Nucleic Acids Research, 2024, 1–7
https://doi.org/10.1093/nar/gkae970
Database issue

DNAproDB: an updated database for the automated and interactive analysis of protein–DNA complexes

Raktim Mitra [1,†], Ari S. Cohen [1,†], Jared M. Sagendorf [1], Helen M. Berman [1,2] and Remo Rohs [1,3,4,5,*]

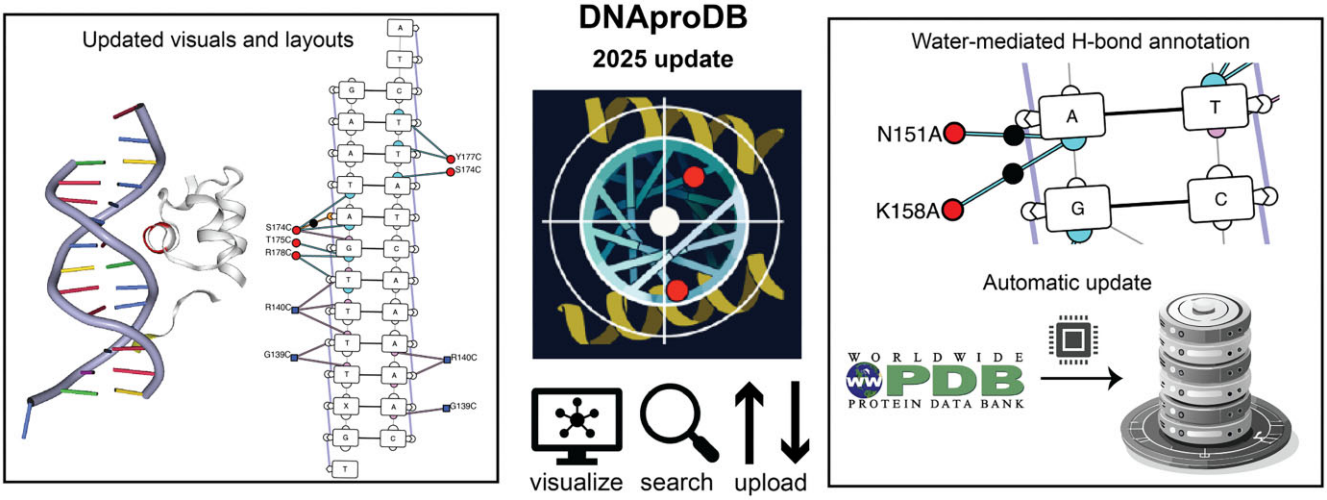
[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[2]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA
[3]Department of Chemistry, University of Southern California, Los Angeles, CA 90089, USA
[4]Department of Physics & Astronomy, University of Southern California, Los Angeles, CA 90089, USA
[5]Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA

[*]To whom correspondence should be addressed. Tel: +1 213 740 0552; Fax: +1 213 821 4257; Email: rohs@usc.edu
[†]The first two authors should be regarded as Joint First Authors.
Present address: Jared M. Sagendorf, Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94158, USA.

Abstract

DNAproDB (https://dnaprodb.usc.edu/) is a database, visualization tool, and processing pipeline for analyzing structural features of protein–DNA interactions. Here, we present a substantially updated version of the database through additional structural annotations, search, and user interface functionalities. The update expands the number of pre-analyzed protein–DNA structures, which are automatically updated weekly. The analysis pipeline identifies water-mediated hydrogen bonds that are incorporated into the visualizations of protein–DNA complexes. Tertiary structure-aware nucleotide layouts are now available. New file formats and external database annotations are supported. The website has been redesigned, and interacting with graphs and data is more intuitive. We also present a statistical analysis on the updated collection of structures revealing salient patterns in protein–DNA interactions.

Graphical abstract



Introduction

Protein–DNA interactions play crucial roles in essential cellular functions like gene regulation, genome packaging, and DNA replication (1,2). Diverse recognition mechanisms underlie these interactions (3–6). Atomic resolution structures of protein–DNA complexes available in the Protein Data Bank (PDB) (7) have been invaluable for understanding these readout mechanisms and provide insight that relate them to func-

Received: September 8, 2024. **Revised:** October 7, 2024. **Editorial Decision:** October 9, 2024. **Accepted:** October 11, 2024

