogen bonds with the guanine's major groove moiety (when the guanine is in a Watson–Crick base-pairing conformation). (**C**) The four plots in this panel show various features that describe properties of DNA–protein interfaces as a whole. The first is the interaction density of the interface. This is the number of nucleotide–residue interactions divided by the number of nucleotides times the number of residues and measures how many interactions are present versus the total number possible and lies between zero and one. Most interfaces are rather sparse, which is typical of double-stranded helical DNA interfaces, while single-stranded DNA interfaces tend to be denser. The next plot shows the protein secondary structure composition of interfaces. The 'α-helix' label refers to interfaces that are made up primarily of α-helices, 'β-sheet' primarily of β-sheets etc. The α-helix and α-helix/β-sheet hybrid are the most common interface type, while a predominantly β-sheet composition is the least common. The final two plots compare the mean number of nucleotide interactions per residue for double-stranded helical DNA and single-stranded DNA interfaces. Single-stranded DNA interfaces overall tend to involve less nucleotide interactions per residue, which may indicate that individual residues contribute less to the overall binding affinity than for double-stranded helical DNA on average.

features for every nucleotide–residue pair are determined using the distribution of values among a large sample of nucleotide–residue interactions. Figure 3B shows mean values of hydrogen bonds and buried solvent accessible surface areas for major groove–side chain interactions for every nucleotide–residue pair type. For more details of how we identify and classify interaction moieties, see Supplementary Methods.

DATABASE AND WEB INTERFACE

The DNAproDB database is a document-oriented database built using MongoDB (22). At the time of publication, it provides data for 4509 PDB entries that contain at least one DNA–protein complex and is updated regularly as new PDB entries are released. Every entry in our database corresponds to an entry in the PDB and all the data we provide are encapsulated in a single JSON document for that entry. The structure of these data files is described in the proceeding sections and in Figure 2 and Supplementary Figure S2. Figure 3 shows various statistics and distributions of selected features that give a brief overview of the DNAproDB database. The DNAproDB database is quite heterogeneous, with structures ranging in size from only two nucleotides to several hundred, DNA tertiary and secondary structures ranging from single-stranded DNA to three-way holiday junctions, and proteins ranging from transcription factors to DNA recombinases. This heterogeneity is completely transparent to the user, however, as every entry follows a standard data format that has been designed to be flexible and to accommodate the wide variation that exists in the structural data set DNAproDB is built on.

Users can access the data in our database in several different ways. First, we offer the entire database for download as a flat file at our download page: https://dnaprodb.usc.edu/download.html. When uncompressed, each line of this file

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

contains a JSON document corresponding to one entry in the database. Second, users can search our database using our search page at https://dnaprodb.usc.edu/search.html. Here, users can construct a query based on a variety of features related to DNA, protein or DNA–protein interactions to generate a data set meeting a specific criterion or set of criteria. The user will be presented with a result page summarizing all the returned entries that meet the specified criteria, and from this page can download data for any or all of the returned entries. Third, users can download data for individual structures from the report page for those structures (see 'Structure reports' for a description of these pages). Finally, we provide a simple RESTful web API where advanced users can query our database directly without needing to use our web-interfaces. We refer interested readers to our documentation at https://dnaprodb.usc.edu/documentation.html for more information about using this API.

Searching the database

Users may search the DNAproDB database in a variety of different ways. The simplest search is to directly provide a list of PDB identifiers that will return all matching entries in our database from the supplied list. Additional features can be specified as a filter, and only structures that meet the additional criteria will be returned. More generally, a user will search for structures based on features describing structural and sequence characteristics rather than knowing the PDB identifiers in advance. Using our search form and combining different features, users can create powerful searches for structures based on characteristics of the DNA, protein or DNA–protein interactions. Our search form provides an easy user interface to build powerful queries on a subset of available features, and includes inputs for DNA features, protein features and DNA–protein interaction features.

