3394–3408 Nucleic Acids Research, 2021, Vol. 49, No. 6 Published online 28 February 2021
doi: 10.1093/nar/gkab113

Defining a novel domain that provides an essential contribution to site-specific interaction of Rep protein with DNA

Katarzyna Wegrzyn[1,†], Elzbieta Zabrocka[1,†], Katarzyna Bury[1], Bartlomiej Tomiczek[1], Milosz Wieczor[2], Jacek Czub[2], Urszula Uciechowska[1], María Moreno-del Alamo[3], Urszula Walkow[1], Igor Grochowina[1], Rafal Dutkiewicz[1], Janusz M. Bujnicki[4,5], Rafael Giraldo[3] and Igor Konieczny[1,*]

[1]Intercollegiate Faculty of Biotechnology of University of Gdansk and Medical University of Gdansk, University of Gdansk, Abrahama 58, 80-307 Gdansk, Poland, [2]Department of Physical Chemistry, Gdańsk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland, [3]Department of Cellular and Molecular Biology, Centro de Investigaciones Biológicas – CSIC, E28040 Madrid, Spain, [4]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Księcia Trojdena 4, 02-109 Warsaw, Poland and [5]Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Umultowska 89, 61–614 Poznan, Poland

Received December 21, 2020; Revised February 04, 2021; Editorial Decision February 07, 2021; Accepted February 10, 2021

ABSTRACT

An essential feature of replication initiation proteins is their ability to bind to DNA. In this work, we describe a new domain that contributes to a replication initiator sequence-specific interaction with DNA. Applying biochemical assays and structure prediction methods coupled with DNA–protein crosslinking, mass spectrometry, and construction and analysis of mutant proteins, we identified that the replication initiator of the broad host range plasmid RK2, in addition to two winged helix domains, contains a third DNA-binding domain. The phylogenetic analysis revealed that the composition of this unique domain is typical within the described TrfA-like protein family. Both *in vitro* and *in vivo* experiments involving the constructed TrfA mutant proteins showed that the newly identified domain is essential for the formation of the protein complex with DNA, contributes to the avidity for interaction with DNA, and the replication activity of the initiator. The analysis of mutant proteins, each containing a single substitution, showed that each of the three domains composing TrfA is essential for the formation of the protein complex with DNA. Furthermore, the new domain, along with the winged helix domains, contributes to the sequence specificity of replication initiator interaction within the plasmid replication origin.

INTRODUCTION

The binding of proteins to nucleic acids is often accomplished via specific motifs. Many of these motifs (e.g. helix-turn-helix (HTH), zinc finger, leucine zipper) are present within protein domains that are directly engaged in interaction with nucleic acids (1). In proteins that initiate DNA replication, usually the HTH motif and its variant, the winged HTH (WH) motif, are present within the domain responsible for DNA binding.

In bacteria, chromosomal DNA replication is initiated by the DnaA protein, which is composed of four domains (2). Domain IV, acting as the DNA-binding domain (DBD), possesses an HTH motif and is responsible for interactions with specific sequences (DnaA boxes) within the origin of replication of the bacterial chromosome (3,4). A crystal structure of the nucleoprotein complex assembled by the DBD of *Escherichia coli* DnaA showed that the HTH motif interacts primarily with the major groove of double-stranded DNA (dsDNA) and that additional contacts are made in the minor groove (3). Binding of DnaA to the DnaA-boxes within the dsDNA results in melting of the duplex in an AT-rich region, named the DNA unwinding ele-

[*]To whom correspondence should be addressed. Tel: +48 58 5236365; Email: igor.konieczny@ug.edu.pl
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Present addresses:
María Moreno-del Alamo, Department of Microbial Biotechnology, National Centre of Biotechnology – CSIC, E28049 Cantoblanco (Madrid), Spain.
Rafael Giraldo, Department of Microbial Biotechnology, National Centre of Biotechnology – CSIC, E28049 Cantoblanco (Madrid), Spain.

ment (DUE), and formation of a single-stranded region (ssDNA), where the replication machinery is assembled. The DnaA boxes are separated from the AT-rich region by a GC-rich sequence and DnaA-trios, sequences required for the stable formation of aDnaA-filament (5). In the ssDNA DUE, additional sequence-specific interaction of the DnaA protein takes place, mediated by amino acid residues of domain III (AAA+; an ATPase associated with various cellular activities) (6,7). Some of the amino acid residues of the AAA+ domain that engage in interactions with ssDNA are part of an initiator-specific motif (ISM), which is characteristic for proteins from the AAA+ family that initiate DNA replication (6). Such an ISM, although different from that found in DnaA, can also be identified in the AAA+ domains of the archaeal and eukaryotic replication initiators, called the ORC (origin recognition complex) proteins (8). However, in contrast to the bacterial DnaA protein, this motif interacts with dsDNA in ORC proteins. Crystallographic data have revealed that the ISM motif of ORC proteins (Orc1–1, Orc1–3) from the archaeon *Sulfolobus solfataricus* interacts with one-fourth of the dsDNA-binding surface at the origin of replication (9). Superposition of the *S. solfataricus* Orc1 protein crystal structure in a complex with DNA (9) onto the crystal structure of the *Drosophila* Orc4 protein suggested that the ISM is also engaged in the formation of a nucleoprotein complex at the origin of replication in eukaryotic replication initiators (10). Similar conclusions were drawn from *in silico* analysis of the crystal structure of human ORC (11). The cryo-EM structure of the ORC complex from *Saccharomyces cerevisiae* showed that ISM, as well as elements of the WH domain, indeed interact with DNA, but the participation of the DNA-binding motifs of the individual ORC proteins (Orc1–Orc5) in the nucleoprotein complex formation differs from their archaeal counterparts (12). This structural analysis also revealed the presence of another motif responsible for DNA binding—the basic patch (BP)—observed in Orc1 (12). The interaction of a patch of basic amino acids with DNA was also predicted earlier for fungal ORC proteins (13), and comparative sequence analyses revealed the conservation of BP in Orc1 proteins from yeast to human (12). Point mutations introduced into this region in the Orc1 protein from *Saccharomyces cerevisiae* indeed identify two residues, K362 and R367, as important for the nucleoprotein complex formation by the Orc1 protein (13).

In addition to the ISM motif, ORC initiators contain a domain with a WH motif, which is the main dsDNA-binding domain. The HTH motif interacts with the major groove, and the β-hairpin wings engage in minor groove binding (9). WH domains are also characteristic of plasmid replication initiators, the Rep proteins (14–18). These plasmid replication initiators are usually composed of two WH domains (Figure 1). One of these domains, the C-terminal WH2 domain, is mainly responsible for binding to specific repeated sequences (iterons) within the dsDNA origin of replication, whereas the second N-terminal domain, the WH1 domain, plays a secondary role in DNA binding by Rep monomers. The WH1 domain provides a dimerization interface when Rep dimers bind to operator sequences to repress *rep* transcription (15,16). The few available crystal structures of monomers of Rep proteins in complex with an iteron DNA sequence revealed that the amino acid residues of helix α4 in WH1 and helix α4′ in WH2 are in contact with dsDNA (16,18). Analysis of point mutants of π, the Rep protein of the plasmid R6K, showed that changes in the amino acid residues located within the α4 and α4′ helices indeed decrease the ability of the protein to interact with dsDNA (18).

