Current Research in Structural Biology 2 (2020) 130–143

Contents lists available at ScienceDirect

Current Research in Structural Biology

journal homepage: www.journals.elsevier.com/current-research-in-structural-biology

Understanding the structural details of APOBEC3-DNA interactions using graph-based representations

J.C.-F. Ng, F. Fraternali *

Randall Centre for Cell and Molecular Biophysics, King's College London, United Kingdom

ARTICLE INFO

Keywords:
APOBEC3
Protein-DNA interaction
Protein structural networks
Structural bioinformatics

ABSTRACT

Human APOBEC3 (A3; apolipoprotein B mRNA editing catalytic polypeptide-like 3) is a family of seven enzymes involved in generating mutations in nascent reverse transcripts of many retroviruses, as well as the human genome in a range of cancer types. The structural details of the interaction between A3 proteins and DNA molecules are only available for a few family members. Here we use homology modelling techniques to address the difference in structural coverage of human A3 enzymes interacting with different DNA substrates. A3-DNA interfaces are represented as residue networks ("graphs"), based on which features at these interfaces are compared and quantified. We demonstrate that graph-based representations are effective in highlighting structural features of A3-DNA interfaces. By large-scale *in silico* mutagenesis of the bound DNA chain, we predicted the preference of substrate DNA sequence for multiple A3 domains. These data suggested that computational modelling approaches could contribute in the exploration of the structural basis for sequence specificity in A3 substrate selection, and demonstrated the utility of graph-based approaches in evaluating a large number of structural models generated *in silico*.

Introduction

Human APOBEC3 (apolipoprotein B mRNA editing catalytic polypeptide-like 3; hereafter abbreviated as A3) enzymes refer to a family of seven cytidine deaminases which causes cytidine-to-uridine (C→U) mutations on single-stranded DNA (ssDNA) molecules. Targeting nascent reverse transcripts of retroviruses such as HIV-1 (human immunodeficiency virus-1) (Sheehy et al., 2002), A3 proteins contribute to restrict retroviral replication and hence have an important role in immune defence (Malim, 2009). The mutagenic function of A3, however, is not restricted to the viral genome, as DNA sequence variations observable in many human cancers are recently found to be attributable to the action of A3 (Alexandrov et al., 2013). A3 activity is found to be specific towards DNA sequence of the pattern 5′-TCA/T-3′, based on which a "mutational signature'' attributable to A3 activity could be extracted and quantified (Figure S1, Supplementary Material) from cancer genomes. While the functional contexts under which A3 mutagenesis occurs are extensively studied (e.g. in Ref. (Ng et al., 2019)), we only begin to understand the structural basis of this process, particularly the details of interaction between A3 proteins and DNA "substrates'' of specific topology and sequence.

This has been made possible with a growing number of A3-nucleic acid crystal structures which are experimentally determined and deposited in the Protein Data Bank (PDB). A3 proteins are almost entirely composed of cytidine deaminase (CDA) domains, which are capable to bind both DNA and RNA (Figure S2A, Supplementary Material). However, only ssDNA could serve as the deamination substrate for all A3 members (Smith et al., 2012). The CDA domains themselves are highly identical (with sequence identity as high as 90%) to one another on sequence terms (Figure S3A, Supplementary Material); A3 proteins vary in terms of the number of CDA domains they contain. While three out of seven human A3 proteins (APOBEC3A [A3A], APOBEC3C [A3C] and APOBEC3H [A3H]) contain only one CDA domain, the remaining four contain two of such domains, with catalytic activity only detectable in the C-terminal CDA, whereas the N-terminal CDA is responsible for oligomerisation and interaction (purportedly non-sequence-specific) with RNA molecules (Salter & Smith, 2018). Structural investigations have identified regions of the A3 CDA structure which are functionally important. Notably, the four loops (loops 1, 3, 5 and 7, hereby referred to as the "gating'' loops) surrounding the active site are known to control both access and specificity of the substrate (Fig. 1A; also Figure S2B, Supplementary Material) (Salter & Smith, 2018; Silvas & Schiffer, 2019;

* Corresponding author.
E-mail address: franca.fraternali@kcl.ac.uk (F. Fraternali).

https://doi.org/10.1016/j.crstbi.2020.07.001
Received 5 May 2020; Received in revised form 17 July 2020; Accepted 21 July 2020

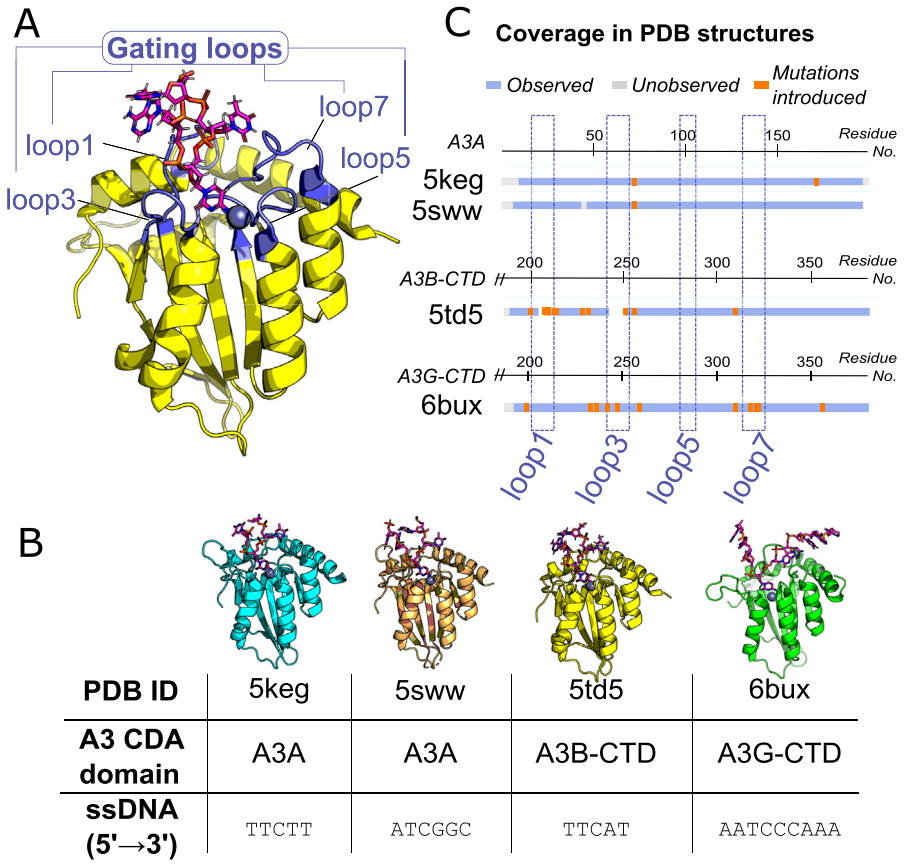

Figure 1. Summary of A3-DNA structures used for this analysis. (A) Annotation of the "gating loops'' on A3 CDAs. The A3B–CTD complex with ssDNA (PDB 5td5) is shown. (B) Co–crystal structures of A3 CDA domains with ssDNA deposited in the Protein Data Bank (PDB) used in this analysis. (C) Structural coverage of A3-DNA structures. Regions observed in the crystal structure are labelled in blue, while residues unobserved are labelled grey. Deleted regions are represented as white gaps; residues labelled orange have been mutated in the construct used to resolve the structure. Data from PDBe-KB (Varadi et al., 2020). The gating loops as labelled in panel (A) are indicated accordingly. See Supplementary Fig. S4 for coverage in all existing PDB structures for these domains.

Hou et al., 2019; Salter et al., 2016). It has already been known that for the A3 CDAs in *apo* state (i.e. without bound nucleic acid substrates), the size of the binding pocket varies, with substrates having much more restricted access to the active site of the C-terminal CDA of APOBEC3B (hereafter A3B-CTD), in comparison to A3A and the C-terminal CDA of APOBEC3G (hereafter A3G-CTD). This is coordinated by stacking interactions between loop1 and loop7 (Shi et al., 2015). These structural details have been further clarified by a handful of recent crystal structures which capture the conformation of A3 CDAs in complex with RNA and ssDNA of different sequences and topologies (Shi et al., 2017a; Shi et al., 2017b; Kouno et al., 2017; Maiti et al., 2018; Fang et al., 2018; Shaban et al., 2018). So far A3A (Shi et al., 2017b; Kouno et al., 2017), as well as A3B-CTD (Shi et al., 2017b) and A3G-CTD (Maiti et al., 2018) have been captured with substrate DNA molecules of different sequences and topologies at the active site (Fig. 1B). The DNA binding mode observed in these structures resembles the way in which the *Staphylococcus aureus* tRNA adenosine deaminase, a distant homologue, binds its substrate tRNA (Losey et al., 2006) (Figure S2A, Supplementary Material). A handful more structures captured nucleic acid molecules bound to A3 CDAs, but distal to the active site (Fang et al., 2018; Shaban et al., 2018). Detailed comparative analysis of A3-DNA complexes would reveal insights into the differences of their DNA-binding behaviours, in terms of both the sequence and the topology of the bound DNA, across different A3 CDAs domains. However, to date none of such studies have been published.

Two important caveats regarding structural studies of the A3 CDAs deserve careful consideration: first, all of the aforementioned A3-DNA structures used constructs which introduced sequence deletions and/or mutations to enhance protein solubility (Fig. 1C). Some structures also have missing residues within the flexible loops. Importantly, these deviations from the wild-type sequence mainly affect the gating loops. Since these loops are important to control substrate access, the binding mode observed using these modified/incomplete constructs might not reflect the true mechanism of substrate recognition. Second, structures of A3 CDAs co-crystalised with their nucleic acid substrates are available to only A3A, A3B-CTD and A3G-CTD to this date. Therefore, a systematic survey across the entire human A3 family would necessitate the need to employ techniques such as homology modelling, to address the lack of structural coverage for all the other A3 family members, and to extract the principles of substrate recognition with full-length, wild-type CDA domains.

Such examination of substrate recognition across the family also necessitates the use of tools which are capable of unbiased extraction and large-scale comparison of features at our interfaces of interest, if one intends to take a combinatorial approach to examine different substrate topologies across the entire panel of A3 CDAs. Here we represent protein structure as graphs of interactions between residues, with Cα atoms of residues as nodes, and edges specified by the pairwise Cα-Cα distances; features local to the interface are therefore known as *subgraphs*. Graphs have been used to represent protein assemblies, generating "atlas'' of possible topological arrangements of assemblies for various protein families (Heal et al., 2018); they are also amenable as a tool to model protein dynamics, by treating residue–residue interactions as molecular springs (Bahar et al., 1997; Bahar et al., 2010; Papaleo, 2015). Here we use graphs simply as representations of protein structures, reducing the complexity of residue interactions maps from 3D to 2D, while retaining important information (both topological and physico-chemical) of protein structural features. These graph-based features also lend themselves to various theoretical measures to quantify, for example, "hub'' characteristics of different residues, revealing the relative importance of specific spatial arrangements (Doncheva et al., 2012; Chakrabarty & Parekh, 2016). These graph-based representations could therefore be harnessed as a tool to extract and compare structural features of our interfaces of interest.