Using our search form, DNA features can be included at several levels—entity level features that describe a DNA structural entity as a whole such as the entity type (e.g. single-stranded), strand-level features that are features of individual DNA strands, such as sequence motif or GC content, and helix level features that are features of helical segments within a DNA structural entity. Protein features include chain-level features such as sequence clusters, UniProt, CATH and GO annotations. These are extremely useful when wanting to include or exclude a particular protein or protein family from a search. Interaction features can be included in the search at the level of individual nucleotide–residue interactions or at the level of global interface properties. Much more complex and detailed searches are possible using the MongoDB query language in combination with our web API, but the searches that can be constructed using our search form cover the majority of use cases.

As one example of the type of searches DNAproDB supports, using the interaction feature inputs a user could search for structures where an arginine forms at least one hydrogen bond with a guanine in the major groove of the DNA and with the arginine belonging to an α-helical secondary structure. Alternatively, the user could simply search for structures where there are any contacts in the DNA major groove with a protein α-alpha helix. This search could be combined with DNA features, such as constraining that the DNA structural entity contains a helical segment between 8 and 20 base pairs in length and the helix conformation be in B-form. The user could also include one or more DNA sequence motifs to match on. Using protein features, we could further restrict this to return complexes containing homeodomain proteins that share a characteristic fold, by specifying the relevant CATH annotation as a protein chain feature. The DNAproDB database provides powerful search capabilities that no other structural database currently offers. Users can search the database to quickly retrieve data for a particular DNA–protein complex, discover new structures or generate data sets based on structural criteria.

Structure reports

DNAproDB presents the data available for any entry in our database or any structure uploaded to our server via a report page. The report page is the central web component of DNAproDB and allows users to explore, visualize and interact with data DNAproDB provides as well as view the structure in three dimensions using NGL viewer (23,24). The report page has been completely redesigned for the newest release of DNAproDB in order to support many new features and upgrades on the backend of our processing pipeline and database. Data are presented to the user based on their selection that is indexed by model, DNA structural entity and one or more protein chains within any protein entities interacting with the selected DNA entity. These selections allow users to step through the data in manageable ways and display as much or as little information as is relevant at one time.

The report page has three major components; the first are tables that display information about protein entities, chains and chain segments, DNA entities, strands, helices and single-stranded segments, and data about DNA–protein interfaces in the structure. The tables present data at the model level, updating any time the model index changes in the user selection. Citation data are also provided with references to the original publication and links to the PDB and NDB entries if the report is for a DNAproDB database entry.

The second component of the report page is our interactive visualizations. We currently offer three visualization types: the residue contact map, the helical contact map and the helical shape plot. The residue contact map shows individual nucleotide–residue interactions, DNA secondary structure, protein secondary structure and interaction moieties all in one figure. The DNA is displayed as a graph, with individual nucleotides being nodes in the graph, and edges between them indicating backbone links, base pairing or base stacking, each with a distinct color. Different base-pairing geometries such as Watson–Crick, Hoogsteen or other non-standard pairing conformations are indicated via the base-pair edges, and other structural features such as backbone breaks, missing phosphates, the DNA strand sense and nucleotide structural moieties (see 'Identification of structural and interaction moieties', Figure 4B and Supplementary Figure S4) are all represented graphically. Protein residues are displayed as small nodes with the node