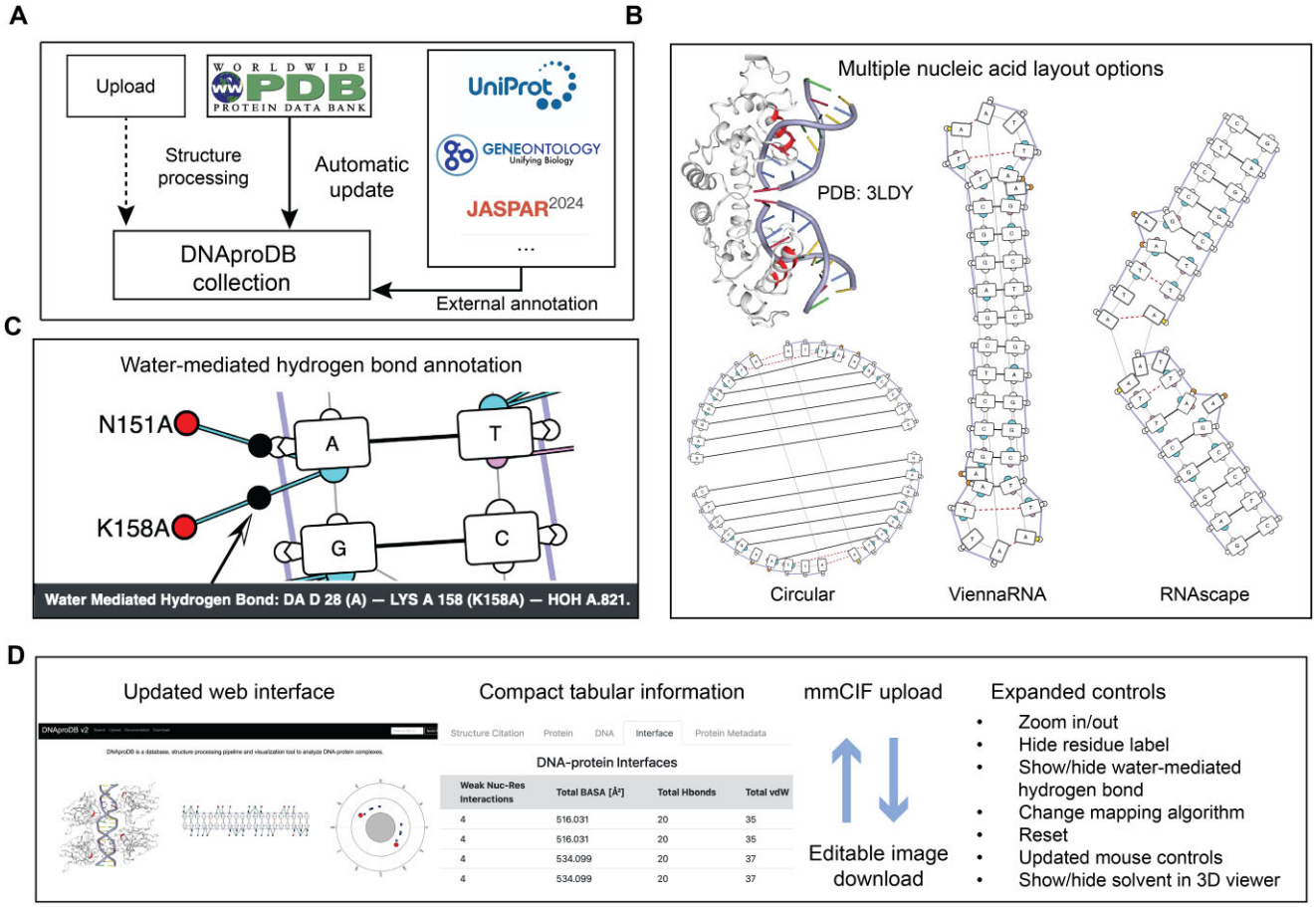


Figure 1. Key aspects of this update to DNAproDB. (**A**) Automatic update and decoupled external annotation incorporation scheme. (**B**) Different nucleic acid layout options with added tertiary structure aware RNAscape layout, shown for PDB ID: 3LDY. (**C**) Water-mediated hydrogen bond annotation. (**D**) Various improvements in other aspects of DNAproDB.

tion. As a computational resource which extensively analyzes such structures and presents their data in publication-quality representations, the DNAproDB web server (8) and database (9) have been a useful resource for biologists, and are linked by tool libraries such as the Nucleic Acid Knowledge Base (NAKB) (10).

This update improves the DNAproDB analysis pipeline, output data presentation, and web interface (Figure 1). The updated analysis pipeline now computes annotations of water-mediated hydrogen bonds, which are known to play an important role (11) in protein–DNA recognition and, in some cases, a very prominent one (12). New PDB structures are automatically processed and incorporated into DNAproDB weekly. The primary interface visualization, 'Residue contact map', now allows users to select a mapping algorithm for nucleic acid layout. In addition to secondary structure-based mapping (13), tertiary-structure aware mapping (14) is now available. Binding specificity data for transcription factors catalogued in the JASPAR2024 database (15) has been integrated. Users can now upload structures in the macromolecular Crystallographic Information File (mmCIF) format and download interface visualizations in an editable figure format. More information regarding these updates, as well as quality-of-life and user-interface improvements, is described in the following sections. The DNAproDB search functionality and documentation have also been expanded.