In this study, we applied graph-based representations of protein structures to study A3 CDAs, with a focus on the gating loop conformations. Since sequence modifications have been introduced to obtain A3-DNA crystal structures, we generated full-length, wild-type A3 CDA models using homology modelling, and grafted DNA substrates of different sequences and topologies onto these models. This generates a panel of A3 CDAs, in both their *apo* states and in complex with DNA. Using these structural models we aim to extract the principles which govern the recognition of the preferred substrates of these CDA domains. By defining residue graphs at the A3-DNA interfaces, we demonstrate that Cα subgraphs capture substantive differences at the gating loops across A3 CDAs, and provide an extra layer of quantitative parameters to describe the conformational landscape of such loops. We also performed a large-scale *in silico* mutagenesis of the bound DNA chain to compare the preference of substrate DNA sequence for multiple A3 domains. These data demonstrate the use of computational modelling in exploring the structural basis for sequence specificity in A3 substrate selection, and showcase graph-based approaches could be effectively applied to evaluate a large number of structural models generated *in silico*.

Materials and Methods

Analysis of selected APOBEC3 structures in the Protein Data Bank

Both structures of APOBEC3 CDAs on their own, as well as in complex with nucleic acid molecules (Table 1), were considered in this work. The two complex structures of A3A with ssDNA (PDB 5keg and 5sww; (Shi et al., 2017b; Kouno et al., 2017)) were largely identical to each other in terms of binding site arrangement and nucleic acid topology, but the sequences of the DNA molecule were slightly different at the + 1 position (Fig. 1B). In Section Homology modelling, *apo* A3 CDA structures used in this work as templates for homology modelling are listed.

Definition of protein-nucleic acid interface

To define, for each PDB structure, interface between protein and nucleic acid (if the structure contains so) and/or interface between monomers (in an oligomeric structure), the POPSCOMP (Kleinjung & Fraternali, 2005) software was used. The algorithm calculates the change in solvent-accessible surface area (SASA) upon complex formation, based on SASA calculations of single protomers implemented in POPS (v2.3; (Cavallo et al., 2003)). Residues with a change in relative SASA > 15% upon complex formation were extracted as interface residues (Fornili et al., 2013).

Homology modelling

Not all A3 CDAs have experimentally determined structures, and for those which do, many of those structures contain designed mutants (in aiding protein expression and crystallisation) and/or missing residues (Shi et al., 2017b; Kouno et al., 2017; Maiti et al., 2018) (Fig. 1C). Therefore, homology modelling was performed on all 11 human A3 CDAs, where the wild-type sequence (UniProtKB) was modelled over either the corresponding PDB structure or that of its closest homologue (see Table 2). T-coffee (Notredame et al., 2000) was used to generate sequence alignments. MODELLER (v9.17; (Webb & Sali, 2017)) was used in a two-step modelling process: first, the automodel mode was used to generate 200 decoys. The decoy with the lowest DOPE score (which used statistical potentials to score the decoys in terms of their energetic favourability; see Ref. (Shen & Sali, 2006)) was selected. Second, this selected decoy was further subjected to loop refinement. Loop definitions were taken from DSSP (Kabsch & Sander, 1983) analysis of this selected structure. 200 loop-refined decoys were generated. Two decoys were selected for grafting different nucleic acid chains (see below, section Graft generation); this gave rise to two ensembles of A3-DNA grafts for each domain, which enables evaluation of how the starting configuration impacts on the quality of grafts generated.

Table 1
APOBEC3-DNA complex crystal structures examined. ssDNA: single-stranded DNA. Residues in the α3 range were used for superposition with APOBEC3 CDA models to generate APOBEC3-DNA grafts.

PDB ID	Protein chain	Nucleic acid chain	α3 range	Molecules
5keg	A	B	105–119	A3A with ssDNA
5sww	A	E	105–119	A3A with ssDNA
5td5	A	C	288–302	A3B-CTD with ssDNA
6bux	A	B	289–304	A3G-CTD with ssDNA

Table 2
APOBEC3 CDA homology models, their corresponding template and regions of interest.

Domain	Template	α3range	Zn^{2+} coordination		
A3A	4xxoA	105–119	H70	C101	C106
A3B-NTD	5tkmA	98–111	H66	C97	C100
A3B-CTD	5cqiA	288–302	H253	C284	C289
A3C	3vowA	98–111	H66	C97	C100
A3D-NTD	5k81A	110–123	H78	C109	C112
A3D-CTD	5hx5A	294–307	H262	C293	C296
A3F-NTD	5k81A	97–110	H65	C96	C99
A3F-CTD	5hx5A	281–295	H249	C280	C283
A3G-NTD	5k81A	98–111	H65	C97	C100
A3G-CTD	3v4kA	289–304	H257	C288	C291
A3H	5w45A	86–99	H54	C85	C88

APOBEC3-nucleic acid grafts

Graft generation

Poses of A3-DNA complexes were taken from the available PDB structures (Table 1 and Fig. 1). Using PyMOL each of these structures was superposed (using the command super) with every modelled A3 CDAs, and generated "grafts'' of A3-nucleic acid complexes. The coordinates of the protein and DNA chains were stored together in one PDB file. The orientation of both the protein and the nucleic acid chains were manually inspected. A total of two ensembles of grafts were generated for each domain, using different starting poses:

- **Ensemble *apo***: The protein chain was taken to be the model with the lowest DOPE score from MODELLER. Note here the protein was modelled in *apo* form, i.e. the binding and configuration of single-stranded DNA was not considered in the scoring and selection.
- **Ensemble *in situ***: The protein chain was taken to be the model (generated using the procedure detailed in section Homology modelling) which best resembles the A3 domain that was resolved in complex with the DNA chain. For example, to generate a graft of the DNA chain in PDB 5td5 onto A3A, the A3A conformation used was the A3A model which has the lowest root-mean-squared deviation (RMSD) when superposed with the A3 domain in 5td5. In comparison with Ensemble *apo*, since the positioning of the DNA was taken into account in selecting the starting protein conformation, this minimises clashes between the two chains in the starting pose.

The grafted poses were subject to cycles of energy minimization and repacking using Rosetta (source code version 2018.33.60351 bundle; (Leaver-Fay et al., 2011)), with the relax protocol and the dna scoring function (Ashworth et al., 2006; Ashworth et al., 2010), restraining around the starting pose. The Rosetta software reads in the atomic coordinates, detects missing atoms and submits the structure to optimization (see Section S1.1 in the Supplementary Materials for further details). Here the protein side-chains underwent repacking while the backbone arrangements for the nucleic acid chains were observed to remain unchanged. Both ensembles were generated using this identical protocol and the same scoring function. In each case, 200 decoys were generated to constitute an ensemble.

Residue graphs

Extraction of residue graphs

For each interface we considered only the entity/entities which is/are (a) protein(s). Functionalities provided in the PDB module of Biopython were used to extract residue graphs and calculate residue–residue distances. We extracted neighbours of the interface residues, defined as all residues whose Cα atoms are within a specific distance threshold from that of the interface residues. This list of Cα′s would be the list of nodes in the graphs. Edges were then drawn if the Cα-Cα distances are within a certain distance cut-off. Here, the cut-offs of [6 Å, 8 Å, 10 Å] have been considered. In Figure S6 (Supplementary Materials) the effect of altering this distance cut-off to the dimension (numbers of nodes and edges, and thus the size) of the graph is illustrated.

These cut-off values span the range of distance cut-offs used in other applications of network construction on protein structural data [e.g. Ref. (Bakan et al., 2011)]. The issue regarding range of residue–residue distances suitable in constructing residue graphs has been explored and discussed elsewhere (Salamanca Viloria et al., 2017). This is related to how nodes are defined: some choose to define centres of mass of (either side-chains or the entire) residues as nodes, rather than using Cα atoms as was done here. Using the former definition Salamanca Viloria and colleagues (Salamanca Viloria et al., 2017) have assessed a diverse set of structures, and suggested an optimal distance cut-off of 5 Å; considering here we refer to Cα-Cα distances, this would correspond to ≈ 7–8Å. The lower bound would be commonly-used thresholds in structural biology for defining contacts, which is in the range of 4–4.5 Å (Salamanca Viloria et al., 2017); that would map roughly to around 6–7Å when Cα-Cα distances are concerned.

All pairwise atomic distances between the Cαs of interface residues and that of the neighbours were calculated. We then generated, for each interface, a residue graph as an representation (See Fig. 3A for an example). This "subgraph'' (i.e. part of the graph describing structure of the entire protein/domain) constitutes the "fingerprint'' of the interface, and represents the basis for comparison with other proteins.

Evaluation of residue graphs

Residue graphs were visualised with PyMOL (v2.1.0; (Schrödinger, 2017)) programmatically using functionalities implemented under the python application programming interface (API). Residue graphs were also analysed quantitatively (by, e.g. calculating centrality measures specific to each node [i.e. residue] in the graph) using the R igraph package. A number of centrality measures are calculated for each residue in the graph; these measures quantify the importance of vertices within a graph, based on the connectivity and therefore the possible flow of information from one vertex (residue) to another. Specifically, four examples of such centrality measures were calculated: (i) degree, (ii) betweenness centrality, (iii) closeness centrality, and (iv) eigencentrality. See section S1.2 in Supplementary Material for a more detailed description of how they are calculated. The majority of these metrics were originally developed to describe social networks, and subsequently borrowed for biological networks; see e.g. Ref. (Freeman, 1978; Freeman, 1980; Borgatti & Everett, 2006; Negre et al., 2018) for detailed account of their derivation and application. Here, the calculation of all four centrality measures listed above were performed using the igraph package in R, on residue graphs defined on A3-DNA crystal complexes, as well as the generated graft poses in the modelling procedure (see section Graft generation).

A geometric approach to evaluate subgraph-nucleic acid interactions

We also evaluated how the subgraphs are oriented with respect to the nucleic acid molecule. Specifically, while the subgraph can be in direct contact with part of the nucleic acid molecule, in certain cases, especially in grafts generated using the *in silico* procedure above, the nucleic acid molecule may not be oriented in contact with a subgraph in a chemically plausible manner, possibly due to the inability to reject such solutions in the sampling procedure. A visual inspection of some poses reveal, that in certain cases the nucleic acid appears to be embedded in a long loop (either at loop1 or loop3) at the CDA domain, i.e. the nucleic acid chain appears to "penetrate'' through the loop, thus creating an "entanglement'' (Fig. 7A). In principle, one should not obtain such unphysical structures, but as it happens the optimization and scoring procedure is not immune to misleading incorrect geometries in these cases. This motivated the development of a geometry-based procedure to evaluate whether a subgraph is a physicochemically plausible interface, by detecting nucleic acid entanglement in the structure. A plausible CDA-substrate interaction was said to occur only if there was no nucleic acid entanglement in the complex, as such entanglements are results of physically unrealistic conformations.

The entanglements were broadly classified into two types:

- **"Chain-entanglement"**: the nucleic acid chain passes through the loop.
- **"Base-entanglement"**: for a certain position within the nucleic acid chain, its base penetrates through the loop.

Here we have used a geometric approach to detect these entanglement phenomena, amounting to an automatic detection without the need of manual inspection. For either case, Cα atoms on the loop defines a triangle on a plane **P** (Fig. 2). One has then to specify two atoms in the nucleic acid chain/base, which in turn define a vector $\vec{v}$. This problem of detecting entanglement is thus reduced to examining how $\vec{v}$ projects to the triangular plane **P**. One then has to compare different definitions of $\vec{v}$ and the triangle on **P** to map the location of the entanglement.