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

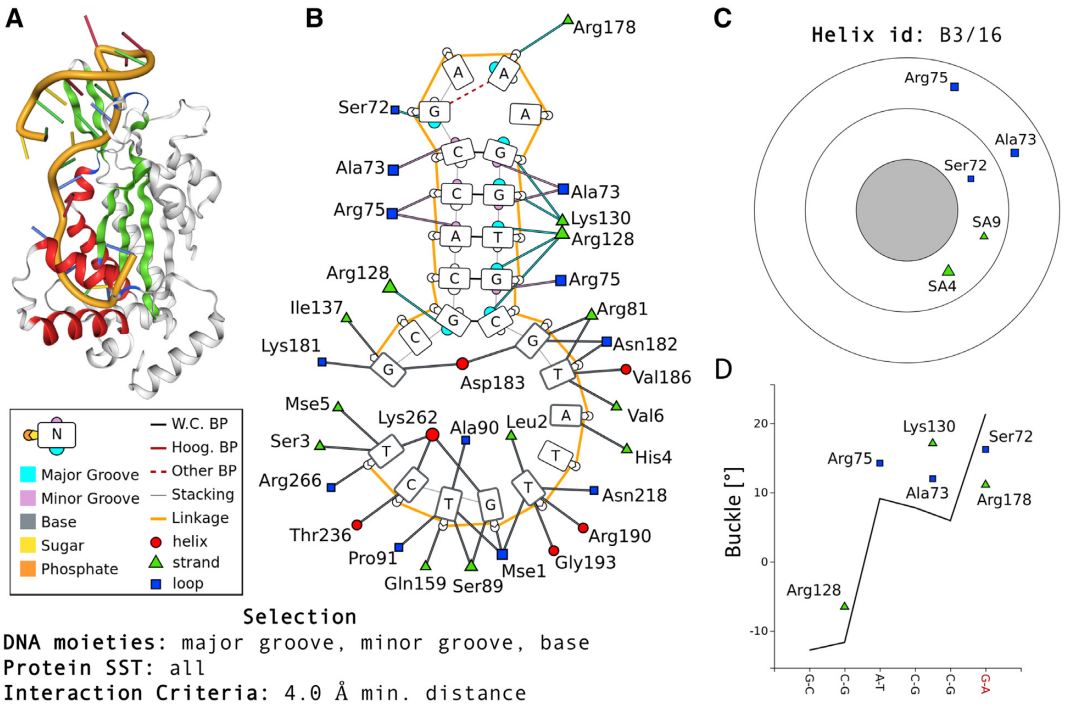

Figure 4. Example visualizations provided by DNAproDB that have been generated through our browser-based tools. DNAproDB visualizations are interactive, customizable and can be exported directly as PNG images. These visualizations were taken from the report page of the conjugative relaxase TrwC–DNA co-crystal structure (PDB ID: 1OMH) (26). Selected DNA moiety interactions were base, major and minor groove, all protein secondary structure types were included, and interactions were filtered by a minimum nucleotide–residue distance of 4.0 Å or less. The size (pixel area) of each residue or SSE symbol has been chosen to indicate the number of nucleotides it interacts with. (**A**) A 3D view of the structure produced by NGL (23,24). The colored regions are the interacting residues that are displayed in the remaining three panels and are colored using the same color scheme. (**B**) DNAproDB uses a graph-based approach for displaying the residue contact map. Information about nucleotide base-pairing, base-stacking and backbone linkages are displayed via edges between nucleotides, and edges between nucleotides and residues (shown here as red circles, blue squares and green triangles; the shapes and colors of the symbols indicating secondary structure) denote an interaction. The sense of the DNA strands can be seen based on the direction of the sugar-phosphate linkages between each adjacent nucleotide in each strand. A G/A mismatch can be seen near the top of the figure, flanked by two flipped-out bases and indicated by a red dashed line signifying the non-Watson–Crick base pair. (**C**) A polar contact map showing major and minor groove interactions in the helical region of the structure. Protein secondary structure elements are shown and plotted according to which DNA moieties they interact with. The major (inner annulus) and minor (outer annulus) grooves are represented by concentric annuli and the position of each interacting protein secondary structure element (SSE) within each annulus is determined by the helicoidal coordinates of the SSE (9). Protein β-sheets are represented by triangles, α-helices are represented by circles and loops by squares. The polar distribution of the SSE interactions reflects that the protein binds to one side of the helix. (**D**) A helical shape plot with the shape parameter Buckle plotted. The extreme value at the last position of the sequence is due to an A/G mismatch that is automatically indicated in red by DNAproDB. The green triangle and blue squares represent β-sheet and loop residues that interact with the DNA and their approximate location along the sequence of the double-helical region. The same residues can be seen interacting with the helical region of the DNA in panel (B).

shape and color representing the residue secondary structure, and edges between residue and nucleotide nodes represent an interaction between the two and which DNA moiety the interaction involves.