We analyzed the expanded DNAproDB structure collection for salient features of protein–DNA interactions (Figure 2). These results (based on a larger sample size in this update) reaffirm previous statistics presented about DNA minor groove recognition (3) and patterns of amino acid-base stacking for single stranded DNA (9). Additionally, we present and discuss examples of the newly added water-mediated hydrogen bond annotations in selected structures (Figure 3).

DNAproDB has been used by experimental biologists to upload, analyze, and present interface visualizations in their work (16). We developed this update to assist their efforts, likely leading to additional contributions from the scientific community. We want to emphasize the increased utility of DNAproDB in light of structure prediction tools like AlphaFold3 (17), RoseTTAFoldNA (18), and RoseTTAFold-AA (19), and binding specificity prediction tools including DeepPBS (20) and rCLAMPS (21). These computational tools hint towards a promising future of protein–DNA structure prediction and design (22). We expect that DNAproDB will be an invaluable tool and assist such efforts.

Update details

Processing pipeline and data update

At the time of its previous release (9), DNAproDB contained a static collection of structures. This resulted in PDB struc-

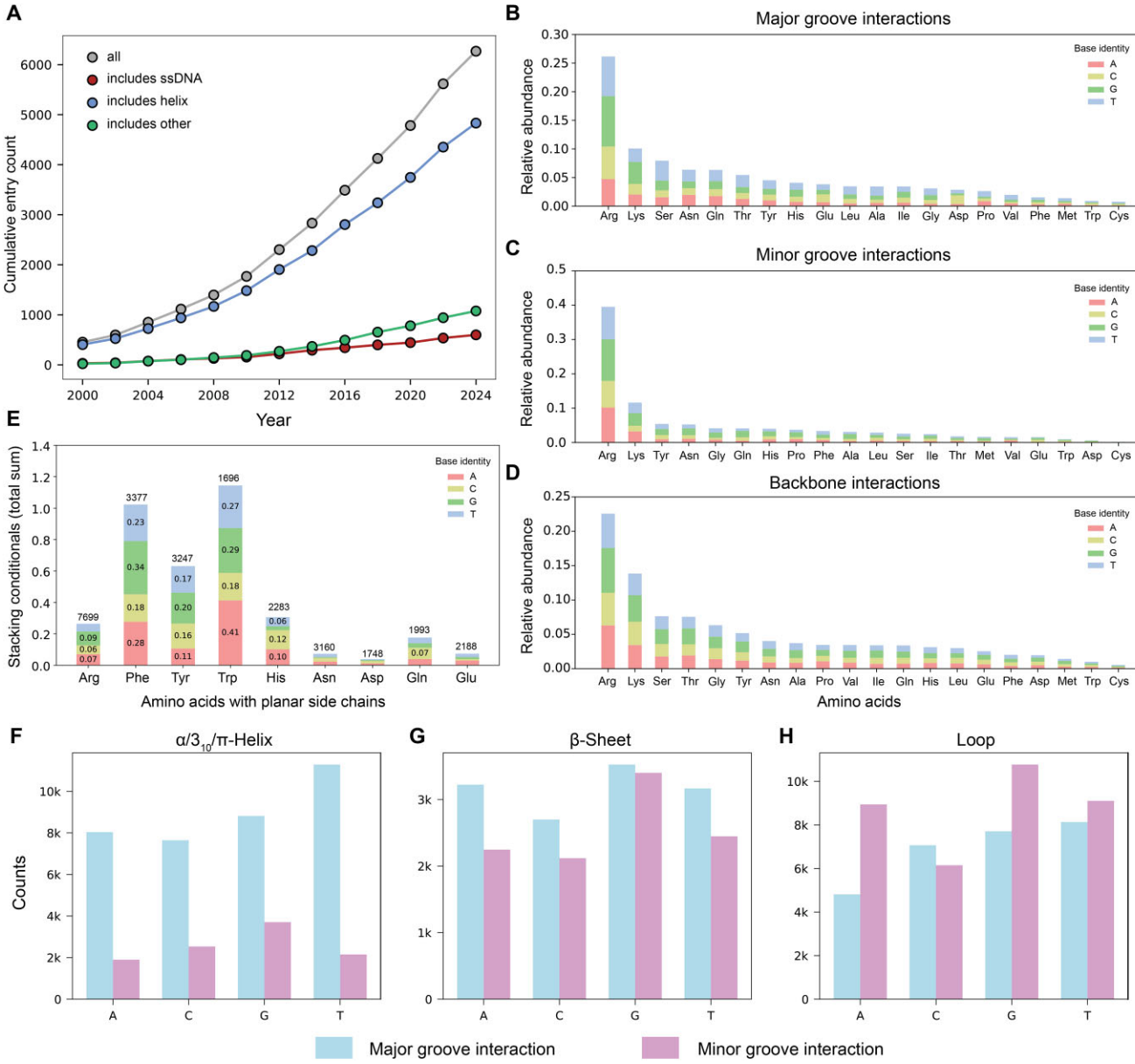


Figure 2. Quantitative analysis of protein–DNA complexes in the DNAproDB collection. (**A**) PDB release years of structures catalogued in the updated DNAproDB collection (as of 7 June 2024). The plot compares the total number of entries for protein–DNA complexes with the number of entries for single-stranded DNA, double-stranded DNA helices, and other DNA conformations. (B–D) Relative abundance of different amino acids interacting with the DNA major groove (**B**), minor groove (**C**), and phosphodiester backbone (**D**). In each case, fraction of interaction with each base is shown in