Testing for entanglement. See Supplementary Materials, Section S1.3 for a detailed description of the mathematical procedures for testing entanglements. For a given protein loop and nucleic acid substrate, different definitions of **P** were considered (by using different residues to define the vertices; see Supplementary Methods S1.3), over every consecutive pair of positions along the nucleic acid substrate to define $\vec{v}$. Therefore, under such scheme we could map the exact location where entanglement occurs. All geometry calculations were performed using the sympy (v1.4) package in Python unless otherwise stated above. Entanglements were also visualised for manual inspection with PyMOL.

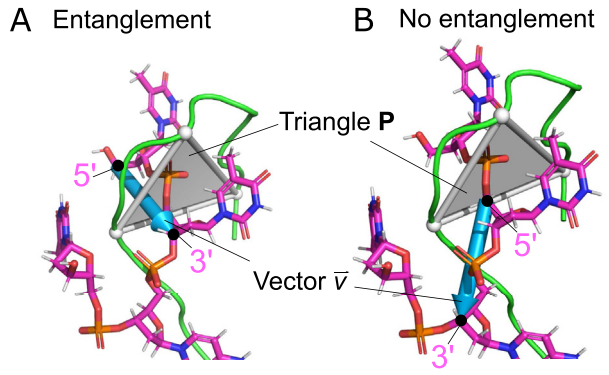

Figure 2. A geometrical approach for automatic protein-nucleic acid entanglement detection. Triangles **P** are defined by Cα atoms on the protein loop of interest (depicted as grey triangles). On the nucleic acid side, a vector $\vec{v}$ (blue arrow) is defined to represent a chain of nucleic acid sequence. This is taken to be the C5′ and C3′ atoms of two adjacent DNA bases in the chain (labelled accordingly here). By examing the geometrical relationships between **P** and $\vec{v}$, this allows for distinguishing between the presence (A) and absence (B) of entanglement. See Figure S5 in the Supplementary Materials for further explanation.

Evaluation of intermolecular interactions

Geometric methods were used to evaluate interactions between the DNA and protein chains in the original A3-DNA complexes and our generated poses. Two types of intermolecular interactions were examined here, namely hydrogen bonding and π-π stacking, as both of them were documented to be important in mediating A3 interaction with ssDNA. Each type of intermolecular interactions was detected as detailed below:

Hydrogen bonds. Possible hydrogen bonds were detected between the DNA and protein chains for each structure examined, using the findhbond tool in UCSF Chimera (v1.13.1; (Pettersen et al., 2004)). Each structure was first loaded into the Chimera software, and the nucleic acid chain was selected. Hydrogen bonds with exactly one end (either the donor or the acceptor atom/atom group) on the nucleic acid chain was detected. We did not restrict the detection to specific residue/atom types, therefore salt bridges are also included in this analysis. Since water molecules were omitted in the Rosetta-simulated poses, hydrogen bonds involving water molecules detected in the original structures were not considered in the analysis. The atoms involved in the hydrogen bonds were tabulated and analysed.

π-π stacking. Unlike hydrogen bonds, the detection of π-π stacking has not been implemented in popular molecular visualisation software like UCSF Chimera or PyMOL. Therefore we implemented a procedure for automatic detection of such interactions, with reference to approaches enlisted in the literature [e.g. Refs. (Ferreira de Freitas & Schapira, 2017)]: possible π-π stacking was detected by first listing residues/bases containing aromatic groups and then examining the pairwise distance between their centres of mass. If this distance is smaller than 6 Å, the angle between the planes defined by the two aromatic groups were detected. Note here a very generous distance cut-off was used (typical distances are in the range of 3.5-4Å; see Ref. (Mills & Dean, 1996; Li et al., 2017), such that imperfect but possible interactions (as artefacts of the modelling and optimisation procedure) to be still detectable. Three atoms were used for each planar residue/base to define such plane (Table 3).

DNA motif optimisation

To map the mechanism of substrate recognition, we reasoned that based on the A3-DNA grafts, by allowing the DNA substrate sequence to vary subject to the local physico-chemical environment, one could en-

Table 3
Atom types used in defining planes for π-π stacking detection. Atom types are defined using worldwide PDB (wwPDB) definitions (Burley et al., 2019).

Biomolecule	Residue/Base	Atom (wwPDB definition)
Protein	HIS	CB CE1 NE2
	TYR	CB CE1 CE2
	TRP	CB CZ1 CZ3
	PHE	CB CE1 CE2
DNA	A & G	N9 C2 C6
	T & C	N1 N3 C5

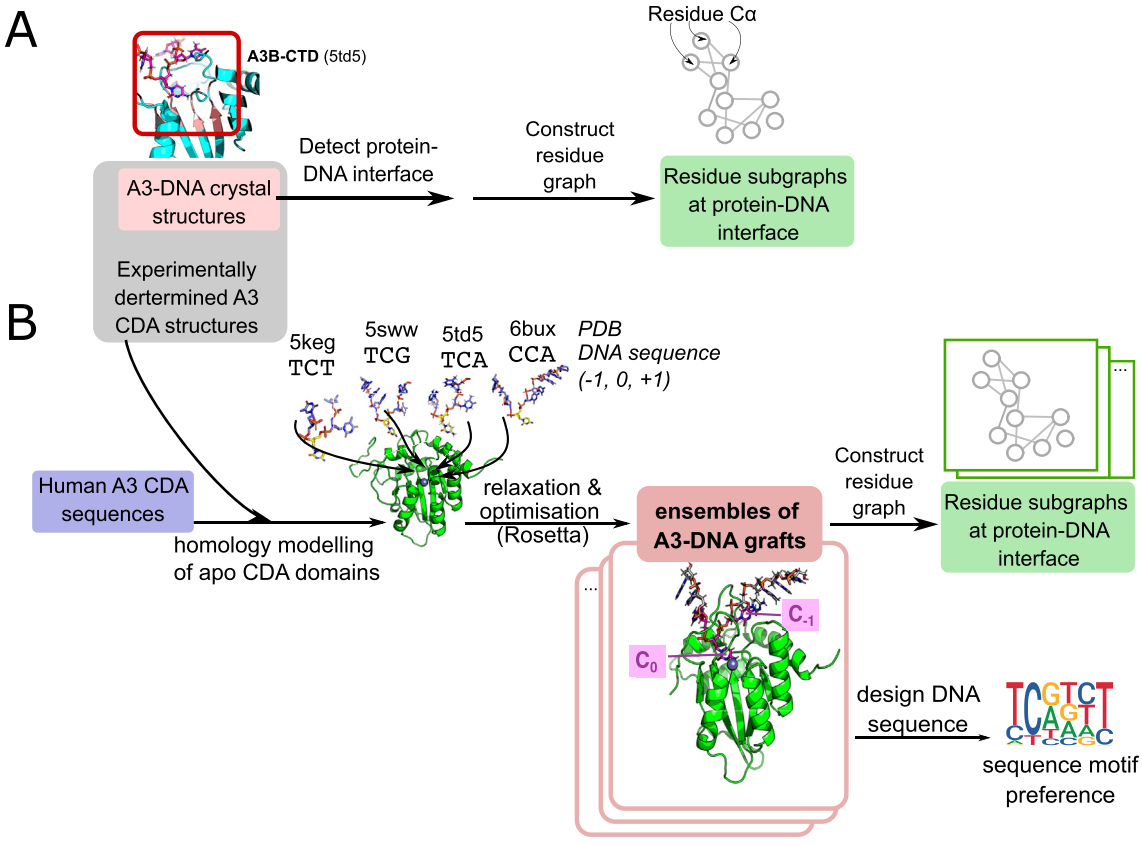

Figure 3. Overall schematic of this study. (A) Protein-DNA interfaces in A3-DNA co–crystal structures were extracted using POPSCOMP (Kleinjung & Fraternali, 2005) and residue subgraphs representing these interfaces were constructed (see Methods). (B) Homology modelling was used to model all human A3 domains. This was followed by grafting onto the modelled structure different nucleic acid substrates used in existing A3-DNA crystal structures. By subsequently allowing the structure to relax and optimise, we generated ensembles of A3-DNA grafts, based on which ensembles of residue-based graphs were extracted. Moreover, these residue-based grafts ensembles were subject to *in silico* design of the substrate DNA sequence. This allows for probing the DNA sequence preference of each A3 CDA domain.

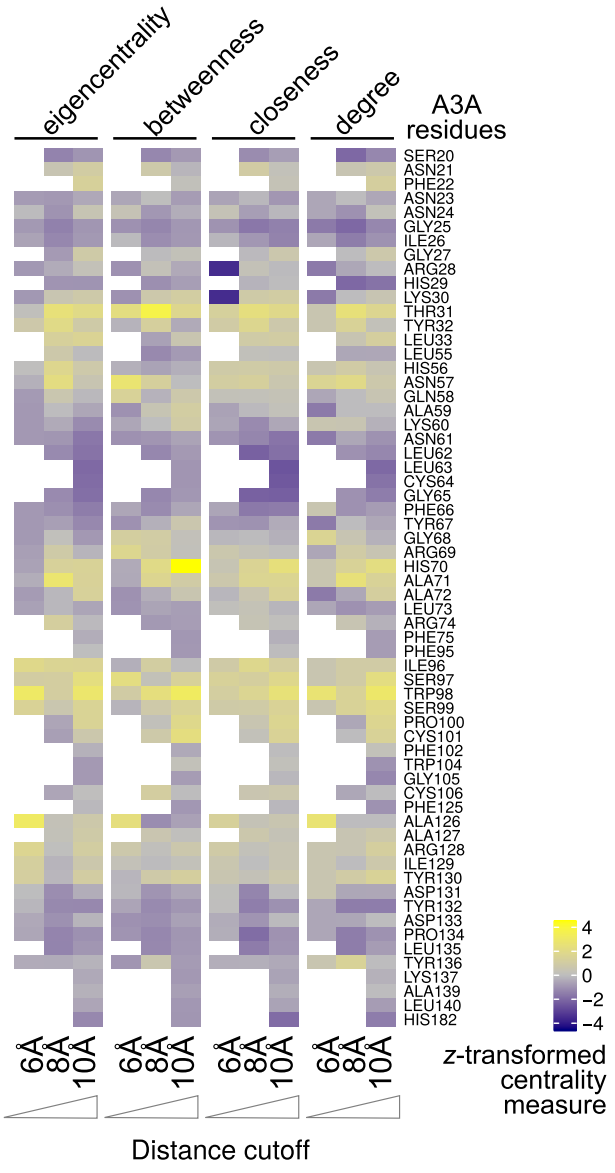

Figure 4. Network parameters calculated on interface residue graphs defined using A3-DNA crystal structures. Here the A3-DNA structure (PDB 5sww) is depicted as an example. The eigencentrality, betweenness, closeness and degrees of each A3A residue in the graph are depicted as heatmaps, comparing their values in graphs defined using different distance cut-offs. These centrality measures were normalised in each graph (i.e. by the columns) using a *z*-transformation, such that those residues with a higher centrality measure would be depicted with a yellow grid, whereas those with lower centrality depicted dark blue. Blank grids indicates that these residues were not included in the relevant graphs. See Figure S8 in Supplementary Material for plots using other A3-DNA crystal structures considered in this work.

gineer the known substrate preference for a given A3 CDA domain. Here an DNA motif optimisation experiment for each of the grafts generated using the identical Rosetta protocol as detailed above for assessing DNA-binding: identical re-packing was performed, but here Rosetta was allowed to mutate any DNA base, including the catalytic C (at position 0) in each DNA pose. For each ensemble of grafts, on the 200 decoys generated, each was allowed for 50 runs for such motif mutation procedure, thus generating in total 20,000 poses (= 2 ensembles × 200 decoys × 50) with varying DNA sequences for each graft.