The helical contact map is a compact visualization where interactions with protein secondary structural elements (SSEs) are plotted along the helical axis for individual DNA helices present in the selection (if none exist then this visualization is not available). Interactions of SSEs with different DNA moieties are represented in concentric annuli and the position of each plotted SSE is given in helicoidal coordinates that is a curvilinear coordinate system defined by the axis of a helical DNA segment. This visualization is a unique and compact way to summarize the coarse-grained interactions of helical DNA regions. See Sagendorf *et al.* (9) for a detailed description of this visualization type.

The helical shape plot is the newest visualization type in DNAproDB and plots DNA shape parameters (such as major and minor groove width, or base-pair shape parameters) for a selected helix within the user selection along the sequence of the helix. In addition, DNA–protein residue interactions are plotted showing approximately where each residue in the interface interacts in sequence space and the secondary structure of that residue. Protein residue interactions can be toggled off if only DNA shape parameters are desired, or they can be displayed to indicate possible DNA shape readout, such as the presence of positively charged residues in regions of narrow minor groove width. Supplementary Figure S5 shows an example of a helical shape plot for minor groove width that illustrates such a readout mechanism for a Hox-Exd heterodimer (PDB ID: 2R5Z) (25).

All of our visualizations are highly customizable and interactive. Custom color schemes, labels and plot orientations can be chosen. Figure 4 shows example visualizations for the conjugative relaxase TrwC bound to a helical segment of DNA with a long single-stranded overhang (PDB

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

ID: 1OMH) (26). The visualized interactions and plot components are determined based on the selected DNA structural entity and protein chains: however, many additional criteria can be defined. Users can decide which DNA moiety interactions should be included or what protein secondary structure types to display. They can filter interactions based on the number of hydrogen bonds, buried accessible surface area, center of mass, mean nearest neighbor or minimum nucleotide–residue distances, or interaction geometry. Supplementary Figure S4 shows a variety of different criteria applied to the same structure in order to filter what interactions are shown. This high degree of customization allows users to create visualizations that display very specific views of the DNAproDB data to highlight particular aspects of the interface. Visualizations can be exported directly as a high-resolution static PNG file for use in publications or presentations. They are also interactive—hovering the cursor over various parts of the visualizations will display additional information and clicking on residues or nucleotides will highlight them in the 3D view of the structure.

The final component of the report page is the data explorer, which allows users to traverse the DNAproDB data file hierarchy and explore the raw data for that structure using a searchable JSON viewer. Every item and visualization on the report page is generated from the data contained in this data file that the user can download from the report page for their own use.

Integration with the Nucleic Acid Database

DNAproDB can also be accessed through the Nucleic Acid Database (NDB) (8,27). Each NDB entry is individually linked to its respective DNAproDB report page under NDB Structural Features. The integration of DNAproDB with the NDB makes DNAproDB report pages directly accessible through the PDB for any DNA containing structure.

DISCUSSION AND CONCLUSION

DNAproDB aims to provide a variety of biophysical and structural features that are useful for analyzing structures of DNA–protein complexes. These include commonly used features such as hydrogen bonding, DNA shape parameters, DNA and protein secondary structure, nucleotide–residue interaction distances and many more. DNAproDB and the features it provides can be used for many purposes. The data we provide can be used as is and DNAproDB can be treated simply as a processing pipeline, simplifying and automating the process of generating the set of features users find relevant to their work. It can be used as a visualization and data exploration tool by generating interactive plots and graphics of DNA–protein interfaces using the web-based tools available from structure report pages. By taking advantage of our database's search capabilities a user can generate sets of PDB entries that meet specific criteria using the variety of available features DNAproDB provides. The data we provide for each entry can also be used in more sophisticated ways for clustering, regression, classification or other statistical analysis using external software tools.