color. (**E**) Conditional probabilities of different protein residues and base forming a stacking geometry. *Y*-axis represents summed values over the bases for each amino acid. Interaction count associated with each amino acid is shown above each stacked bar. (F–H) Counts of interactions with different bases, categorized by major and minor groove for secondary structure classes: helix (includes $\alpha/3_{10}/\pi$-helix) (**F**), sheet (β-sheet) (**G**) and loop residues (**H**).

tures released after the most recent DNAproDB update being unavailable. In this update, we have addressed this limitation by implementing an automatic update pipeline (Figure 1A). Every week, the pipeline queries the PDB for newly released structures, downloads and processes them, and adds them to the DNAproDB collection.

In addition, the structure processing pipeline has been decoupled from any external annotation dependencies. This allows external annotations to be updated without reprocessing each structure or affecting the user experience. Annotations from the JASPAR2024 database (15) (incorporating the most recent binding specificity matrix ID and logo) have been included whenever applicable.

The asymmetric unit molecular weight cutoff, which determines whether a structure is included in the collection, has been expanded from 250 to 1500 kDa, increasing the number of structures available for analysis. The latest collection size as of 7 June 2024, is 6731 structures. This set has been analyzed and was included in the results presented in Figure 2.

Originally, a large part of the processing pipeline was written using Python 2 (23). We redesigned the backend processing pipeline to ensure compatibility with Python 3 (24).

Expanding its functionality, DNAproDB now calculates and annotates water-mediated hydrogen bonds between protein and DNA within this update. The program HBPLUS (25), with the '-h' option set to 3 Å, and the '-d' option set to 3.5 Å,