Statistics and data visualisation

All data visualisation have been performed in the R statistical computing environment (v3.4.4). Plots were generated using plotting functionalities in base R and the ggplot2 package (Wickham, 2016). Sequence logos were generated using the ggseqlogo package (Wagih, 2017) in R. Heatmaps were produced using the gplots package (Warnes et al., 2019).

Results

In this work we study the DNA-binding interface of the A3 enzymes, using a comparative approach by analysing all the 11 human A3 CDAs. Homology modelling techniques were applied to model A3 CDA domains without experimentally determined structures, as well as modelling the binding of DNA substrates of different sequences and topologies with these domains. Importantly, using such approach we could consider full-length, wild-type A3 CDA domains, thereby overcoming the sequence modifications present in existing A3-DNA crystal structures. Structural features were assessed by using networks ("graphs") to represent the structure of a protein. These graph-based representations lend themselves to application of various metrics developed to describe graph topologies, to identify, for instance, important residues acting as "hubs'' of the graph. To define a graph, Cα atoms of residues are taken as nodes; edges are only drawn between residues of a pairwise Cα-Cα distance within a certain threshold; we will explore below the impact of altering such threshold to define graphs of different stringency. Information about the exact amino acids (identity and residue position) are retained in the nodes (Fig. 3A). This results in a coarse-grained representation of the spatial arrangement of a protein interface of interest. Using the various functionalities available in the Rosetta suite, this also allows for examination of ensembles of A3-DNA interactions (obtained by repacking protein side-chains), and designing optimal substrate sequence for different A3 CDAs (Fig. 3B).

Graph topologies at the A3 DNA-binding interface

We applied graph-based representations to compare DNA-binding interfaces of A3 CDAs, making use of a panel of recently determined crystal complexes of various A3 CDAs, with DNA substrates of different sequences and topologies (Fig. 1B; (Shi et al., 2017b; Kouno et al., 2017; Maiti et al., 2018)). Using POPSCOMP (Kleinjung & Fraternali, 2005) to define protein-DNA interfaces, subgraphs were generated for these protein structural regions (Fig. 3A).

We first compared the effect of distance cut-offs on graph properties. As one would intuitively expect, graphs of larger sizes would be extracted if a larger distance cut-off is used (Figure S6, Supplementary Materials). Moreover, when different node centrality measures (see section Residue graphs in Methods; also Section S1.2, Supplementary Material) are considered, a more continuous trend of these values could be observed in graphs generated under larger distance cut-offs (Fig. 4), in comparison to graphs defined using small cut-off values, where large jumps in centrality values between adjacent residues can often be observed. As we will see later, centrality measures correlate with conventional structural features, connecting graph features with commonly used concepts in protein structural analysis.

Graph extraction on ensembles of homology models

Whereas the resolved A3-DNA crystal complexes have proved to be applicable in this analysis, a few issues remain to be addressed. First, all of these structures were resolved on modified protein sequence of the A3 CDA in question (Fig. 1C), with some introducing substitutions at particular amino acid positions, and others deleting or grafting part of the flexible loops around the active site (Shi et al., 2017b; Kouno et al.,

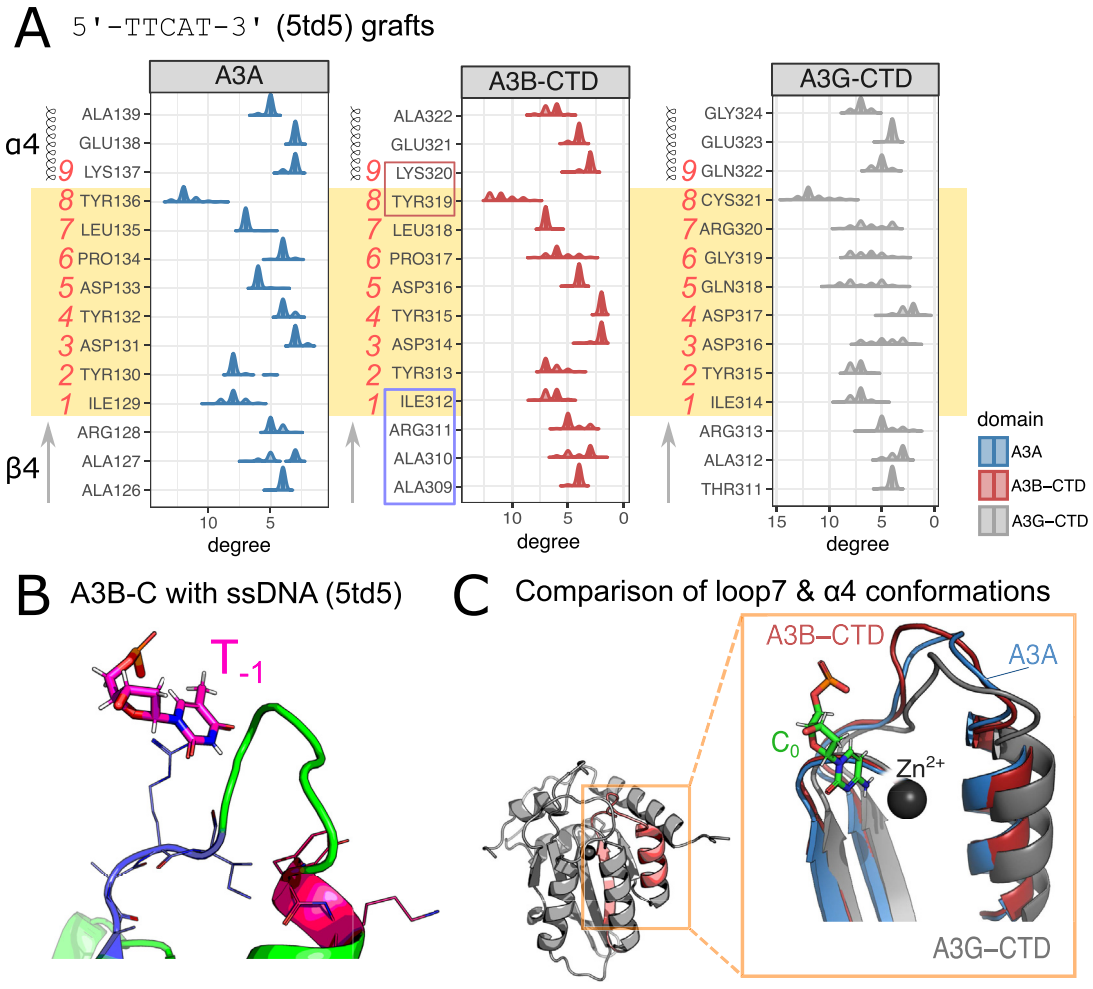

Figure 5. Graph analysis of loop7 of A3-CDAs grafted with DNA. (A) Degree distributions in loop7 residue subgraphs from Ensemble *in situ*. Here grafts of 5′-TTCAT-3' (from PDB 5td5) onto A3A, A3B-CTD and A3G-CTD are considered. Residues in loop7 are highlighted in yellow. Cα graphs were constructed with distance cutoff of 8 Å. As subgraphs are extracted from an ensemble of 200 poses, here these are depicted as density curves. Residues were numbered for illustration later (e.g. Figure 6). For A3B-CTD, some residues were shadowed (purple-blue/pink); these are depicted in panel (B). Secondary structure annotation follows that of an A3A crystal structure (PDB 4xxo). Poses depicted here were from Ensemble *in situ*, i.e. generated starting from the configuration which best resembles the 5td5 crystal structure. See Supplementary Fig. S9 for comparison of other centrality measures, and results from Ensemble *apo*. (B) "Hub'' residues on A3B-CTD crystal structure complexed with single-stranded DNA (ssDNA, PDB 5td5). Residues with high degrees in panel (A) are depicted with sticks and colour-coded as indicated in (A). The −1 ssDNA position was also indicated. (C) A comparison of conformations around loop7 (highlighted in inset) of A3A, A3B-CTD and A3G-CTD. Loop7 of A3A (blue) and A3B-CTD red are more compact than that of A3G-CTD (grey). Consequently for A3G-CTD the following α-helix shifts outward, away from the deamination site represented by the catalytic C_0 and the zinc ion (black sphere).

2017; Maiti et al., 2018). Therefore, the structural details captured in these structures may not represent the true contexts of substrate binding. Second, parts of these loops around the active site are involved in crystal contacts (Figure S7, Supplementary Material). Third, these structures represent a snapshot of interaction between the nucleic acid chain and the CDA domain; this is important in our application here, since for A3 CDAs substrate recognition involves flexible loops. Therefore, it is necessary to inspect *ensembles* of structures to obtain a more representative sample of the conformational diversity of these regions. We therefore used *in silico* tools to construct structural ensembles of wild-type, full-length A3 CDA domains (Fig. 3B, also see Methods), by first performing homology modelling with MODELLER (Webb & Sali, 2017; Sali et al., 1995) to "repair'' these structures; this step placed the wild-type sequence into the modified positions. We then re-introduced the coordinates of the DNA chain back to these "repaired'' structures, and used the Rosetta suite (Leaver-Fay et al., 2011) for molecular modelling, where multiple poses were generated and evaluated with scoring functions designed for protein-DNA interactions (Ashworth et al., 2006; Ashworth et al., 2010). These produced ensembles of full-length A3 CDAs complexed with DNA, based on which graph extraction could be performed (Fig. 3B). Considerations have been given to minimise possible clashes between the incoming DNA chain and the full-length domains, which were modelled on their own (see Methods, section Graft generation). These models generally feature similar shapes in terms of their DNA-binding interface (Root--mean-squared deviation [RMSD] in the range of 0.5-2Å; Figure S3, Supplementary Material) while revealing substantial conformational diversity (RMSD up to 4-5Å) in some cases. We observed distributions of centrality measures per residue, reflecting differences within the ensemble (Fig. 5A). Interestingly, hinges of flexible loops appear to be "hub'' residues in these subgraphs with high centrality values (Fig. 5A–5B). The difference between residues within and outside of the