Point mutations that change the DNA-binding ability have also been described for the TrfA protein, the replication initiator of the broad host range plasmid RK2 (19,20). This Rep protein is expressed in bacterial cells in two forms: a full-length protein with a molecular mass of 44 kDa (TrfA-44) and a shorter one with a molecular mass of 33 kDa (TrfA-33). The short protein, due to an internal translational starting point, begins at residue 98. Depending on the host bacterium, both of these forms can initiate replication of the RK2 plasmid (21,22). Although TrfA protein can sequence-specifically interact with single iteron sequence (23), it triggers plasmid DNA replication only by binding to the five iterons located within the plasmid replication origin (*oriV*) and subsequent formation of a nucleoprotein complex at one specific strand of ssDNA exposed in the DUE region of *oriV* (24). According to a structural prediction, the C-terminal region of TrfA-33 (190–382 aa) contains two WH domains (17). In the available *trfA* mutant variants (25), mutations resulting in proteins with either decreased (TrfA P314L, TrfA P314S) or increased (TrfA E361K) DNA binding were identified (19,20). These substitutions affecting the protein interaction with DNA are located in the WH2 domain (Figure 1). However, several mutations affecting the DNA binding are also located in the region of unknown structure preceding the WH1 domain (Figure 1) (e.g. substitutions P151S, R169H and A171T (19,20)). This region (residues 98–190) is long when compared to other Rep proteins (Figure 1). In this work, we investigated the possible role and structure of this additional N-terminal region of TrfA in the formation of a nucleoprotein complex with dsDNA.

MATERIALS AND METHODS

Bacterial strains, plasmids and nucleotides

Bacterial strains and plasmids used in this work are listed in Supplementary Table S1. *Escherichia coli* cells were transformed with plasmids using standard procedures. Oligonucleotides used in this work are listed in Supplementary Table S2. Biotinylated dsDNA fragments, used in the SPR analysis, were obtained by hybridization of two complementary single-stranded oligonucleotides (Metabion) ItT and ItB, pUCT and pUCB, as well as 1itT and 1itB. Plasmids (pSUMO-DBD and pSUMO-WH1WH2) were constructed by cloning the *trfA* gene fragments encoding the DBD and WH1WH2 domains into a pET SUMO vector. The cloning was performed using the Champion™ pET SUMO Expression System Kit (Invitrogen) according to the manufacturer's manual. The *dbd* and *wh1wh2* gene fragments were prepared by PCR amplification using pAT30 plasmid as a template. Oligonucleotides used for amplification of *dbd* (DBD-top and DBD-bot) and *wh1wh2* (WH1-WH2-top and WH1-WH2-bot) DNA fragments are listed

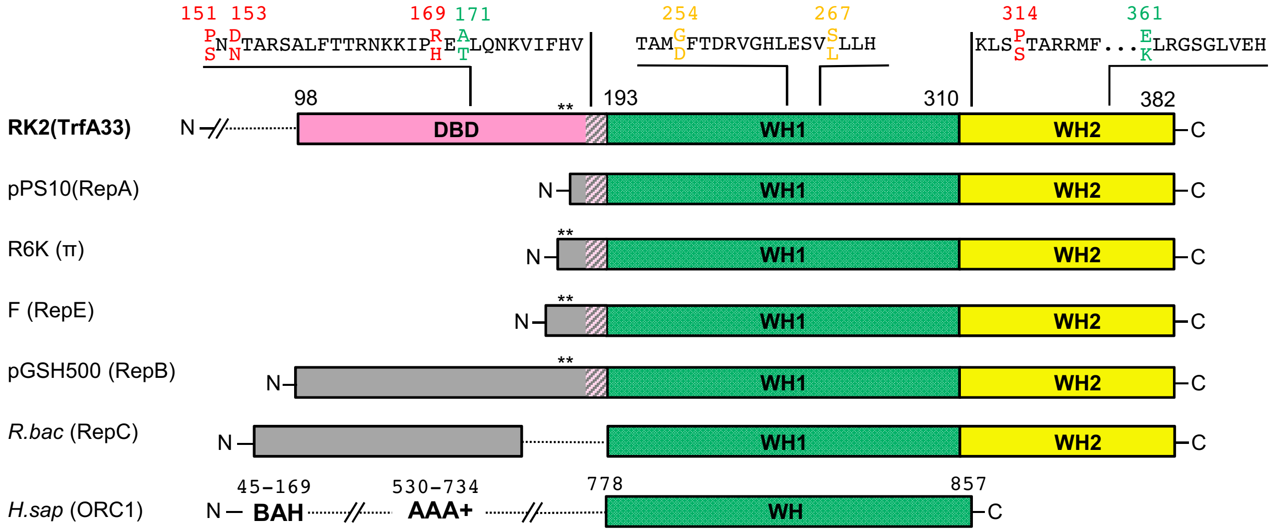

Figure 1. Comparison of domain organization of WH proteins. Coloured blocks indicate sequence conservation, based on multiple sequence alignment (MSA). Winged helix (WH) domains are indicated, WH1 in green and WH2 in yellow, putative DBD in pink. Conserved hydrophobic block is marked with stripes (in TrfA33 176–192 aa). Presence of positively charged basic patch is indicated with a star (*). Point mutations changing the TrfA33 DNA-binding ability are indicated, inhibitory in red, enhancing in green, inhibiting dimerization in orange.

in Supplementary Table S2. After amplification of the specific PCR products (153 bp *dbd* and 579 bp *wh1wh2* PCR products), the ligation reactions were performed according to the manufacturer's manual.

The constructions of pET SUMO-DBD/P151S, pAT30*trfA* G254D/S267L/P151S/E361K, pAT30*trfA* G254D/S267L/A171T/P314S and pAT30*trfA* G254D/S267L/R103E, pAT30*trfA* G254D/S267L/T105E, pAT30*trfA* G254D/S267L/K165E, pAT30*trfA* G254D/S267L/P168E, pAT30*trfA* G254D/S267L/R169E, pAT30*trfA* G254D/S267L/N234E, pAT30*trfA* G254D/S267L/R236E, pAT30*trfA* G254D/S267L/K280E and pAT30*trfA* G254D/S267L/R347E plasmids were performed through site-directed mutagenesis based on the following templates: pET SUMO-DBD (using P151S-F and P151S-R oligonucleotides), pAT30*trfA* G254D/S267L/P151S (using E361K-F and E361K-R oligonucleotides), pAT30*trfA* G254D/S267L/A171T (using P314S-F and P314S-R oligonucleotides) and pAT30*trfA* G254D/S267L (using R103E-F and R103E-R, T105E-F and T105E-R, K165E-F and K165E-R, P168E-F and P168E-R, R169E-F and R169E-R, N234E-F and N234E-R, R236E-F and R236E-R, K280E-F and K280E-R, R347E-F and R347E-R oligonucleotides), respectively (Supplementary Table S2).