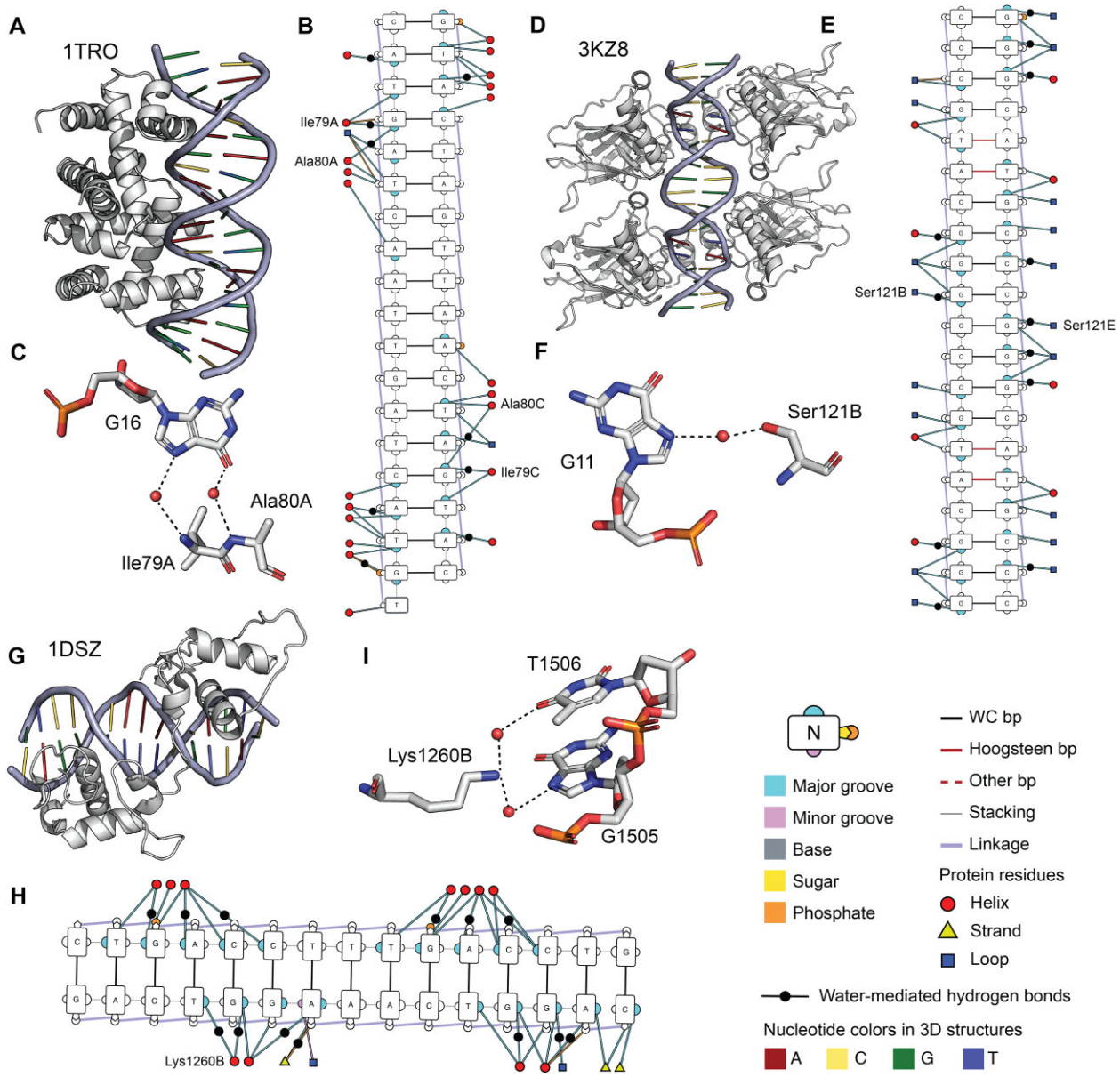


Figure 3. Water-mediated hydrogen bond annotation in DNAproDB. Selected examples of water-mediated hydrogen bond annotations as reflected in the updated DNAproDB. (**A–C**) Trp repressor/operator complex (PDB ID: 1TRO) (**D–F**) p53 tetramer with Hoogsteen base pairs (PDB ID: 3KZ8) (**G–I**) RXR-RAR DNA-binding complex (PDB ID: 1DSZ). In each of the three cases, the 3D structure of the respective complex is shown in (A, D, G). The DNAproDB 'Residue contact map' is shown (with only selected protein residues annotated) in (B, E, H). Atomic views of selected water-mediated hydrogen bond interactions are shown in (C, F, I), respectively.

and with the remaining parameters kept as default, is used to detect hydrogen bonds. Custom scripts were written to determine water-mediated interactions via shared water molecules between hydrogen-bonded pairs (see Data Availability).

Visualization

We updated the 'Residue contact map' and 3D structure (Figure 1B) visualizations presented in DNAproDB in several ways. The nucleic acid backbone color used in these components has been changed to a more visually pleasing metallic blue-gray color, compared to the previously used yellow-orange color.

In addition to the existing secondary structure-based and circular layouts, an RNAscape (14) based layout for placing nucleic acids has been computed and added to the 'Residue contact map'. This new layout is more representative of tertiary structure compared to the other two representations (Figure 1B). An option to switch between these different layouts is available.

During this update, some Python 2 version utilities for secondary structure-based layout computation were discontinued. We replaced these utilities with analogous Python 3 versions provided by the 'Forgi' package (26).

Water-mediated hydrogen bonds have now been incorporated as an interaction edge in the 'Residue contact map'. These are indicated by a black circle (Figure 1C) in the interaction map. Hovering over the water-mediated contacts will present further information (e.g. residue number of the water molecule involved). An option to hide these interactions is also

available. The '3D viewer' component displays the structure without solvent and a button to show solvent alongside the structure is included.

Web interface and user experience

Since its inception, we have continuously provided support for DNAproDB users and taken note of their feedback. In this update, we redesigned the web interface based on this information (Figure 1D). The home page and 'Quick Search' field now have suggestions for PDB IDs to explore. This can be helpful for a first-time user. Instructions and explanations for different components, which were previously written directly on the page, are now available as pop-up components upon mouse hover. Report pages for each PDB entry now prominently display the title of the entry. The information tables have been rearranged in a modern and tabular fashion, resulting in a clearer representation of information.

DNAproDB offers many customization features for the 'Residue contact map'. However, these options were often overlooked by users due to their non-prominent placement on the website. We have redesigned the user interface to make basic options like rotation, zooming, download, and switching between the layout algorithms easily accessible directly above the visualization. Buttons to access further customization options ('Chart options' and 'Interface selection') are prominently placed. The options within the 'Chart options' tab have been expanded. Within the 'Interface selection' tab, basic options (model, entity, chain, moiety selection) are shown first. Additional options are presented as advanced options. Mouse-based interaction controls for the '3D viewer' and 'Residue contact map' have been made analogous, to the extent possible.