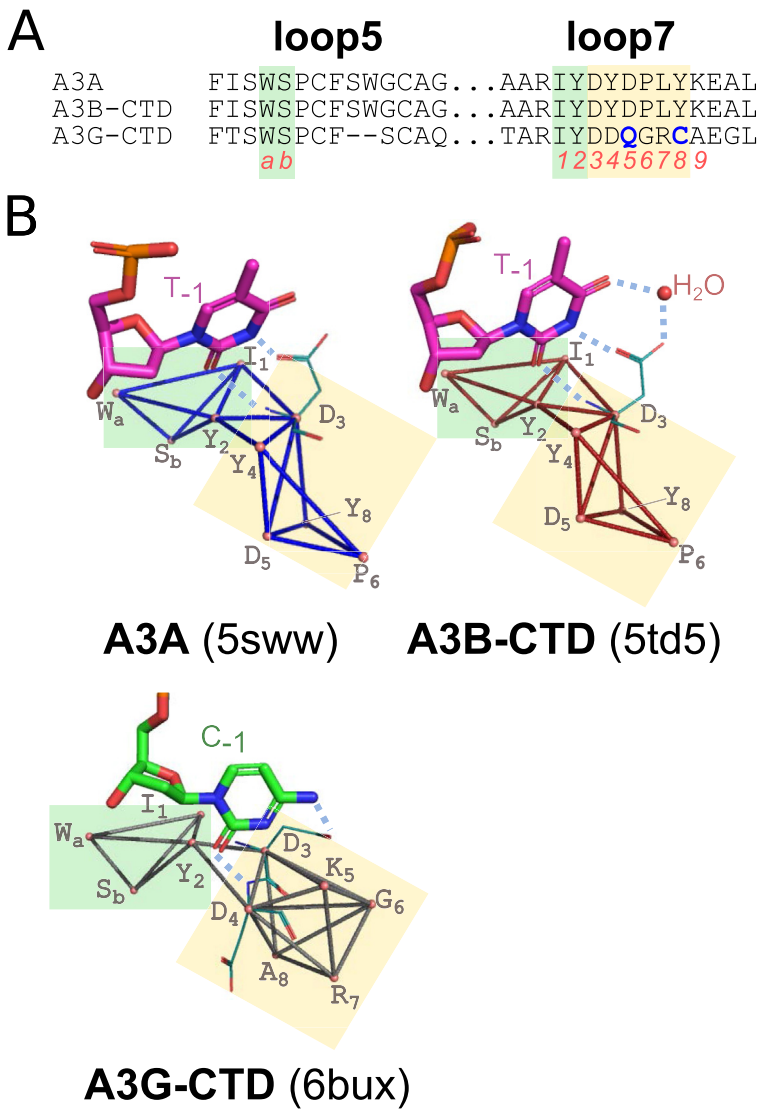

Figure 6. Residue graphs at the catalytic site in A3-DNA crystal structures. (A) A sequence alignment of A3A, A3B-CTD and A3G-CTD at loop5 and loop7. The deaminated cytidine binds to loop5 and loop7 dictates DNA base preference at the −1 position. Residues depicted in graphs in panel (B) were highlighted in green or yellow and labelled with letters/numbers for convenience. (B) Residue subgraphs of the DNA interface for A3A, A3B-CTD and A3G-CTD. Residues were labelled using the scheme illustrated in (A). Conservation of the catalytic motif (subgraphs shaded green) manifests as identity in topology in 3D; variations in the sequence translates to variations in 3D topology in the subgraphs extracted (shaded yellow). Here the base at position −1 was represented by sticks. Side chains responsible for base recognising was depicted as turquoise thin lines. Hydrogen bonds responsible for recognition of the −1 base were taken directly from the literature describing these crystal structures (Shi et al., 2017b; Maiti et al., 2018), represented by blue dashed lines.

hinges, however, appears to be less pronounced for A3G-CTD (Fig. 5A). Inspecting the conformation of loop7, A3A and A3B-CTD appears more compact than A3G-CTD, which leads to the following α-helix (*α*4) shifting outwards from the catalytic site (Fig. 5C). These demonstrate that graph analyses can provide complementary details to conventional structural analyses, and potentially offer insights into subtle differences between the activity of different A3 CDA domains.

As graphs could become prohibitively large for individual inspection for larger distance cut-offs (see above), here graphs extracted using a small distance cut-off (6 Å) are displayed (Fig. 6). These graphs reveal a complex "web'' of cycles of residue arrangements at the catalytic site, mapping to loop5 and loop7 of the CDA domain (Fig. 6). Notably, by visualising these subgraphs obtained from different A3 domains, patterns of conservation and evolution of residues local to the catalytic site can be observed. For example, the subgraph visualisation highlighted the conservation at the A3 catalytic motif, as well as variations in loop7, in both the sequence (identities of the nodes; Fig. 6A) and conformational (topology of the graph) senses. Such visualisation also rationalises the positioning of backbone and side-chains important in recognition of the substrate (Fig. 6B). Altogether, these applications demonstrate the usage of graph-based representation to reveal important structural features of the A3-DNA interface.

Modelling the impact of substrate sequences and topologies on A3 DNA-binding functions

So far the DNA-binding interface graphs presented above have been obtained from particular A3 CDAs with structures in complex with ssDNA resolved experimentally. We have further performed homology modelling to obtain structural models for all A3 CDA domains (see Methods, section Homology modelling). Using the procedure illustrated in Fig. 3B, we have generated ensembles of A3-DNA "grafts'' amenable to comparisons of possible A3-DNA interactions with different sequences and topologies on the substrate, which allows for examination of the suitability of these substrates to different A3 CDAs.

Nucleic acid entanglement in A3 CDAs

It became apparent, from manually inspecting these grafts and their optimised poses, that not all of these grafts are physicochemically realistic: in some generated structures the nucleic acid molecule is "entangled'' within the loops around the catalytic site of the CDA domain, in particular loop1 and loop3 (Fig. 7A). In general the following two extreme cases could be observed: either the nucleic acid chain entirely "penetrates'' the protein loop ("chain-entanglement", as illustrated in Fig. 7A), or an individual base on the nucleic acid chain is "wrapped''

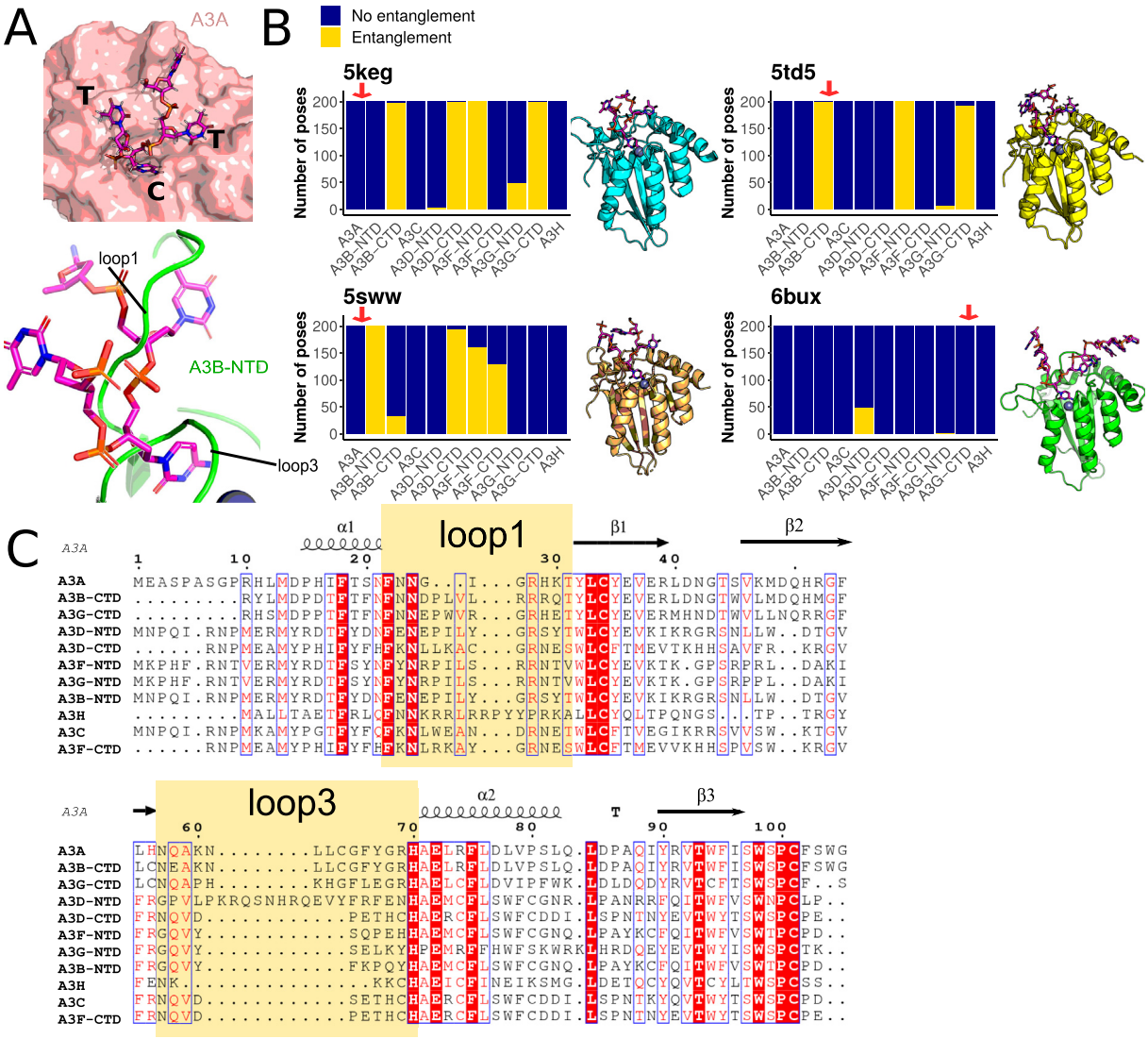

Figure 7. Nucleic acid entanglement in A3-DNA grafts. (A) Depiction of the positioning of DNA construct in A3 domain structures. Whereas certain domains could fit in the DNA sequence as reported (e.g. A3A with 5′-TCT-3′, top in pink), in other cases (e.g. N-terminal CDA of A3B [A3B-NTD]; bottom in green) the DNA molecule "protrudes'' into loop1 and/or loop3. (B) A survey of entanglement for all 4 DNA constructs considered here against all 11 human A3 CDA domains. Pink arrows indicated the identity of the protein chain as resolved in the original PDB entry. See methods for a detailed account of an automatic procedure to detect nucleic acid entanglement, and Figure S10 in Supplementary Material for counts specifically for loop1 and loop3. Here results from Ensemble *in situ* (i.e. starting from pose which resembles the original A3-DNA complex) are shown; see Supplementary Fig. S10 for results from Ensemble *apo* where the position of the bound DNA was not considered (see Methods). (C) Sequence alignment of the A3 CDA domains showing variations at loop1 and loop3 (highlighted). Alignment was generated using T-coffee (Notredame et al., 2000). The secondary structure shown corresponds to that of an A3A crystal structure (PDB 4xxo) (Bohn et al., 2015).

around by residues in the loop ("base-entanglement"). "Chain-entanglement'' is unrealistic, since the nucleic acid substrate would then be impossible to dissociate from the enzyme. A "base-entanglement'' refers to configurations that hold the base *within* a protein loop and do not allow for reaching out the catalytic pocket; this is incompatible with our current knowledge regarding the substrate binding mechanisms of A3 CDAs (Shi et al., 2017b; Kouno et al., 2017; Maiti et al., 2018), according to which the nucleic acid substrate is positioned via weak inter-atomic contacts with residues which define the catalytic pocket. Such entanglements are therefore likely artefacts generated through the *in silico* modelling process, indicating a limitation in homology modelling procedures in determining a proper conformations relevant to enzyme activity; they also imply that these domains might at least require large structural rearrangement (which was not achievable by the adopted Rosetta procedure) of the gating loops to accommodate such substrate, or that they might not bind DNA of such topologies at all.