A key design consideration of DNAproDB was the ability for users to take our data and perform analyses independent of the web-based tools that we have developed. The choice of our data structure and format makes it easy for users with only a limited knowledge of JSON to parse, search and analyze the DNAproDB data. In Figure 5 we show three examples of different types of analyses done using DNAproDB data. In each case, we used the flat DNAproDB database file (which is available to users through our website) in combination with external software tools. A full description of the methods used to perform these analyses is provided in Supplementary Data. In Figure 5A, we examine an 'interaction motif' for the residue tyrosine. The heat map shown is for a three-nucleotide minor groove interaction where tyrosine interacts with exactly three nucleotides in the DNA minor groove and displays a strong signal for a C-C-G interaction. The right panel offers a plausible explanation for the presence of this motif—the hydroxyl group of the tyrosine can form bidentate hydrogen bonds with the O2 atom of cytosine and the N2 atom of guanine along the C/G base pair's minor groove edge. These favorable interactions may increase the likelihood of tyrosine binding to a CG-rich region of the DNA minor groove. Such an interaction motif may be relevant for proteins that bind DNA sequences containing CG regions and contain tyrosine in their binding domains and is a good example of how we can infer biologically relevant information from structural analysis using DNAproDB.

In Figure 5B we show how features of the DNA–protein interface vary by the biological function of the protein, with proteins of different functions occupying distinct regions of the feature space. Every point in this plot is a protein chain that interacts with DNA from a DNAproDB entry and the color represents a biological function or process the protein is involved in (based on Gene Ontology annotations). Correlations can be seen between the function of the protein chain and the way it interacts with DNA when the interface features are projected to a low dimensional space. Several distinct, though partially overlapping, clusters are evident in this principal component analysis (PCA). The PCA plot in Figure 5B indicates that DNAproDB features can capture differences in the binding mechanisms and characteristics of these proteins, which are at least partially related to their biological function. This is important since we expect that proteins that bind different forms of DNA and under different circumstances should have noticeably distinct binding mechanisms, and the set of features that describe the DNA–protein binding should reflect those differences accordingly.

In Figure 5C, we calculate the probability of protein residues with planar side chains to form stacking interactions with different nucleotide bases in single-stranded DNA. These probabilities can be used to roughly estimate the free energy of stacking via

$$\Delta G^{\text{stack}}_{N,R} = -RT \ln \frac{P(\text{stack}|N,\ R)}{1 - P(\text{stack}|N,\ R)}$$

where N is a nucleotide type and R is a residue type. Naturally, we do not expect the sampling from a limited number of structural examples to be good enough for an accurate estimation; however, the notable differences in stacking probabilities may be relevant to the binding specificities of cer-