The download option now supports the editable Scalable Vector Graphics (SVG) format. DNAproDB currently displays Watson-Crick, Hoogsteen, and other base-pairing geometries via correspondingly stylized base-pairing edges (e.g. Hoogsteen base-pairing in p53 tetramer–DNA complex (5) reflected in Figure 3E). For additional analysis of non-Watson-Crick base-pairing geometries, a link to the RNAscape webserver (14) has been included in each report page. Clicking this link will redirect the user to the RNAscape website and automatically run it on the desired structure.

The 'Documentation' page has been updated to include troubleshooting instructions and a detailed description of the report page and visualizations presented by DNAproDB. The 'Search' page has been reorganized, and a new search category 'Additional Options' has been added. Through this category, users can search structures based on gene names, JASPAR IDs, or Gene Ontology entry identifiers.

Quantitative analysis of readout features

Entries in the DNAproDB collection (as of 7 June 2024) encompass protein–DNA structures including single-stranded DNA (ssDNA), double-stranded DNA helices (dsDNA), and other conformations (e.g. G-quadruplex). We quantified the growth of such entries over time based on their PDB release dates, which reflects an exponential trend (Figure 2A). Fewer entries contain ssDNA and other conformations compared to dsDNA. However, recent years (2016 onwards) demonstrate a steady growth in ssDNA entries (Figure 2A).

Studies on protein–DNA structures have revealed consistent patterns in protein residue–DNA interaction frequencies (3,27). We sought to quantify similar statistics in the updated collection of DNAproDB. To this end, we computed relative abundances of different amino acids interacting with the major groove (Figure 2B), minor groove (Figure 2C), and phosphodiester backbone (Figure 2D). Relative abundance for a residue (R) is the fraction of occurrence of this protein residue interacting with a DNA moiety relative to other residues.

$$\text{Relative abundance } (R) \;=\; \frac{|\mathit{Interactions\ involving}\ R|}{\sum_R |\mathit{Interactions\ involving}\ R|}$$

This is computed separately for the major groove, minor groove, and DNA backbone. Each of these values in Figure 2B-E is further subdivided into fractions per DNA base, shown in four colors. For the major groove, we see an abundance of residues able to perform recognition via hydrogen bonds, with arginine (Arg) and lysine (Lys) residues showing the greatest presence (Figure 2B). For the minor groove, this preference for arginine and lysine is even stronger relative to other residues (Figure 2C). This agrees with the observation that the minor groove is more electronegative (3), favoring positively charged amino acid sidechains while repelling negatively charged sidechains [e.g. aspartic acid (Asp), glutamic acid (Glu) etc.].

For amino acid residues (R) with a planar side chain component (i.e. able to form a stacking interaction with a base ($B \in [A, C, G, T]$) in single-stranded DNA), interaction geometries (g) can be of three different types: $g \in [\mathit{stack},\ \mathit{pseudo\ pair},\ \mathit{other}]$ (based on SNAP (28)). Stacking conditionals $P(g = \mathit{stack} \mid R,\ B)$ were computed for major and minor groove interactions as a fraction of the counts of *stack* geometry against counts for all geometries. i.e.

$$P(g = \mathit{stack} \mid R,\ B) = \frac{|g = \mathit{stack},\ R, B|}{\sum_g |g,\ R,\ B|}$$

This term sums to 1 when summed over g (not for $R,\ B$). This information is presented in Figure 2E in the form of a stacked bar chart. The total height of each stacked bar (i.e. for each amino acid) is $\sum_B P(g = \mathit{stack} \mid R,\ B)$. The pattern visible in this data conforms with the previously computed version in (9) while encompassing a larger sample size.

DNAproDB also provides annotations and a visualization ('Helical contact map') reflecting how various secondary structure elements of a protein interact with the major and minor groove of DNA. We quantified these interactions to reveal statistical patterns (Figure 2F-H). We compute instances of helical secondary structures (including α-helices, π-helices and 3_{10}-helices) interacting with the four primary DNA bases in either the major or minor groove (Figure 2F). There is a clear preference for protein contacts through α-helices in the major groove, reflecting the use of a recognition helix by many protein families (29). On the other hand, for β-sheets, major and minor groove interactions are comparable in number, with a slight preference for the major groove (Figure 2G). The 'Loop' category reflects residues appearing in loop regions of proteins interacting with DNA. Minor groove interactions are slightly more favored in this case (Figure 2H). In all cases, guanine (G) is the most favored DNA base that is contacted.

Water-mediated hydrogen bonds

As described previously, the updated DNAproDB processing pipeline detects and visually annotates water-mediated hy-

drogen bond interactions between protein and DNA (Figure 1B). This feature improves the accuracy and relevance of the DNAproDB visualization for some structures. For example, the co-crystal structure of the Trp repressor/operator complex (PDB ID: 1TRO, Figure 3A, Residue contact map: Figure 3B) reflects a protein–DNA recognition scheme without any direct hydrogen bonds in the major and minor groove. Instead, DNA recognition occurs via water-mediated hydrogen bonds (Figure 3B) (12). A detailed view of two protein backbone nitrogen atoms (belonging to Ile79 and Ala80) recognizing G11 in this manner is presented in Figure 3C. This type of recognition scheme was previously not reflected in DNAproDB.