To automate the process of detecting such entanglement phenomena, we used a geometrical approach to detect whether a part (either individual bases or the DNA chain) of the DNA molecule is entangled in planes defined by residues in loop1 and loop3. Briefly (for a detailed description see Methods), atomic coordinates are treated in a three-dimensional Cartesian system. The pipeline defines triangles using triples of residues, and finally examines the location of the DNA base/chain relative to the planes as defined by the triangles. These inform geometrical tests to verify whether the DNA base/chain is entangled in a loop. These tests considered every possible combination of vector and triangle definitions. They would therefore indicate whether and where protein-DNA entanglements occur in each generated pose.

Surveying across all combinations of A3 CDAs and nucleic acid sequences/- topologies examined here, entanglement of the DNA appears to be fairly prevalent (Fig. 7B), with a linear DNA topology (in PDB 6bux) appearing more receptive (less entanglement observed) across the CDAs

in general, in comparison to the U-shaped topologies adopted in the other PDB entries examined here. Two observations are notable: firstly, the results shown in Fig. 7B is based on Ensemble *in situ* of the grafts, computed using the configuration which best resembles the experimentally resolved A3-DNA complexes as the starting pose. Entanglement is expectedly more prevalent in Ensemble *apo* of the grafts (Figure S10, Supplementary Materials), where no consideration of the native A3-DNA interaction has been taken in estimating the ensemble (and hence more likely to encounter clashes between the protein and DNA chains). Secondly, entanglement is also evident for combinations of nucleic acid molecule and A3 CDA domain which has been experimentally resolved (labelled in Fig. 7B; albeit with a modified CDA domain sequence), suggesting that substantial rearrangement local to the protein-DNA interface may be necessary for a full-length, wild-type CDA domain. Entanglement occurs at both loop1 and loop3 of the CDA domain (Supplementary Fig. S10). The existence of entanglement in these grafts could be a possible explanation for the difference in activity of these domains: one could speculate that the variability in loop1 and loop3 (Fig. 7C) could be a protective mechanism against excessive deamination, by limiting their ability to approach substrates (see Discussion).

Probing the basis of substrate preference of A3 CDAs

While all A3 CDAs are homologous and share identical structural fold, variations exist with regards to their preferred DNA substrate sequence. The majority of catalytically active CDA domains prefer the sequence of 5′-TC-3’ (where C is the mutated base); A3G C-terminal CDA (A3G-CTD) is the sole exception, preferring 5′-CC-3′ instead (Yu et al., 2004; Langlois et al., 2005; Chen et al., 2006; Logue et al., 2014). The Rosetta toolset includes widely used protocols for structure-based protein and nucleic acid designs (Leaver-Fay et al., 2011; Alford et al., 2017). Here we ask whether these functionalities could be exploited to assess substrate preference of A3 CDAs, by designing the optimal substrate sequence for different A3 CDAs based on the graft poses we have generated (Fig. 3B).A large-scale mutagenesis of the DNA substrate sequence was carried using Rosetta (section DNA motif optimisation, Materials and Methods). For each A3 CDA, a large number of designs were generated (for 2 ensembles × 200 poses × 50 designs each, yielding a total of 20,000 for each CDA domain), allowing for sufficient sampling to explore DNA sequence preferences. Owing to existing knowledge on sequence preferences, here we focus on comparing A3A, A3B-CTD and A3G-CTD, the three domains

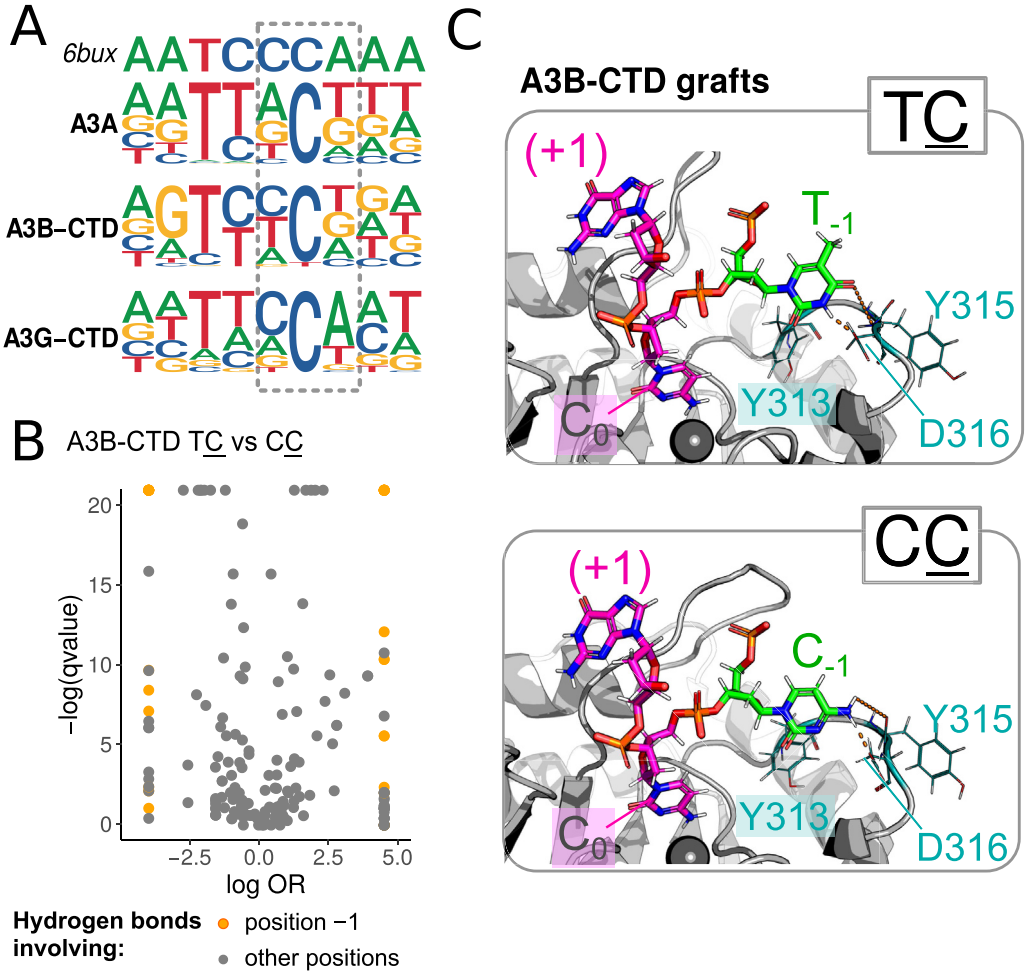

Figure 8. Optimising DNA motif bound to A3 CDA domains. (A) Substrate preference of A3A, A3B-CTD and A3G-CTD, simulated using the DNA topology in PDB 6bux as the starting sequence. The [-1, 0, 1] positions were highlighted (dashed lines). These sequence logos were generated using the results from Ensembles 1 and 2 (see Methods, section Graft generation) combined. (B) A comparison of hydrogen bonds found in poses with 5′-TC against those found in 5′-CC. Here a volcano plot was depicted, illustrating the enrichment (quantified using an odds ratio (OR) of 5′-TC over 5′-CC) and the significance level (Fisher's exact test; corrected using the Benjamini-Hochberg method). Hydrogen bonds enriched in 5′-TC would carry a positive *logOR*, whereas those in 5′-CC would be negative. Hydrogen bonds involving the DNA position −1 are highlighted in orange. (C) Visualisation of hydrogen bonds on A3B-CTD poses. Only hydrogen bonds involving the −1 position unique to either 5′-TC and 5′-CC are highlighted (in orange); those common to both and/or involving the sugar-phosphate backbone are not shown for simplicity. The −1 position is coloured green. Notably, the hydrogen bond interactions shown here are distinct to those reported in crystallographic studies where substrates of either sequence were used (see Supplementary Fig. S11).

with best established analyses on substrate sequence preference (Silvas & Schiffer, 2019; Shi et al., 2017b; Silvas et al., 2018), with the longer, more linear 5′-CCA-3′ A3G-CTD substrate (PDB 6bux; (Maiti et al., 2018)), on which the least entanglement has been observed in our grafts (Fig. 7). We find that some designs on the A3B-CTD grafts successfully alter 5′-CC-3′ into 5′-TC-3' (Fig. 8A); roughly 38% of the poses were designed into 5′-TC-3′, while a similar proportion of poses retained 5′-CC-3′ as does the original substrate (Table 4). However, for A3A, which is highly identical to A3B-CTD, base preference at the −1 position were different; this could possibly explained by difference in the loop1 sequence between A3A and A3B-CTD (Fig. 7C), both in terms of the length of loop1 and the fact that A3B-CTD has more positively-charged residues at this site. For A3G-CTD, it is expected that the motif preference in the generated grafts should be identical to the input substrate (i.e. 5′-CC-3′), since this input substrate was resolved in complex with A3G-CTD; around 55% of the poses retain 5′-CC-3'.

We sought to identify intermolecular interactions unique to A3B-CTD bound to 5′-TC, as such features would possibly contribute to recognise preferred substrates. As reported previously, preference for 5′-TC *vs* 5′-CC are dictated by hydrogen bonds and π-stacking with the side-chains of loop7 residues (Shi et al., 2017b; Kouno et al., 2017; Hou et al., 2019). We compare the occurrence of these intermolecular interactions in designs of 5′-TC *vs* 5′-CC (see Methods, section 2.4.4). While this analysis identifies interactions enriched in 5′-TC over 5′-CC (Fig. 8B and Table 5), most of these interactions involve either other bases in the ssDNA molecule, or the sugar-phosphate backbone at position − 1. A small number of such interactions involve atom groups distinct to either C_{-1} or T_{-1} (highlighted in Fig. 8C). These hydrogen bonds are different from those reported in crystallographic studies (Shi et al., 2017b); this is expected due to slight difference in the positioning of the −1 nucleotide in A3B-CTD and A3G-CTD complexed with ssDNA (Figure S11E, Supplementary Materials). Hydrogen bonds mediated by water molecules, which could be important in dictating substrate preference (Shi et al., 2017b), were omitted here due to the omission of water molecules in our models (Supplementary Fig. S11). Therefore, while some attempts in this motif optimisation experiment has reproduced the sequence preference of A3B-CTD as reported in structural studies, the procedure gives a mechanistic explanation behind such sequence preference which is alternative to earlier reports. With regards to the negative findings on A3A (Fig. 8B), these probably imply that more sophisticated scoring methods and descriptors of protein-ssDNA interactions are required to model A3-DNA interactions more properly (see Discussion).

Taken together, residue subgraphs demonstrate to be an effective way to visually highlight and compare conformational differences; the isomorphism between subgraphs and conventional descriptions of protein structural elements also enables studies of this kind of problem at a larger scale, where one can move from inspecting individual protein structures to a large–scale graph extraction, thereby enabling a comprehensive survey of structural features over a sizeable family of homologous proteins, as in this case of A3 CDAs.