Protein purification

All TrfA protein variants were overproduced in *E. coli* ArcticExpress (DE3) strain. In the experiments presented in this study, highly purified proteins (95% purity or higher) were utilized. All TrfA variants used in the tests were N-terminally histidine-tagged 33 kDa versions of TrfA with G254D/S267L mutation, resulting in constitutively active monomer of the protein. The use of G254D/S267L mutated TrfA protein variants eliminated the need of protein activation prior to each experiment. In the whole text, the TrfA wt refers to TrfA with only G254D/S267L mutations. Purification of the His-tagged TrfA variants (including TrfA wt, TrfA R103E, TrfA T105E, TrfA P151S, TrfA K165E, P168E, R169E, TrfA A171T, N234E, R236E, K280E, TrfA P314S, TrfA E361K, TrfA P151S/E361K, TrfA A171T/P314S, TrfA R347E) was performed essentially as described previously (26). All SUMO fusion proteins (SUMO-DBD, SUMO-WH1WH2 and SUMO-DBD/P151S) were overproduced in *E. coli* BL21(DE3) strain. The overproduction and purification of all SUMO fusion proteins were performed according to the manufacturer's manual.

In vivo binding assay

The *in vivo* binding assay was performed as described in Cereghino *et al.* (19). The *E. coli* JM109 (pAT388) strain was transformed using pAT30 plasmids containing the *trfA* gene and grown on Luria Agar (LA) (Amp, Cm) plates. Overnight cultures of JM109 (pAT388), carrying the *trfA* wild-type or mutant gene, grown in Luria Broth (LB) medium with Amp and Cm, were diluted 1:100 in LB and grown for 2 h at 37°C. Cultures were placed on ice while serial dilutions (10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} in LB) were prepared. Five-microliter aliquots were deposited on agar plates supplemented with Amp and Cm and plates with Amp, Cm and Sp. The plates were incubated at 37°C for 24 h (Amp, Cm) or 48 h (Amp, Cm, Sp). Strain growth was compared on plates with and without Sp. The assay was repeated three times.

SPR analysis

Standard surface plasmon resonance (SPR) analyses using Biacore T200 (GE Healthcare) were performed as described in the manufacturer's manual. DNA binding by analyzed proteins was studied using a 5′-biotinylated dsDNA fragment containing: five RK2 iterons, unspecific DNA of pUC19 or a single iteron with the linker sequence, immobilized on a streptavidin matrix-coated Sensor Chip SA (GE Healthcare). All oligonucleotides were commercially synthesized (Metabion, Germany) (Supplementary Table S2). The dsDNA was immobilized on the sensor surface to yield a final value of ∼60 RU for a DNA fragment containing a sequence of five iterons and for unspecific DNA of pUC19; and to 25 RU for DNA fragments containing a single iteron with the linker sequence. Experiments were run at 25°C and the running buffer was HBS-EP (150 mM NaCl, 10 mM HEPES pH 7.4, 3 mM EDTA, 0.05% Surfactant P20). In binding experiments, the buffer flow rate was set to 15 μl/min, while in kinetic experiments the buffer flow rate was 30 μl/min. Obtained data were analyzed using Biacore T200 Evaluation Software (GE Healthcare, USA). The results are presented as sensorgrams obtained after subtraction of the background response signal from a reference flow cell and from a control experiment with buffer injection.

Mass spectrometry (MS)

The reaction mixtures for the chemical crosslinking experiment contained 1 μM annealed DNA with the internal modification C2dT (5′-dimethoxytrityl-5-[N-(trifluoroacetylaminoethyl)-3-acrylimido]-2′-deoxyuridine (Glen Research, USA)), 56 μM TrfA, 40 mM HEPES-KOH (pH 7.8), 100 mM NaCl, 4 mM ATP and 11 mM magnesium acetate. The mixtures were prepared on ice, followed by incubation at 32°C for 0.5 h. The samples were crosslinked with 2.28 mM di-succinimidyl glutarate (DSG, Thermo Fisher) for 45 min at 37°C, and the reaction was stopped by addition of 50 mM Tris (pH 8.0). The mixtures were analyzed by SDS-PAGE followed by Coomassie blue staining. Bands corresponding to nucleoprotein complexes were excised, washed with ammonium bicarbonate (ABC) buffer, dehydrated with acetonitrile, reduced with 10 mM DTT, alkylated with 50 mM iodoacetamide (IAA) and washed again with ABC buffer. DNA digestion was performed with 250 U benzonase (Sigma-Aldrich) for 2 h at 37°C. Proteolysis was achieved by adding 12.5 ng/μl trypsin for overnight digestion at 37°C. Crosslinked fragments were eluted from the gels by extraction with 0.5% (vol/vol) trifluoroacetic acid (TFA) in acetonitrile (ACN). The samples were vacuum dried and dissolved in 10 μl of 0.1% TFA. The samples were then purified on ZipTip columns followed by TiO_2 purification and elution with 20 μl of 0.3 M ammonia. The obtained samples were analyzed by MALDI-TOF/TOF with an Autoflex III mass spectrometer (Bruker).

The UV crosslinking reaction mixtures contained 1 μM annealed DNA with the internal modification 5-Br-dU (5-bromo-2-deoxyuridine (Thermo Scientific)), 56 μM TrfA, 40 mM HEPES-KOH (pH 7.8), 100 mM NaCl, 4 mM ATP and 11 mM magnesium acetate. The mixtures were assembled on ice, followed by incubation at 32°C for 0.5 h. UV irradiation was performed at 120 000 μJ/cm^2 in a crosslinker (CL-1000 UV crosslinker, UVP, USA; 302 nm lamp). Subsequently, the reaction mixtures after UV crosslinking were not separated in SDS-PAGE but directly precipitated with ethanol. The samples were mixed with DHB matrix dissolved in 50% ACN with 0.1% TFA and analyzed by MALDI TOF/TOF 5800+ (AB Sciex). Data were analyzed using ProteinPilot (AB Sciex, USA) and MASCOT (Matrix Science Inc., USA) software. Peptide masses were compared with those predicted (FindPept and PeptideMass) for tryptic digestion of TrfA, allowing for missed cleavages. Orphan peaks were classified as potential crosslinked peptide pairs, and these peaks were identified by comparing their experimental masses with those calculated for any pair of tryptic peptides, including at least one internal undigested Lys residue plus 98.1 Da (the mass of the reacted DSG bridge) (27).

The sequences of the synthesized oligonucleotides (Centro Investigaciones Biológicas or Thermo Scientific) used were as follows: for chemical crosslinking—WT1, WT2, 3T, 6T, 13T, 30T, 33T, 35T, 37T and 40T oligonucleotides (Supplementary Table S2); for UV crosslinking—WTT, WTB, 53T, 56T, 57T, 63T, 64T, 58B, 64B, 69B, 71B and 78B oligonucleotides (Supplementary Table S2).

Phylogenetic analysis and determination of sequence divergence

Protein sequences of the Rep; TrfA; Orc1, 2, 3, 4, 5; and Cdc6 families were obtained from the Pfam database, version 31.0 (28). Sequences were aligned separately using Clustal Omega v1.2.2 with default parameters (29). The alignment was corrected manually. Profile alignments were performed for the WH1 and WH2 domains prior to a final round of multiple sequence alignments, followed by manual trimming of incomplete sequences. Sequence similarity searches were performed using hmmsearch using an *e*-value threshold $<10^{-3}$. Profile hidden Markov models (HMMs) were constructed with the hmmbuild program from the HMMER package (30). The HMM profile of the putative DBD was constructed based on sequence alignment of full-length TrfA homologs from the Pfam database (31). Redundant sequences were removed prior to the final alignment based on sequence identity >95%.