Similarly, protein residues interacting with DNA only through water-mediated hydrogen bonds were also not displayed in the 'Residue contact map'. One such example is the p53 tetramer structure (PDB ID: 3KZ8 (5), Figure 3D, Residue contact map: Figure 3E). This structure illustrates serine residues (Ser121) near the tetramerization interfaces involved in water-mediated hydrogen bonds with the major groove edge of two G bases (shown for one selected base in Figure 3F). As this is the sole mode of interaction for these two residues, they were omitted from the visualization in the previous DNAproDB version (9). In this update, these interactions are correctly shown. A variety of complex interaction geometries are possible when water-mediated hydrogen bonds are involved. One such example can be found in interactions of the RXR/RAR DNA-binding domain heterodimer in complex with the retinoic acid response element (PDB ID: 1DSZ (30), Figure 3G, Residue contact map: Figure 3H). The lysine residue (Lys1260) is involved in recognizing consecutive bases (G and T) through water-mediated hydrogen bonds involving two different water molecules. This update to DNAproDB allows exploring such recognition schemes.

Discussion

DNAproDB, since its inception in 2017 (8), has been a valuable resource for the structural biology community. Its comprehensive analysis pipeline, covering diverse aspects of protein–DNA binding, outputs data that can be readily used in downstream analysis by the user (9). DNAproDB also provides interactive and publication-quality visualizations. In this update, we improved DNAproDB in multiple aspects. New structures released since the last update in 2020 (9) have been incorporated, resulting in a much larger collection. The pipeline has been future-proofed via the new automatic update feature. The backend implementation has been upgraded to Python 3, ensuring a long-lasting lifespan for DNAproDB.

A key scientific improvement in the analysis pipeline is the incorporation of water-mediated hydrogen bond calculation. Interest in water-mediated interactions has been growing. This is evidenced by the CASP16 challenge for predicting solvent shells around the *Tetrahymena* ribozyme structure (31). Currently, these interactions are not well modeled by structure prediction and analysis tools (17–20,32,33). We expect that this added feature in DNAproDB will advance the field in understanding readout mechanisms.

Visualizations have been improved by enabling tertiary structure-aware nucleic acid layouts, incorporation of water-mediated hydrogen bond indicators, better customizability, and other visual improvements. The website has been redesigned, and data presentation has been improved. Structure files in mmCIF format can now be uploaded, which was previously unsupported. Altogether, these updates result in an improved DNAproDB, which we expect to continue serving the structural biology community for the foreseeable future.

Data availability

DNAproDB and associated data are freely available for all users at https://dnaprodb.usc.edu/.

The pipeline and frontend implementations are available through figshare at https://doi.org/10.6084/m9.figshare.27263145, and via GitHub at https://github.com/timkartar/DNAproDB and https://github.com/ariscohen/DNAproDB_frontend.

Acknowledgements

The authors acknowledge Luigi Manna for setup and maintenance of DNAproDB and thank the Rohs lab members for their support and valuable feedback.

Author contributions: Conceptualization (R.M., J.M.S., H.M.B., R.R.), methodology (R.M., A.S.C., J.M.S., H.M.B., R.R.), visualization (R.M., A.S.C.), software (R.M., A.S.C., J.M.S.), manuscript writing (R.M., A.S.C., R.R.), supervision (R.R.).

Funding

Andrew J. Viterbi Fellowship in Computational Biology and Bioinformatics (to R.M.); National Institutes of Health [R35GM130376 to R.R.]; Human Frontier Science Program [RGP0021/2018 to R.R.]. Funding for open access charge: National Institutes of Health [R35GM130376].

Conflict of interest statement

None declared.

References

1. Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.
2. Lai,W.K.M. and Pugh,B.F. (2017) Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat. Rev. Mol. Cell Biol.*, **18**, 548–562.
3. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.
4. 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.
5. Kitayner,M., Rozenberg,H., Rohs,R., Suad,O., Rabinovich,D., Honig,B. and Shakked,Z. (2010) Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.*, **17**, 423–429.
6. Chiu,T.P., Rao,S. and Rohs,R. (2023) Physicochemical models of protein–DNA binding with standard and modified base pairs. *Proc. Natl. Acad. Sci. U.S.A.*, **120**, e2205796120.
7. wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
8. 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.