Table 4
Comparison of base preference at position −1 for motif-optimised A3-DNA poses. Numbers shown represent the percentages of poses with each DNA base at position −1. Poses of A3 CDAs in complex with Rosetta-designed sequence motives are considered here. Results from both Ensemble *apo* and Ensemble *in situ* (see Methods, section Graft generation) are combined and tabulated here.

Domain	% poses with base at position −1			
	A	C	G	T
A3A	46.87	6.38	27.91	18.85
A3B-CTD	20.31	38.38	3.42	37.89
A3G-CTD	23.81	55.15	14.63	6.42

Table 5
Comparison of π-π stacking occurrences in A3B-CTD 5′-TC and 5′-CC grafts. π-π stacking between the protein and DNA chain was detected using a geometrical method detailed in the Method section. Here we report the odds ratio (OR) comparing the occurrences of such interactions in A3B-CTD DNA grafts with either a C or a T at the −1 position. A positive OR indicates the interaction is enriched in the 5′-TC grafts. Results are based on graft Ensembles 1 and 2 combined (see Methods, section Graft generation). Statistical significance was assessed using a Fisher's exact test, the *p*-values from which were corrected for multiple comparisons using the Benjamini-Hochberg method.

protein residue	DNA base	odds ratio (OR)	*q*-value
HIS253	C_0	0.129	0
TYR250	C_0	0.848	2.52×10^{-2}
TYR250	C_{-1}	0.000	1.29×10^{-48}
TYR250	T_{-1}	∞	1.49×10^{-53}
TYR313	C_{-1}	0.000	2.44×10^{-7}
TYR313	C_{-2}	9.636	1.47×10^{-5}
TYR315	C_{-1}	0.000	2.71×10^{-11}
TYR315	T_{-1}	∞	1.69×10^{-10}

Discussion

This study has demonstrated the usefulness of the use of a graph-theoretical approach to describe protein interfaces, here illustrated with the DNA-binding interface in A3 CDAs. The approach revealed suitable to effectively screen, identify and compare interfaces from a large number of structural complexes. Each residue graph is essentially a representation of the input structure; in this sense graph-based representations are insensitive to the number of available crystal structures for a given domain of interest. Notably, in combination with *in silico* modelling of structural ensembles, the use of graph-based representations enables comparisons of the relative importance of the residues in protein interfaces (Fig. 5). Per-residue evaluation of protein-DNA interfaces have previously yielded scoring systems useful for the prediction of DNA-binding propensities given a protein surface (Corsi et al., 2020); here our subgraphs represent the spatial arrangement amongst residues, which will add an extra layer of description for such interfaces. The use of graph theory has been recently an area of active research in structural bioinformatics, for example in cataloguing patterns of molecular assemblies (Heal et al., 2018). Here geometry and graph theory are applied to the problem of analysing protein interfaces, with successes to highlight, both visually and numerically, similarities and differences across A3 CDAs in the DNA-binding interface. Our interface graphs embed residue information and allow for efficient comparisons; other important elements of the interface, e.g. water molecules, are not included, which deserve attention and careful interpretation of these graphs in generating biological insights. Nevertheless, these methods are capable to push the analysis of protein interfaces to a proteome-wide scale, where a large number of structural poses could be screened *in silico*, here illustrated with analysing thousands of grafted poses of A3 CDAs with different ssDNA topologies. While structural modelling engines like MODELLER and Rosetta have been the focus in the structural biology community to model protein structures, analysis of the large amount of data generated deserves more detailed analyses of specific geometric features of interaction interfaces or binding pockets; methods such as graph-based representations could fill this gap.

In the case of the A3 CDA domains, the subgraph extraction method coupled with an *in silico* modelling pipeline described here have generated new insights with regards to substrate recognition. Subgraphs specific to the DNA-binding interface highlight, both qualitatively and quantitatively, similarities and differences across the human A3 family. Using homology modelling and grafting of various DNA topologies, it has become apparent that the A3 CDA domains exhibit selectivity towards specific DNA sequences and topologies, in that for some combinations entanglement of the DNA chain persisted (Fig. 7). Importantly, this existed even in the case where the starting conformation was selected to

accommodate the DNA chain *in situ*. Homology modelling is inherently blind in assessing whether conformations are relevant for biological activity. This points to an area of improvement in the modelling pipeline and calls for an effective rejection protocol to be in place, so that such solutions could have been filtered away from the modelling procedure. On the other hand, the long flexible loops around the catalytic site (Fig. 7C) could also be a naturally evolved feature of the CDA domains, to limit excessive DNA binding, rendering these positions susceptible to persisting entanglements in our ssDNA grafting procedure. Moreover, the fact that entanglement exists could imply that there is a limit, beyond which extrapolation of the DNA-binding mode of one specific A3 CDA domain to other family members could be unrealistic.

We have also illustrated the importance of considering modifications introduced in the crystal structures, particularly in interpreting the atomistic details of DNA binding in the case of the A3 CDA domains. Many existing A3-DNA co–crystal structures are resolved on a modified A3 CDA domain, either through point mutation or grafting of entire protein loop(s) from another family member (Shi et al., 2017b; Kouno et al., 2017; Maiti et al., 2018). Here we have "repaired'' these hybrid constructs using homology modelling, and subsequently sampled ensembles of structures (Fig. 3B). These results demonstrated that the binding modes revealed by these crystallographic studies to the wild-type CDA domain might need to be revisited when considering wild type CDA structures, since entanglement could also, surprisingly, be observed for these domains on which complexes with single-stranded DNA were obtained and reported (Fig. 7B). On the topology of substrate, a more linear substrate appears to be accommodated more universally across A3 CDAs in comparison to U-shaped ones, although this may also be related to redundancy (continuous string of cytosine) in the DNA sequence (Fig. 1B). Modelling of full-length, wild-type CDAs helps in rationalising previously reported findings, for example the comparison of accessibility of the catalytic pocket: while crystallographic studies pointed to the importance of loop1 and loop7 in controlling substrate access (Shi et al., 2015), loop3 is typically removed or mutated in these experimental studies. Here we demonstrate that loop3 also poses a significant steric barrier for ssDNA in our substrate grafting experiments (Fig. 7). The behaviour of A3 CDA domains *in vitro* render sequence alterations necessary in order to obtain crystal structures (Shi et al., 2017b; Kouno et al., 2017; Maiti et al., 2018), but here our results suggest that caution has to be exercised in interpreting the structural details as revealed by these resolved structures. In these cases, *in silico* approaches are important in sampling possible conformations of full-length, wild-type domains which otherwise could not be obtained with only experimental approaches.

Furthermore, on the subject of substrate preference, we attempted a motif optimisation experiment (Fig. 8). This successfully engineered different preferred motives for A3B-CTD and A3G-CTD, suggesting that the structural arrangements around the catalytic site for these two domains optimise its specificity against substrates harbouring the 5′-TC-3′ or the 5′-CC-3′ motif respectively, as described in *vitro* experiments and clinical tumour genomes. Since the −1 base is positioned differently in A3B-CTD and A3G-CTD, poses which successfully incorporate the A3B-CTD preferred motif engage in a different set of intermolecular interactions (Fig. 8D) as compared to the A3B-CTD-ssDNA co-complex crystal structure (Fig. 6B). Existing scoring functions for modelling protein-DNA interactions appear to be optimised with protein structures in complex with double-stranded DNA (Ashworth & Baker, 2009). While this is the best option available for this analysis, scoring functions which are better optimised with single-stranded DNA, and/or DNA sequence-specific scoring methods (Paillard & Lavery, 2004) might help in evaluating the grafting poses. Moreover, due to limitations in the modelling procedures, water molecules were omitted, which has been known as important contributor in intermolecular interactions in the A3-DNA setting (Shi et al., 2017b; Kouno et al., 2017; Maiti et al., 2018). This analysis might also be improved by using more sophisticated methods to model and assess protein-DNA interactions, by, e.g. using more comprehensive force-fields and atomistic molecular dynamics (MD) simulations. Such approaches, applied on the grafts generated here, could be useful in assessing the relevance of binding modes revealed in these structural studies of A3 with single-stranded DNA of different sequences and topologies.

An important consideration in understanding the catalytic mechanism of A3 proteins, especially for those containing two CDA domains, is on the assembly of the two domains with respect to one another. This has been a long-standing question in the field, but recently a full-length crystal structure of rhesus A3G has been reported, albeit in the absence of ssDNA (Yang et al., 2020). Possibly, in the future a complex structure of a full-length double-domain A3 in complex with ssDNA will be resolved, and this will shed light onto the structural basis behind the differential level of catalytic activity between the NTD and the CTD, and how the cooperativity of these two domains could contribute to modulate the level of mutagenesis catalysed by A3 proteins.

Conclusions

This work demonstrates that graph-based representations are effective in comparing important structural features of the A3 CDA domains, coupled with large-scale modelling of structural ensembles to fill the gap in unresolved A3 CDA domain structures. Using molecular modelling we have analysed the structural underpinning of the substrate preference of A3B-CTD, which reveal alternative intermolecular interactions to coordinate substrate binding (Fig. 8). Taking advantage of the rich information that residue subgraphs embed, these could be harnessed as a cataloguing tool to represent the structural organisation at various protein interfaces. This requires subgraph extraction over many structures which contain a particular type of interface, and a robust statistical approach to select a small set of subgraphs which are representative enough to reconstruct the entirety of the interface of interest. In this way, these graph-based representations will have the potential to be developed into a system of descriptors for protein interfaces.

Author statement

Joseph Chi-Fung Ng: Conceptualization; Data curation; Formal analysis; Writing, reviewing and editing manuscript. **Franca Fraternali**: Conceptualization; Funding acquisition; Supervision; Reviewing and editing manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank the Fraternali laboratory for fruitful discussions. This work was by the Croucher Foundation Hong Kong (scholarship to J.C.F.N.), the Medical Research Council (MR/L01257X/1 to F.F.) and the Biotechnology and Biological Sciences Research Council (BB/T002212/1 to F.F. and J.C.F.N.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crstbi.2020.07.001.

References

Sheehy, A.M., Gaddis, N.C., Choi, J.D., Malim, M.H., 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral vif protein. Nature 418 (6898), 646–650. https://doi.org/10.1038/nature00939.