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

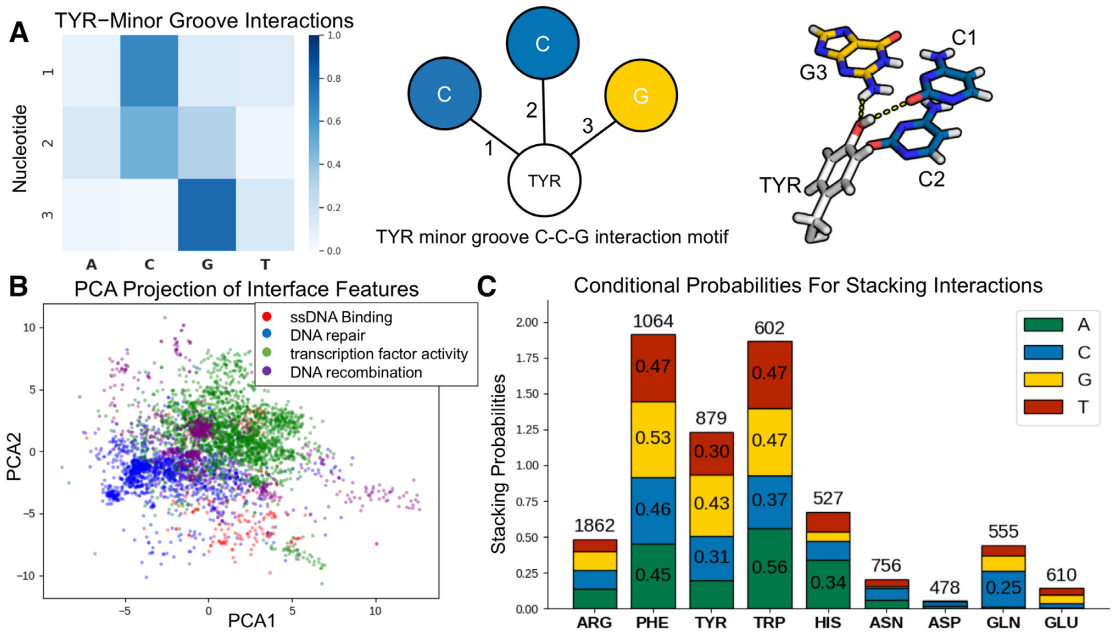

Figure 5. Several example analyses done using data provided by the DNAproDB database. (**A**) Shown here is a three nucleotide 'interaction motif' for the residue tyrosine interacting in the minor groove of helical DNA. The heat map was generated by taking all available instances (99; interactions were filtered for sequence redundancy of the interacting protein chain) of tyrosine residues that bind in the DNA minor groove and interact with exactly three nucleotides simultaneously (identified from the list of nucleotide–residue interactions provided under interface features). The three interacting nucleotides were sorted by center of mass distance (an interaction feature) and placed into bins 1, 2 and 3 from the nearest nucleotide to the furthest. In this way, we capture the radial distribution of the 3D structure of the interacting nucleotide triplet. The frequency of each nucleotide type (A, C, G or T) was then plotted for each distance bin. In this example, we see a strong signal for a C-C-G interaction motif. We note that this is not the same as a sequence motif because these nucleotides need not be adjacent or even on the same strand. On the right side of the panel an example is shown of the structure of this interaction motif with a tyrosine forming two hydrogen bonds between its OH group and the N2 atom of the guanine and the O2 atom of cytosine. See Supplementary Data for a more detailed description of the analysis. (**B**) The first and second principal components from a PCA projection of 134 DNAproDB features (or features derived from them) describing the DNA–protein interface for 4758 different interfaces (which were broken down by protein chain). Each point in the plot corresponds to one interface. The plots are colored according to the GO annotations of the protein chain and are grouped into four categories based on the annotated biological function of the protein. Correlations can be seen between the protein chain annotations and the way it interacts with DNA as captured by the DNAproDB features when projected to this low dimensional space. Several distinct, albeit overlapping, clusters can be seen that correspond to the different biological functions of the involved protein chain. See Supplementary Data for a more detailed description of the analysis. (**C**) Probabilities for a particular nucleotide–residue interaction to be in a stacking conformation. More precisely, these are conditional probabilities of $P(\text{stack}|N, R)$ where N is a nucleotide type and R is a residue type. DNAproDB assigns a geometry for every nucleotide–residue interaction identified using SNAP, a component of the 3DNA program suite (10). The residues for which probabilities are shown are those with planar side chains so that a stacking conformation can be defined. The conditional probabilities for each residue to stack with each nucleotide are shown as a stacked bar chart, with the numbers inside the bars indicating the probability values and the numbers overtop of the bars the total number of interactions used for that residue type. Only interactions with nucleotides in a single-stranded conformation were included in this analysis. It is interesting to note the variation in stacking probabilities between different nucleotides for a given residue, most notably with tyrosine preferring guanine stacking and histidine strongly preferring adenine stacking. See Supplementary Data for a more detailed description of the analysis.

tain single-stranded DNA-binding proteins—preferring sequences which, among other mechanisms, can form stacking interactions with favorable energies. DNAproDB makes sampling structural data easy, and one can perform many kinds of analyses thanks to the large volume of structure-based data available in our database.