12 Orc, 16 Rep and 8 TrfA sequences were chosen for phylogenetic analysis based on sequence divergence and literature data. One thousand ML searches were performed using RAxML v.8.2.10 (32) with 100 rapid bootstrap replicates under the LG model of amino acid substitution and GAMMA model of rate heterogeneity with four discrete rate categories and the estimates proportion of invariable sites (LG + I + G) (33), which was determined to be the best-fit model for all families by ProtTest v3.2 following the Akaike criterion (34).

Structure prediction

The refinement procedure was initially subdivided into two independent tasks: (i) *ab initio* modeling of the DBD given the lack of structural homologs and (ii) homology modeling of WH1-WH2 domains and their proper translational positioning on the iteron sequence.

For the first task, the Rosetta module for *ab initio* folding was used to generate 100 000 initial models of the 95-aa domain (35). From the 30 top-scoring models, four were selected that yielded a 'folding funnel', i.e. had the most other high-scoring models with a RMSD of 0.75 nm or less. These were used for further MD-based refinement in which equilibrium runs were performed to assess the domains' stability over 2 μs. After ∼500 ns one of them lost its secondary structure so that only three were assessed after the extensive equilibration. For assessment, three structures from the last 200 ns were submitted to the PROSESS webserver, and a single structure was chosen that consistently ranked highest in all key parameters (overall quality, covalent bond quality, packing quality and torsion quality) (36).

For the second task, structural prediction for residues 140–382 of TrfA was carried out using the I-TASSER (37) server based on known homologs. The dsDNA was docked using ZDOCK (38) and the best complex was selected based on similarity to RepE–dsDNA complex (PDB code: 1REP). The position on iteron DNA was assessed by shifting the DNA sequence using the mutate_bases utility of X3DNA (39). Five nucleoprotein variants with different DNA shifts were assessed during 800-ns equilibrium MD runs. From these, a single model in which favorable contacts formed and DNA became bent was selected for further refinement. Now, seven different sequence shifts were generated again and sampled in MD using three criteria: (i) DNA bending energy, evaluated using the harmonic model developed in the lab of Modesto Orozco (40); (ii) distance restraints derived from the crosslinking experiment; (iii) amino acid-nucleobase (i.e. sequence-specific) contacts between WH1/WH2 and the conserved iteron sequence. Criteria (i) and (ii) were evaluated from 200-ns equilibrium runs, while (iii) was evaluated from a reweighted 200-ns metadynamics run that sampled the said number of potentially sequence-specific intermolecular contacts. Finally, a single shift was chosen that yielded satisfactory results across all criteria.

The two models, obtained with the methodology described above, were combined in Modeller (41), with a minimal-distance restraint to position the DBD near the experimentally determined crosslinking site. A total of 200 models were produced, and 10 top-scoring ones were subjected to 50-ns refinement with additional DRMSD-based restraints on segments of the backbone (helicity of the two major-groove facing α-helices; the bottom β-sheet; the equilibrated DBD). Out of the 10 models, 3 produced an extended protein–DNA interface for all domains, and remained stable in terms of secondary structure in non-restrained regions. These three structures were used for the final round of replica exchange solute tempering (REST) enhanced-sampling refinement, in which 12 replicas (4 copies of each structure) were simulated for 100 ns each with the above set of restrains (42). Samples from the lowest temperature trajectory were clustered to select the most probable structure of the complex. This selected structure was then subjected to extensive 2-μs unrestrained equilibration, followed by another round of 100 ns REST and 500 ns equilibration. Finally, two rounds of bias-exchange metadynamics (43) targeting the worst-scoring backbone and side-chain dihedral angles (as indicated by the Molprobity rotamer score (44)) were conducted, yielding the final ensemble of candidate refined structures. All MD simulations were performed with Gromacs (45) in explicit solvent, using the Amber99SB-ILDN (46,47) force field with the BSC1 correction for DNA (48). Metadynamics runs were performed using the Plumed plugin (49).

After selecting 50 structures with least dihedral outliers (as judged by Molprobity (44)), the thermodynamic effect of a number of selected amino acid substitutions on DNA affinity of TrfA was estimated for each of them using the mCSM webserver (50). TrfA protein variants containing those amino acids substitutions were purified and analyzed for DNA interaction using SPR. The SPR-derived experimental data together with data obtained with the mCSM webserver were used to select the best final model with the lowest-MSE (mean standard error). This way, the final model combines insights from (i) positioning along the DNA sequence, (ii) presence of crosslinks observed in the experiment, (iii) structure quality indicators and (iv) experimental and predicted affinity changes due to selected mutations.

RESULTS

Available data demonstrating that amino acid substitutions affecting the interaction of the TrfA protein with DNA are located both within and outside the WH domains (19,20) (Figure 1) that hint at the intricacy of the structure of the protein–DNA complex. Therefore, one could speculate that the interaction of TrfA with DNA depends on elements other than the already identified WH domains (WH1 and WH2, residues 190–382). To investigate this possibility, we first analyzed the sequence homology between TrfA and other plasmid DNA replication Rep initiators, with a focus on similarities in domain organization and/or specific motifs. We identified 1887 distinct Rep proteins and 154 TrfA-like protein sequences in NCBI database using probabilistic profile searches (see 'Materials and Methods' section). The N-terminal region is long in all TrfA-like proteins and significantly more conserved than in other Rep proteins (e.g. RepE from plasmid F, RepA from plasmid pPS10, π from plasmid R6K). The long N-terminal region was also identified in some RepB and RepC families and eukaryotic ORC. However, our analysis revealed that there were no significant sequence similarities between the N-terminal region of TrfA and these or other proteins in the NCBI database (e-value $<10^{-3}$). Interestingly, in some Rep proteins, we identified a short hydrophobic region located before the WH1 domain. This region shares a degree of similarity with a sequence within the N-terminal TrfA region (Figure 1). We thus asked how the TrfA-like proteins are related to other replication initiators. To answer this question, we performed phylogenetic analysis of the representative DNA replication initiators using the maximum likelihood method (ML) (Supplementary Figure S1). Our ML analysis of the WH domains revealed a close monophyletic relationship among all TrfA-like proteins containing extended N-termini. Moreover, the phylogenetic analysis indicated that the TrfA-like proteins are most closely related to RepC, an IncQ-type replication protein C (51,52), and belong to a common protein family together with other plasmid replication initiation factors (Supplementary Figures S1, S2). However, N-terminal region characteristic for the TrfA-like class of proteins is highly conserved and contains regions with amino acid compositions typically compatible with mostly α-helical secondary structures. This finding indicates the structural integrity of the protein and the importance of the conserved domain structure for the molecular function of TrfA.