9. Sagendorf,J.M., Markarian,N., Berman,H.M. and Rohs,R. (2020) DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.*, **48**, D277–D287.
10. Lawson,C.L., Berman,H.M., Chen,L., Vallat,B. and Zirbel,C.L. (2024) The Nucleic Acid Knowledgebase: A new portal for 3D structural information about nucleic acids. *Nucleic Acids Res.*, **52**, D245–D254.
11. Reddy,C.K., Das,A. and Jayaram,B. (2001) Do water molecules mediate protein–DNA recognition? *J. Mol. Biol.*, **314**, 619–632.
12. Otwinowski,Z., Schevitz,R.W., Zhang,R.-G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp represser/operator complex at atomic resolution. *Nature*, **335**, 321–329.
13. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithm. Mol. Biol.*, **6**, 26.
14. Mitra,R., Cohen,A.S. and Rohs,R. (2024) RNAscape: geometric mapping and customizable visualization of RNA structure. *Nucleic Acids Res.*, **52**, W354–W361.
15. Rauluseviciute,I., Riudavets-Puig,R., Blanc-Mathieu,R., Castro-Mondragon,J.A., Ferenc,K., Kumar,V., Lemma,R.B., Lucas,J., Chèneby,J., Baranasic,D., *et al.* (2024) JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **52**, D174–D182.
16. Webb,J.A., Farrow,E., Cain,B., Yuan,Z., Yarawsky,A.E., Schoch,E., Gagliani,E.K., Herr,A.B., Gebelein,B. and Kovall,R.A. (2024) Cooperative Gsx2–DNA binding requires DNA bending and a novel Gsx2 homeodomain interface. *Nucleic Acids Res.*, **52**, 7987–8002.
17. Abramson,J., Adler,J., Dunger,J., Evans,R., Green,T., Pritzel,A., Ronneberger,O., Willmore,L., Ballard,A.J., Bambrick,J., *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**, 493–500.
18. Baek,M., McHugh,R., Anishchenko,I., Jiang,H., Baker,D. and DiMaio,F. (2024) Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods*, **21**, 117–121.
19. Krishna,R., Wang,J., Ahern,W., Sturmfels,P., Venkatesh,P., Kalvet,I., Lee,G.R., Morey-Burrows,F.S., Anishchenko,I., Humphreys,I.R., *et al.* (2024) Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, **384**, eadl2528.
20. Mitra,R., Li,J., Sagendorf,J.M., Jiang,Y., Cohen,A.S., Chiu,T.P., Glasscock,C.J. and Rohs,R. (2024) Geometric deep learning of protein–DNA binding specificity. *Nat. Methods*, **21**, 1674–1683.
21. Wetzel,J.L., Zhang,K. and Singh,M. (2022) Learning probabilistic protein–DNA recognition codes from DNA-binding specificities using structural mappings. *Genome Res.*, **32**, 1776–1786.
22. Glasscock,C.J., Pecoraro,R., McHugh,R., Doyle,L.A., Chen,W., Boivin,O., Lonnquist,B., Na,E., Politanska,Y., Haddox,H.K., *et al.* (2023) Computational design of sequence-specific DNA-binding proteins. bioRxiv doi: https://doi.org/10.1101/2023.09.20.558720, 21 September 2023, preprint: not peer reviewed.
23. Van Rossum,G. and Drake Jr,F.L. (1995) *Python Reference Manual*. Centrum voor Wiskunde en Informatica, Amsterdam.
24. Van Rossum,G. and Drake,F.L. (2009) *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.
25. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
26. Thiel,B.C., Beckmann,I.K., Kerpedjiev,P. and Hofacker,I.L. (2019) 3D based on 2D: Calculating helix angles and stacking patterns using forgi 2.0, an RNA Python library centered on secondary structure elements. *F1000Res.*, **8**, 287.
27. Lin,M. and Guo,J. (2019) New insights into protein–DNA binding specificity from hydrogen bond based comparative study. *Nucleic Acids Res.*, **47**, 11103–11113.
28. 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.
29. Garvie,C.W. and Wolberger,C. (2001) Recognition of specific DNA sequences. *Mol. Cell*, **8**, 937–946.
30. Rastinejad,F., Wagner,T., Zhao,Q. and Khorasanizadeh,S. (2000) Structure of the RXR–RAR DNA-binding complex on the retinoic acid response element DR1. *EMBO J.*, **19**, 1045–1054.
31. Kryshtafovych,A., Schwede,T., Topf,M., Fidelis,K. and Moult,J. (2023) Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins Struct. Funct. Bioinf.*, **91**, 1539–1549.
32. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
33. Sagendorf,J.M., Mitra,R., Huang,J., Chen,X.S. and Rohs,R. (2024) Structure-based prediction of protein–nucleic acid binding using graph neural networks. *Biophys. Rev.*, **16**, 297–314.

Received: September 8, 2024. **Revised:** October 7, 2024. **Editorial Decision:** October 9, 2024. **Accepted:** October 11, 2024