The structures of biological macromolecules are a rich source of information, and DNAproDB makes use of the wealth of available data to enable structure-based computational biology. DNAproDB processes structures provided by the PDB in a unique way and makes the features and annotations we generate available to be used by researchers for further studies. The improvements and updates we have made to DNAproDB have enriched our database with a much larger number of available entries. The addition of new features, improved data organization, a better set of visualization tools and more user-friendly web interfaces make the newest release of DNAproDB a much stronger tool for structural analysis of DNA–protein complexes.

DATA AVAILABILITY

Additional resources such as example code and detailed documentation can be found at https://dnaprodb.usc.edu/documentation.html.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

A collaboration of the authors with Xiaojiang Chen initiated the inclusion of single-stranded DNA–protein com-

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019

plexes in DNAproDB leading to many of the updates discussed. The authors thank all current members of the Rohs laboratory for comments, suggestions and testing of DNAproDB. The authors also thank Luigi Manna for administrating the server hosting DNAproDB and Catherine Lawson for linking to DNAproDB entries from the Nucleic Acid Database (NDB).

FUNDING

National Institutes of Health [R01GM106056, R01GM087986 (in part), R01HG003008 (in part), R35GM130376 to R.R., R01GM085328 to H.M.B.]; Rose Hills Foundation [to N.M.]; Human Frontier Science Program [RGP0021/2018 to R.R.]. Funding for open access charges: National Institutes of Health [R35GM130376].
Conflict of interest statement. None declared.

REFERENCES

1. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
2. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
3. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide protein data bank. *Nat. Struct. Biol.*, **10**, 980.
4. Norambuena,T. and Melo,F. (2010) The Protein-DNA interface database. *BMC Bioinformatics*, **11**, 262.
5. Contreras-Moreira,B. (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Res.*, **38**, D91–D97.
6. Zanegina,O., Kirsanov,D., Baulin,E., Karyagina,A., Alexeevski,A. and Spirin,S. (2016) An updated version of NPIDB includes new classifications of DNA–protein complexes and their families. *Nucleic Acids Res.*, **44**, D144–D153.
7. Laskowski,R.A., Jabłońska,J., Pravda,L., Vařeková,R.S. and Thornton,J.M. (2018) PDBsum: structural summaries of PDB entries. *Protein Sci.*, **27**, 129–134.
8. Coimbatore Narayanan,B., Westbrook,J., Ghosh,S., Petrov,A.I., Sweeney,B., Zirbel,C.L., Leontis,N.B. and Berman,H.M. (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res.*, **42**, D114–D122.
9. 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.
10. 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.
11. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
12. Dawson,N.L., Lewis,T.E., Das,S., Lees,J.G., Lee,D., Ashford,P., Orengo,C.A. and Sillitoe,I. (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.*, **45**, D289–D295.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
14. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
15. ECMA International (2017) *Standard ECMA-404. The JSON Data Interchange Format*. Ecma Publication, ISO/IEC 21778.
16. Westbrook,J.D. and Fitzgerald,P.M.D. (2009) *Structural Bioinformatics*. 2nd edn. John Wiley & Sons, Inc., Hoboken, New Jersey, pp. 271–291.
17. Word,J.M., Lovell,S.C., Richardson,J.S. and Richardson,D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.
18. Dolinsky,T.J., Nielsen,J.E., McCammon,J.A. and Baker,N.A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.*, **32**, W665–W667.
19. Dolinsky,T.J., Czodrowski,P., Li,H., Nielsen,J.E., Jensen,J.H., Klebe,G. and Baker,N.A. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, **35**, W522–W525.
20. 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.
21. 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.
22. Plugge,E., Hawkins,T. and Membrey,P. (2010) *The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing*, Springer-Verlag New York Inc., NY.
23. Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
24. 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.
25. 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.
26. Guasch,A., Lucas,M., Moncalián,G., Cabezas,M., Pérez-Luque,R., Gomis-Rüth,F.X., de la Cruz,F. and Coll,M. (2003) Recognition and processing of the origin of transfer DNA by conjugative relaxase TrwC. *Nat. Struct. Biol.*, **10**, 1002–1010.
27. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz889/5587632 by PPD Development LP user on 20 October 2019