Taking into consideration the amino acid composition and conservation of the TrfA N-terminal region, which precedes the WH1, we hypothesized that it might constitute a separate, DBD. To verify this hypothesis, we used a combined approach utilizing bioinformatical, biochemical and mutant phenotypical analysis to explore the structural requirements for the interaction of TrfA with dsDNA. First, we tested whether the putative DBD itself could interact with DNA. Despite several attempts to purify this domain as a separate polypeptide, we were unable to obtain a preparation that was suitable for DNA-binding assays. Instead, we cloned and purified protein chimeras consisting of SUMO and the TrfA domains (SUMO-DBD and SUMO-WH1WH2). We also purified SUMO-DBD/P151S, a protein carrying a substitution that prevents the TrfA protein to interact with dsDNA (20) (Figure 1). The purified proteins were examined using surface plasmon resonance (SPR; 'Materials and Methods' section) to test the interaction with dsDNA containing five iterons (Figure 2AB). The analysis was conducted only for comparison of the tested proteins and to verify their ability to interact with DNA. The SUMO-DBD and SUMO-WH1WH2 both formed complexes with biotinylated DNA fragments attached to a streptavidin-coated sensor chip (Figure 2B). The response when SUMO-DBD protein was bound to five iterons sequence was similar to that obtained with SUMO-WH1WH2, although the molecular weight of SUMO-DBD (19 kDa) is 1.3 times smaller than that of SUMO-WH1WH2 (25 kDa). Therefore, a relatively high number of SUMO-DBD molecules attached to the dsDNA fragment. No increase in the response signal was detected in the experiment with SUMO-DBD/P151S (Figure 2B), indicating that the P151S substitution results in the inability of a putative DBD to interact with dsDNA. No response was observed in the negative control reaction with the SUMO protein alone (Figure 2B).

To further analyze the nucleoprotein complex established by TrfA and dsDNA, we decided to experimentally detect the protein residues in contact with DNA. Chemical or UV crosslinking followed by MALDI MS analysis was used ('Materials and Methods' section; Figure 3). In experiments involving chemical crosslinking with DSG, a crosslinking reagent that reacts specifically with primary amines (53), we used dsDNA fragments containing the sequence of the second iteron and the associated flanking regions (Figure 3A–J). Each dsDNA fragment used in the experiment contained a modification of one particular thymine, substituted with C2dT. Among all of the designed dsDNA fragments (Supplementary Figure S3A and B), those that formed stable crosslinked complexes with TrfA (3T, 37T, 40T) were analyzed using mass spectrometry (MS). In MALDI-TOF analysis (Figure 3A–C), after subtraction of peaks corresponding to the TrfA protein alone and to protein contaminants (keratin, benzonase, trypsin, MALDI matrix and other *E. coli* proteins), the remaining peaks should correspond to peptides that form nucleoprotein complexes. Based on the obtained masses, a single sequence was predicted to be crosslinked (to oligonucleotide 37T): LMCGSDSTRVK (Figure 3C–E and Supplementary Figure S3B,C), located in the WH2 domain (339–349 aa). The sequence of this peptide was confirmed by MALDI-TOF/TOF. We assume that most likely the

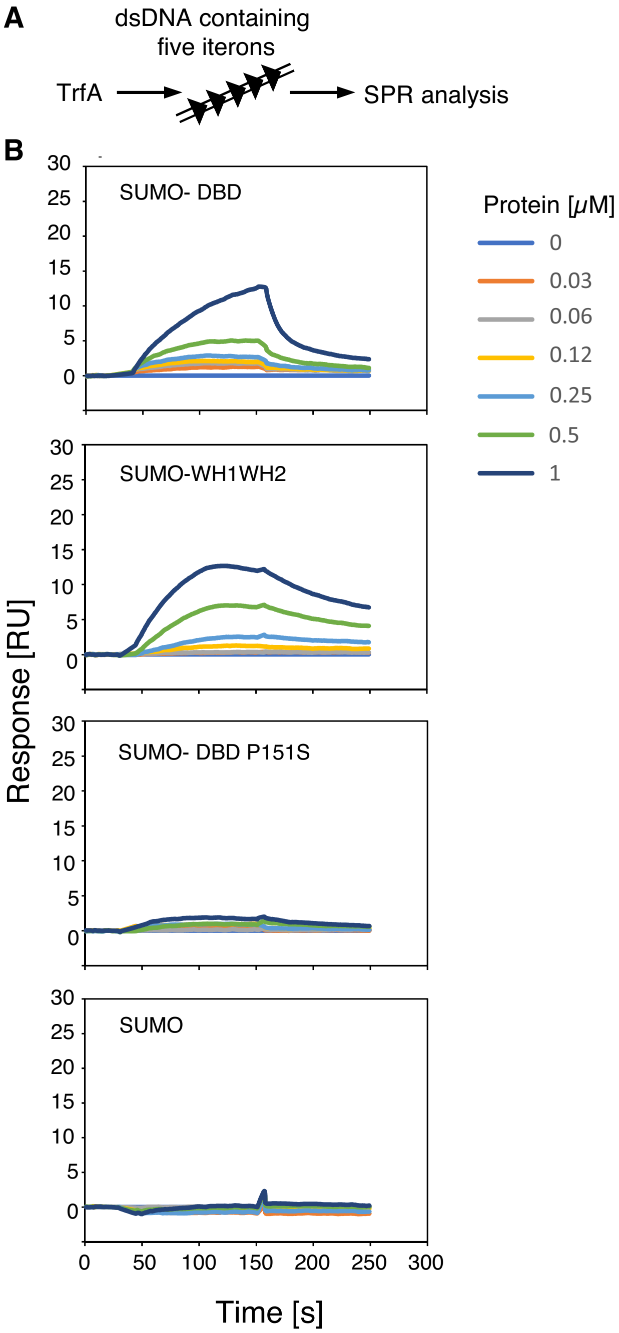

Figure 2. SPR analysis of dsDNA binding by TrfA protein constructs. (**A**) Schematic representation of SPR analysis. (**B**) The SPR analysis of dsDNA binding was performed using the following proteins: SUMO-DBD, SUMO-WH1WH2, SUMO-DBD/151S and SUMO ('Materials and Methods' section). Sensorgrams show the SPR analyses of binding of each protein variants to a double-stranded DNA fragment containing five RK2 iterons. Injections contained the indicated concentrations of protein variants in HBS-EP buffer. HBS-EP was also used as a running buffer.