Malim, M.H., 2009. APOBEC proteins and intrinsic resistance to HIV-1 infection. Phil Trans Roy Soc Lond B Biol Sci 364 (1517), 675–687. https://doi.org/10.1098/rstb.2008.0185.
Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., et al., 2013. Signatures of mutational processes in human cancer. Nature 500 (7463), 415–421. https://doi.org/10.1038/nature12477.
Ng, J.C.F., Quist, J., Grigoriadis, A., Malim, M.H., Fraternali, F., 2019. Pan-cancer transcriptomic analysis dissects immune and proliferative functions of APOBEC3 cytidine deaminases. Nucleic Acids Res 47 (3), 1178–1194. https://doi.org/10.1093/nar/gky1316.
Smith, H.C., Bennett, R.P., Kizilyer, A., McDougall, W.M., Prohaska, K.M., 2012. Functions and regulation of the APOBEC family of proteins. Semin Cell Dev Biol 23 (3), 258–268. https://doi.org/10.1016/j.semcdb.2011.10.004.
Salter, J.D., Smith, H.C., 2018. Modeling the embrace of a mutator: APOBEC selection of nucleic acid ligands. Trends Biochem Sci 43 (8), 606–622. https://doi.org/10.1016/j.tibs.2018.04.013.
Silvas, T.V., Schiffer, C.A., 2019. APOBEC3s: DNA-editing human cytidine deaminases. Protein Sci 28 (9), 1552–1566. https://doi.org/10.1002/pro.3670.
Hou, S., Silvas, T.V., Leidner, F., Nalivaika, E.A., Matsuo, H., Yilmaz, N.K., et al., 2019. Structural analysis of the active site and DNA binding of human cytidine deaminase APOBEC3B. J Chem Theor Comput 15 (1), 637–647. https://doi.org/10.1021/acs.jctc.8b00545.
Salter, J.D., Bennett, R.P., Smith, H.C., 2016. The APOBEC protein family: united by structure, divergent in function. Trends Biochem Sci 41 (7), 578–594. https://doi.org/10.1016/j.tibs.2016.05.001.
Shi, K., Carpenter, M.A., Kurahashi, K., Harris, R.S., Aihara, H., 2015. Crystal structure of the DNA deaminase APOBEC3B catalytic domain. J Biol Chem 290 (47), 28120–28130. https://doi.org/10.1074/jbc.M115.679951.
Shi, K., Demir, Ö., Carpenter, M.A., Wagner, J., Kurahashi, K., Harris, R.S., et al., 2017a. Conformational switch regulates the DNA cytosine deaminase activity of human APOBEC3B. Sci Rep 7 (1), 17415. https://doi.org/10.1038/s41598-017-17694-3.
Shi, K., Carpenter, M.A., Banerjee, S., Shaban, N.M., Kurahashi, K., Salamango, D.J., et al., 2017b. Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. Nat Struct Mol Biol 24 (2), 131–139. https://doi.org/10.1038/nsmb.3344.
Kouno, T., Silvas, T.V., Hilbert, B.J., Shandilya, S.M.D., Bohn, M.F., Kelch, B.A., et al., 2017. Crystal structure of APOBEC3A bound to single-stranded DNA reveals structural basis for cytidine deamination and specificity. Nat Commun 8, 15024. https://doi.org/10.1038/ncomms15024.
Maiti, A., Myint, W., Kanai, T., Delviks-Frankenberry, K., Rodriguez, C.S., Pathak, V.K., et al., 2018. Crystal structure of the catalytic domain of HIV-1 restriction factor APOBEC3G in complex with ssDNA. Nat Commun 9 (1), 2460. https://doi.org/10.1038/s41467-018-04872-8.
Fang, Y., Xiao, X., Li, S.-X., Wolfe, A., Chen, X.S., 2018. Molecular interactions of a DNA modifying enzyme APOBEC3F catalytic domain with a single-stranded DNA. J Mol Biol 430 (1), 87–101. https://doi.org/10.1016/j.jmb.2017.11.007.
Shaban, N.M., Shi, K., Lauer, K.V., Carpenter, M.A., Richards, C.M., Salamango, D., et al., 2018. The antiviral and cancer genomic DNA deaminase APOBEC3H is regulated by an RNA-mediated dimerization mechanism. Mol Cell 69 (1), 75–86. https://doi.org/10.1016/j.molcel.2017.12.010 e9.
Losey, H.C., Ruthenburg, A.J., Verdine, G.L., 2006. Crystal structure of Staphylococcus aureus tRNA adenosine deaminase TadA in complex with RNA. Nat Struct Mol Biol 13 (2), 153–159.
Varadi, M., Berrisford, J., Deshpande, M., Nair, S.S., Gutmanas, A., Armstrong, D., et al., 2020. PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res 48 (D1), D344–D353.
Heal, J.W., Bartlett, G.J., Wood, C.W., Thomson, A.R., Woolfson, D.N., 2018. Applying graph theory to protein structures: an atlas of coiled coils. Bioinformatics 34 (19), 3316–3323. https://doi.org/10.1093/bioinformatics/bty347.
Bahar, I., Atilgan, A.R., Erman, B., 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding Des 2 (3), 173–181. https://doi.org/10.1016/S1359-0278(97)00024-2.
Bahar, I., Lezon, T.R., Yang, L.-W., Eyal, E., 2010. Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys 39, 23–42. https://doi.org/10.1146/annurev.biophys.093008.131258.
Papaleo, E., 2015. Integrating atomistic molecular dynamics simulations, experiments, and network analysis to study protein dynamics: strength in unity. Frontiers in molecular biosciences 2, 28. https://doi.org/10.3389/fmolb.2015.00028.
Doncheva, N.T., Assenov, Y., Domingues, F.S., Albrecht, M., 2012. Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc 7 (4), 670–685. https://doi.org/10.1038/nprot.2012.004.
Chakrabarty, B., Parekh, N., 2016. Naps: network analysis of protein structures. Nucleic Acids Res 44 (W1), W375–W382. https://doi.org/10.1093/nar/gkw383.
Kleinjung, J., Fraternali, F., 2005. POPSCOMP: an automated interaction analysis of biomolecular complexes. Nucleic Acids Res 33, W342–W346. https://doi.org/10.1093/nar/gki369. Web Server issue.
Cavallo, L., Kleinjung, J., Fraternali, F., 2003. POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. Nucleic Acids Res 31 (13), 3364–3366.
Fornili, A., Pandini, A., Lu, H.-C., Fraternali, F., 2013. Specialized dynamical properties of promiscuous residues revealed by simulated conformational ensembles. J Chem Theor Comput 9 (11), 5127–5147. https://doi.org/10.1021/ct400486p.
Notredame, C., Higgins, D.G., Heringa, J., T-coffee, 2000. A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302 (1), 205–217. https://doi.org/10.1006/jmbi.2000.4042 arXiv:10964570.
Webb, B., Sali, A., 2017. Protein structure modeling with MODELLER. Methods Mol Biol 1654, 39–54. https://doi.org/10.1007/978-1-4939-7231-9_4. Clifton, N.J.
Shen, M.Y., Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. Protein Sci 15 (11), 2507–2524.
Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (12), 2577–2637.
Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., et al., 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487, 545–574.
Ashworth, J., Havranek, J.J., Duarte, C.M., Sussman, D., Monnat, R.J., Stoddard, B.L., et al., 2006. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 441 (7093), 656–659.
Ashworth, J., Taylor, G.K., Havranek, J.J., Quadri, S.A., Stoddard, B.L., Baker, D., 2010. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. Nucleic Acids Res 38 (16), 5601–5608.
Bakan, A., Meireles, L.M., Bahar, I., 2011. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics 27 (11), 1575–1577.
Salamanca Viloria, J., Allega, M.F., Lambrughi, M., Papaleo, E., 2017. An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass. Sci Rep 7 (1), 2838.
Schrödinger, L.L.C., November 2017. The PyMOL molecular graphics system, version 2.0.
Freeman, L.C., 1978. Centrality in social networks: conceptual clarification. Soc Network 1 (3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7.
Freeman, L.C., 1980. The gatekeeper, pair-dependency and structural centrality. Qual Quantity 14 (4), 585–592. https://doi.org/10.1007/BF00184720. URL.
Borgatti, S.P., Everett, M.G., 2006. A graph-theoretic perspective on centrality. Soc Network 28 (4), 466–484. https://doi.org/10.1016/j.socnet.2005.11.005.
Negre, C.F.A., Morzan, U.N., Hendrickson, H.P., Pal, R., Lisi, G.P., Loria, J.P., et al., 2018. Eigenvector centrality for characterization of protein allosteric pathways. Proc Natl Acad Sci USA 115 (52), E12201–E12208.
Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., et al., 2004. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem 25 (13), 1605–1612.
Ferreira de Freitas, R., Schapira, M., 2017. A systematic analysis of atomic protein–ligand interactions in the pdb. Med. Chem. Commun. 8, 1970–1981. https://doi.org/10.1039/C7MD00381A. URL.
Mills, J.E., Dean, P.M., 1996. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. J Comput Aided Mol Des 10 (6), 607–622.
Li, M., Goncearenco, A., Panchenko, A.R., 2017. Annotating mutational effects on proteins and protein interactions: designing novel and revisiting existing protocols. Methods Mol Biol 1550, 235–260.
Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Costanzo, L.D., et al., 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res 47 (D1), D520–D528.
Wickham, H., 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag New York. URL. https://ggplot2.tidyverse.org.
Wagih, O., 2017. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics 33 (22), 3645–3647.
Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., et al., 2019. Gplots: various R programming tools for plotting data, r package version 3.0.1.1. URL. https://CRAN.R-project.org/package=gplots.
Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., Karplus, M., 1995. Evaluation of comparative protein modeling by MODELLER. Proteins 23 (3), 318–326. https://doi.org/10.1002/prot.340230306.
Bohn, M.F., Shandilya, S.M.D., Silvas, T.V., Nalivaika, E.A., Kouno, T., Kelch, B.A., et al., 2015. The ssDNA mutator APOBEC3A is regulated by cooperative dimerization. Structure 23 (5), 903–911.
Yu, Q., Chen, D., König, R., Mariani, R., Unutmaz, D., Landau, N.R., 2004. APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. J Biol Chem 279 (51), 53379–53386. https://doi.org/10.1074/jbc.M408802200.
Langlois, M.-A., Beale, R.C.L., Conticello, S.G., Neuberger, M.S., 2005. Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/g anti-retroviral cytidine deaminases provides insight into their DNA target site specificities. Nucleic Acids Res 33 (6), 1913–1923. https://doi.org/10.1093/nar/gki343.
Logue, E.C., Bloch, N., Dhuey, E., Zhang, R., Cao, P., Herate, C., et al., 2014. A DNA sequence recognition loop on APOBEC3A controls substrate specificity. PloS One 9 (5), e97062. https://doi.org/10.1371/journal.pone.0097062.
Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., et al., 2017. The rosetta all-atom energy function for macromolecular modeling and design. J Chem Theor Comput 13 (6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.
Silvas, T.V., Hou, S., Myint, W., Nalivaika, E., Somasundaran, M., Kelch, B.A., et al., 2018. Substrate sequence selectivity of APOBEC3A implicates intra-DNA interactions. Sci Rep 8 (1), 7511.
Chen, H., Lilley, C.E., Yu, Q., Lee, D.V., Chou, J., Narvaiza, I., et al., 2006. APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. Curr Biol 16 (5), 480–485. https://doi.org/10.1016/j.cub.2006.01.031.

Corsi, F., Lavery, R., Laine, E., Carbone, A., 2020. Multiple protein-DNA interfaces unravelled by evolutionary information, physico-chemical and geometrical properties. PLoS Comput Biol 16 (2), e1007624.
Ashworth, J., Baker, D., 2009. Assessment of the optimization of affinity and specificity at protein-DNA interfaces. Nucleic Acids Res 37 (10), e73.
Paillard, G., Lavery, R., 2004. Analyzing protein-DNA recognition mechanisms. Structure 12 (1), 113–122.
Yang, H., Ito, F., Wolfe, A.D., Li, S., Mohammadzadeh, N., Love, R.P., et al., 2020. Understanding the structural basis of HIV-1 restriction by the full length double-domain APOBEC3G. Nat Commun 11 (1), 632.