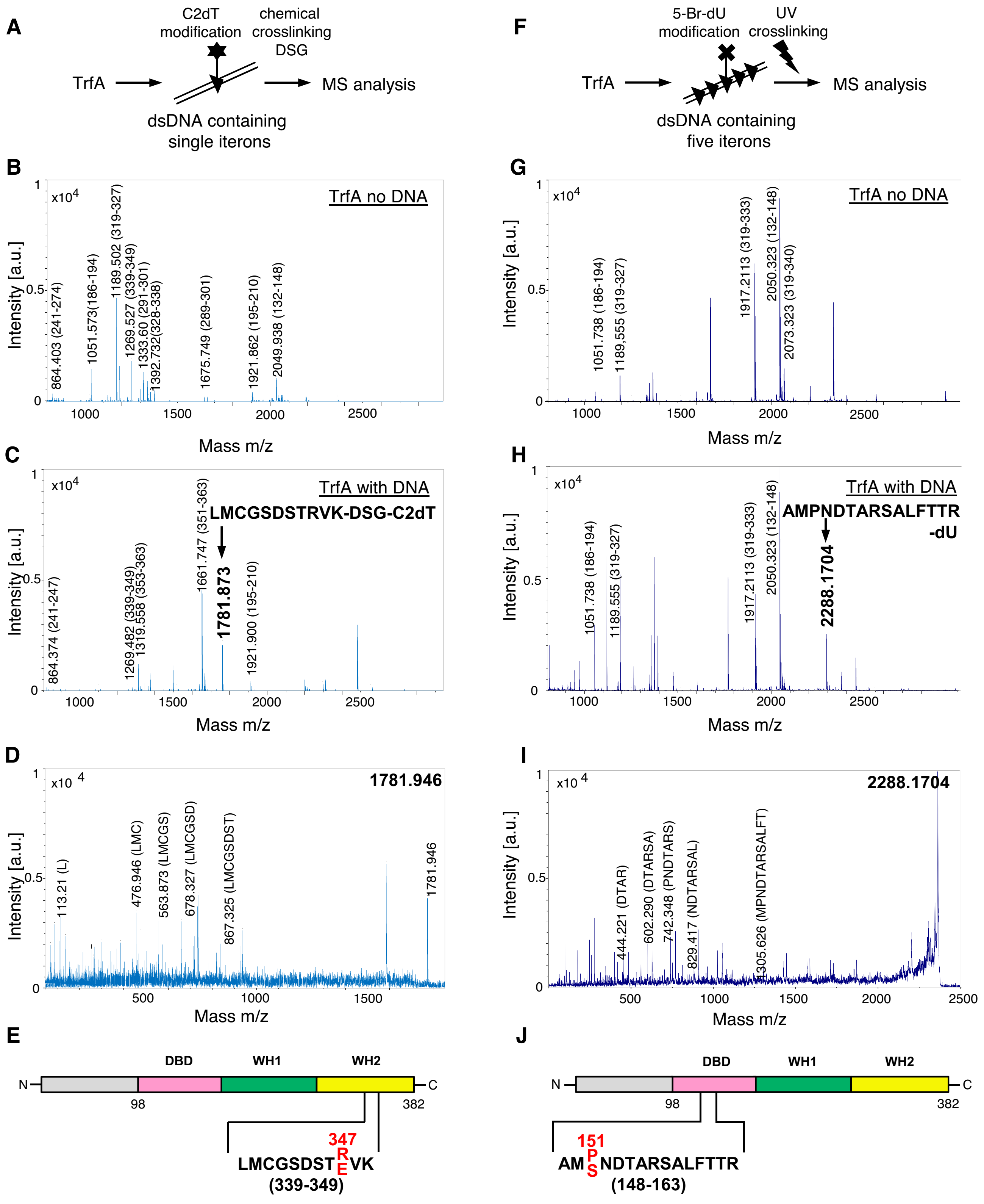

Figure 3. Identification of TrfA residues interacting with DNA. (**A**) Single iteron target scheme. (**B**) MALDI TOF spectrum of trypsin-digested TrfA alone and (**C**) TrfA crosslinked with DNA. Highlighted peak indicates a potential crosslink product. (**D**) MALDI TOF/TOF spectrum of peptide LMCGSDSTRVK fragmentation. (**E**) Sequence and location of the identified peptide marked on the TrfA scheme. (**F**) Five iterons target scheme. (**G**) MALDI TOF spectrum of trypsin-digested TrfA alone and (**H**) TrfA crosslinked with DNA. Highlighted peak indicates a potential crosslink product. (**I**) MALDI TOF/TOF spectrum of peptide AMPNDTARSALFTTR fragmentation. (**J**) Sequence and location of the identified peptide are marked on the TrfA scheme. Substitution affecting DNA–TrfA interactions are colored in red.

crosslink was generated with lysine or arginine residues. For peptides crosslinked to the 3T and 40T oligonucleotides, the sequences predicted after MALDI-TOF analysis, were not confirmed by MALDI-TOF/TOF.

Next, we performed UV crosslinking analysis using dsDNA fragments containing the sequences of all five iterons and the corresponding flanking regions (Figure 3F–J). The UV irradiation can induce a formation of covalent bond between 5-Br-dU and the particular amino acid residues (54,55). Ten DNA fragments were tested, each with a different thymine in the third iteron changed to 5-Br-dU. Of these labeled dsDNA fragments (Supplementary Figure S4A and B), only six formed stable crosslinked complexes with TrfA, as observed in silver stained polyacrylamide gels (53T, 56T, 57T, 58B, 69B, 71B; Supplementary Figure S5), were further analyzed using MS. For these six selected dsDNA fragments, crosslinking reactions with TrfA were performed, and the whole reaction mixtures were precipitated and prepared for MALDI-TOF analysis. Three peaks were identified in the MS: one peak for a peptide with the predicted sequence AMPNDTARSALFTTR (located in the putative DBD, 148–163 aa) crosslinked to the 53T DNA fragment, and two peaks for a peptide with the predicted sequence NKKIPR (also located in the putative DBD, 163–169 aa) crosslinked to the 56T and 57T DNA fragments. To confirm the predicted peptide sequences, the primary peaks were further analyzed using MALDI-TOF/TOF. The fragmentation spectrum was obtained only for the AMPNDTARSALFTTR peptide (Figure 3I,J; Supplementary Figure S4B and C). The low quality of the fragmentation spectra obtained for NKKIPR (Supplementary Figure S6) did not allow identification of the predicted sequence. Nevertheless, in the predicted NKKIPR sequence, we selected substitutions (K165E and R169E) for biochemical analysis to independently confirm the importance of these residues for TrfA interaction with DNA (Supplementary Figure S6B and description below). The lysine and arginine residues that are present in the identified peptides we considered as most likely candidates for interaction with DNA, although other amino acid residues could also be UV crosslinked with 5-Br-dU (54,55).

Both mass spectrometry crosslinking analysis and SPR data on SUMO-DBD binding to iterons indicated the importance of the TrfA N-terminal region for DNA binding (see above). However, the structure of this region remained undefined. We thus aimed to predict the structure and arrangement of the N-terminus of TrfA when it is in complex with iteron DNA. This region was previously predicted to be intrinsically disordered (17) and most likely causes difficulties in TrfA purification and obtaining preparations suitable for crystallographic analysis. To gain insight into the full structure of the TrfA–dsDNA complex, we applied an integrative structure prediction approach combining homology modeling of WH1-WH2 domains and a fragment of DBD (140–382 aa) and *de novo* modeling of remaining fragment of DBD (98–140 aa), guided by experimental crosslinking/MS data. After model building, selection and optimization, the full structure of the complex was assembled with a minimal-distance restraint to position the DBD near the experimentally determined crosslinking site, and subjected to extensive MD-based refinement (for a full description of the model construction workflow see Supplementary Figure S7 and 'Materials and Methods' section). On the basis of rotamer quality in the final structural ensemble, 50 best structures were selected. The models include WH1 and WH2 domains, and a new DBD composed of six α-helices and two antiparallel β-strands. In all of 50 structures the N-terminal loop of DBD interacted with the minor groove of DNA, while helices 5 and 6 and the interhelical loop contacted the major groove of DNA. To verify our predicted candidate structures and select the most plausible model, amino acids positioned at the DNA–protein interface were selected for mutagenesis (R103E, T105E, R163E, K165E, K166E, P168E, R169E, N234E, R236E, K280E, R347E, V348E). Additionally, we chose the residues previously identified (19,20) as possibly involved in TrfA interaction with DNA (P151S, E361K, A171T, P314S) (Figure 1). Altogether 13 TrfA protein variants were purified, each containing a single amino acid as we were not able to obtain the R163E, K166E and V348E constructs. These proteins, along with the wild-type TrfA, were subjected to CD spectra analysis (Supplementary Figure S8 A and B), SPR-measurements of DNA interaction analysis (Supplementary Figure S9 and Table S3) and *in vitro* replication tests (Supplementary Tables S4 and S5). The CD spectra analysis did not show significant differences between purified TrfA protein variants (Supplementary Figure S8). The SPR analysis showed that K165E, R169E, N234E, R236E, R347E and P151S TrfA variants were defective in iteron binding, while two other (R103E, K280E) showed substantially decreased ability to interact with DNA (Supplementary Figure S9 and Table S3). We also observed reduction, especially emphasized by binding constants, of DNA binding for TrfA P314S (Supplementary Table S3). TrfA variants T105E, P168E, A171T and E361K retained the ability to interact with DNA or even bound to DNA with higher affinity (Supplementary Figure S9 and Table S3). TrfA protein variants defective or having reduced DNA-binding ability also could not initiate plasmid DNA replication *in vitro* or had substantially reduced replication activity (Supplementary Tables S4 and S5). Interestingly, replication activity was to some extend reduced by substitutions resulting with elevated DNA binding (e.g. A171T and E361K). Most likely optimal protein DNA-binding avidity, not too low and not too high, is required for the maximal replication activity. We used SPR-based binding kinetics to calculate the change in binding affinity for each of the TrfA variants with respect to the wild-type protein ($\Delta\Delta G_{\mathrm{exp}}$), and compared it with structure-based predictions ($\Delta\Delta G_{\mathrm{pred}}$). Out of 50 candidate structural models, we selected the one with the lowest root-mean square error (RMSE) to experimental data (Supplementary Figure S10 and Figure 4).

The obtained data indicated that besides WH1 and WH2, the interaction of TrfA protein with DNA involves the new N-terminal DBD. Motivated by the fact that substitutions affecting the interaction of TrfA with DNA are mostly located in the novel DBD and in WH2, we asked how the two domains affect the protein avidity to interact with DNA. We tested whether the amino acid substitutions in separate domains of TrfA could have compensatory effects. Compensation might suggest that the interaction of TrfA with dsDNA is based on binding sites located in the distinct do-

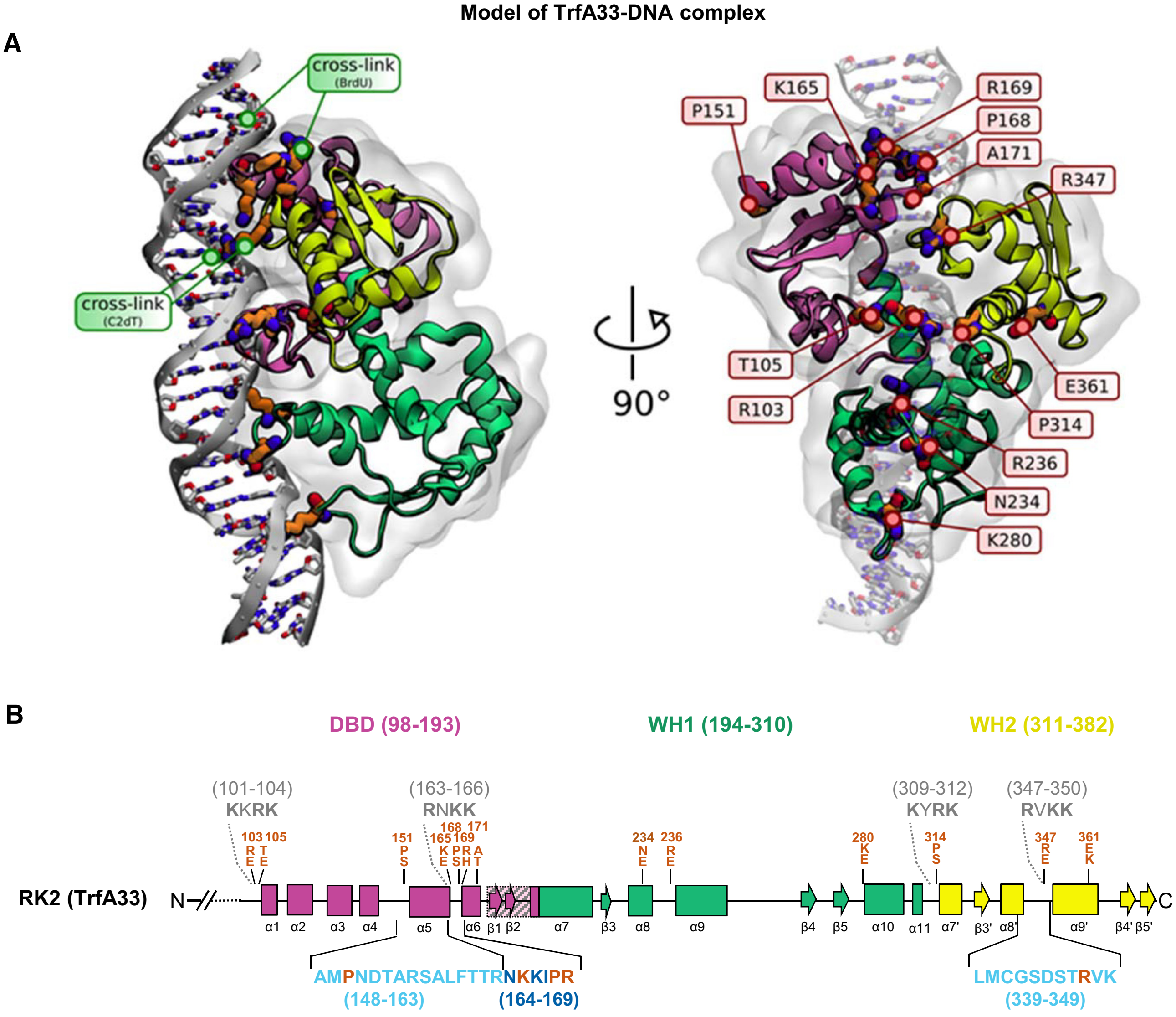

Figure 4. Structural model of TrfA protein. (**A**) Structural model of TrfA33 in a complex with double-stranded DNA. Positions of the introduced amino acid substitutions affecting the TrfA affinity to dsDNA are indicated (orange). Positions of crosslinks between amino acid residues of TrfA and bases of DNA are marked as green spots. (**B**) Schematic representation of TrfA protein, with predicted alpha helices (α) and beta strands (β) represented as colored boxes and arrows, respectively. The hydrophobic region of TrfA protein is marked as a dashed box. The basic patches identified in TrfA sequence are marked in gray (R/KnR/KK). In both model and schematic representation, the three TrfA domains predicted by amino acid sequence analysis are marked as follows: new DBD domain in magenta, WH1 domain in green and WH2 domain in yellow. Peptide sequence predicted by MALDI-TOF MS for interaction with dsDNA (dark blue), sequence identified and confirmed by MALDI-TOF/TOF MS as peptide sequences in contact with dsDNA (light blue), amino acid substitutions affecting TrfA interaction with DNA (orange).

mains. The TrfA protein variants containing substitutions within the WH2 or in DBD (TrfA P151S, TrfA A171T, TrfA P314S and TrfA E361K), and double mutant proteins with substitutions in both the WH2 domain and DBD (TrfA P151S/E361K, TrfA A171T/P314S) (Supplementary Figure S8) were used in the experiments. We performed *in vivo* and *in vitro* experiments to analyze the interactions of the TrfA protein variants with DNA. *In vivo* tests were based on the binding assay described by Cereghino *et al.* (19) (Figure 5AB, Supplementary Figure S11). In this assay, resistance to spectinomycin results from binding of TrfA to its binding site—two iteron sequences—within the strong constitutive promoter PconII. When the iteron sequences remains unbound, the PconII promoter is active, preventing the expression of the *aadA* gene from the complementary strand and resulting in sensitivity to spectinomycin. The obtained results, shown in Figure 5B, clearly demonstrated defectiveness in DNA binding as a result of the single substitutions P151S (DBD) and P314S (WH2) and compensation by E361K (WH2) and A171T (DBD), respectively. The sensitivity to spectinomycin, and the resulting lack of cell growth on a medium supplemented with this antibiotic, was not caused by the lack of TrfA protein in cells expressing the protein variants P151S and P314S, as shown by the western blot analysis (Supplementary Figure S12). The *in vitro* SPR tests confirmed the results obtained in

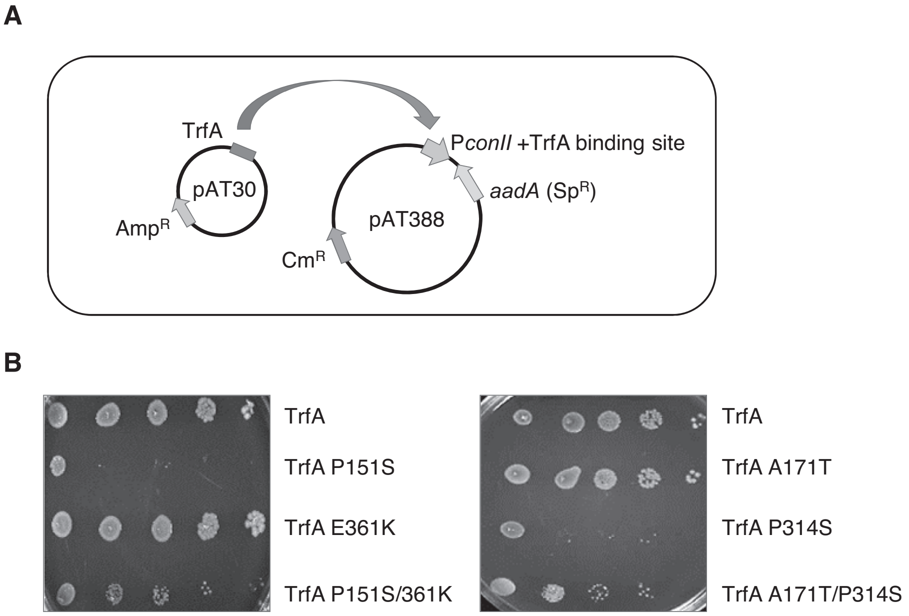

Figure 5. *In vivo* binding of TrfA protein variants to double-stranded DNA. (**A**) The scheme of the experiment. *Escherichia coli* JM109 cells with the pAT388 plasmid (providing chloramphenicol resistance), harboring the binding site for TrfA within the PconII promoter, were transformed with the pAT30 plasmid derivatives (providing ampicillin resistance), containing the genes for the TrfA variants. Binding of TrfA with its specific binding site results in inhibition of the PconII promoter and expression of the *aadA* gene from the complementary strand (thus providing spectinomycin resistance). (**B**) Serial dilutions of cell suspensions grown on LA medium containing ampicillin, chloramphenicol and spectinomycin. The growth of cells in spectinomycin indicates TrfA binding within the specific DNA sequence. The lack of growth indicates that the mutations introduced to TrfA result in the inhibition of protein binding to DNA.

the *in vivo* binding assay: we observed compensatory effects of E361K or A171T substitutions in TrfA P151S/E361K and TrfA A171T/P314S mutants tested for interaction with DNA (Supplementary Figure S9 and Supplementary Table S3). The compensation was more pronounced in case of TrfA P151S/E361K. The compensatory effect of E361K and A171T substitutions on activity of TrfA P151S and TrfA P314S proteins variants, respectively, was also observed in *in vitro* replication assay (Supplementary Table S5). The ability to initiate plasmid DNA replication was increased for TrfA P151S/E361K and TrfA A171T/P314S comparing to TrfA P151S and TrfA P314S proteins (Supplementary Table S5).

We also asked how the TrfA domains contribute to sequence specificity of TrfA interaction with DNA. To answer this question, we performed the SPR analysis with SUMO fusion proteins (SUMO-DBD, SUMO-WH1WH2). The fusion proteins and the whole 33kDa TrfA were tested for interaction with specific (containing five iterons) or unspecific (pUC19) DNA (Figure 6). The obtained results showed that although SUMO-DBD and SUMO-WH1WH2 bind DNA fragments equally well regardless of the presence of iterons, the TrfA protein interacts with the iteron-containing DNA much more effectively compared to pUC19 (Figure 6B and D). In our experiment only the whole TrfA, consisting of DBD, WH1 and WH2, was capable to interact with DNA in a sequence-specific mode. It must be pointed out that when the interaction tests were performed with dsDNA containing a single iteron, the TrfA interaction was substantially decreased and we were not able to detect reasonable response signal from SUMO-WH1WH2 or SUMO-DBD (Supplementary Figure S13). That results emphasized the essentiality of all TrfA domains for the protein interaction with DNA and postulated (23) cooperativity of TrfA interaction with DNA containing multiple binding sites.

DISCUSSION

Our analysis characterized a new family of replication initiation proteins with a unique domain composition that is crucial for interactions with DNA. We showed that members of this new class of Rep proteins, which we called the TrfA-like family, contain three domains: the previously described WH1 and WH2, and an additional DBD, newly identified in this work as containing a distinct structural arrangement. Our experimental data, combined with *in silico* structure prediction, showed that all three domains are essential for the sequence-specificity of TrfA interaction with DNA.

The identified TrfA-like proteins were annotated to numerous different bacterial species. This result was expected because the *trfA* gene, as a part of a broad host range replicon, can be broadly distributed among diverse groups of bacteria. The sequence differences among TrfA-like proteins are possibly a result of an ongoing evolutionary speciation in different bacterial hosts. In *Shewanella oneidensis,* rapid adaptation via mutations in the replication initiation gene *trfA1* was shown to reduce the fitness cost of TrfA1 due to changes in the interaction with the host DNA helicase DnaB (56). Moreover, strains expressing evolved TrfA1 variants showed higher growth rates than those expressing ancestral TrfA1. The described mechanism may contribute to the diversity of TrfA-like proteins and, consequently, can

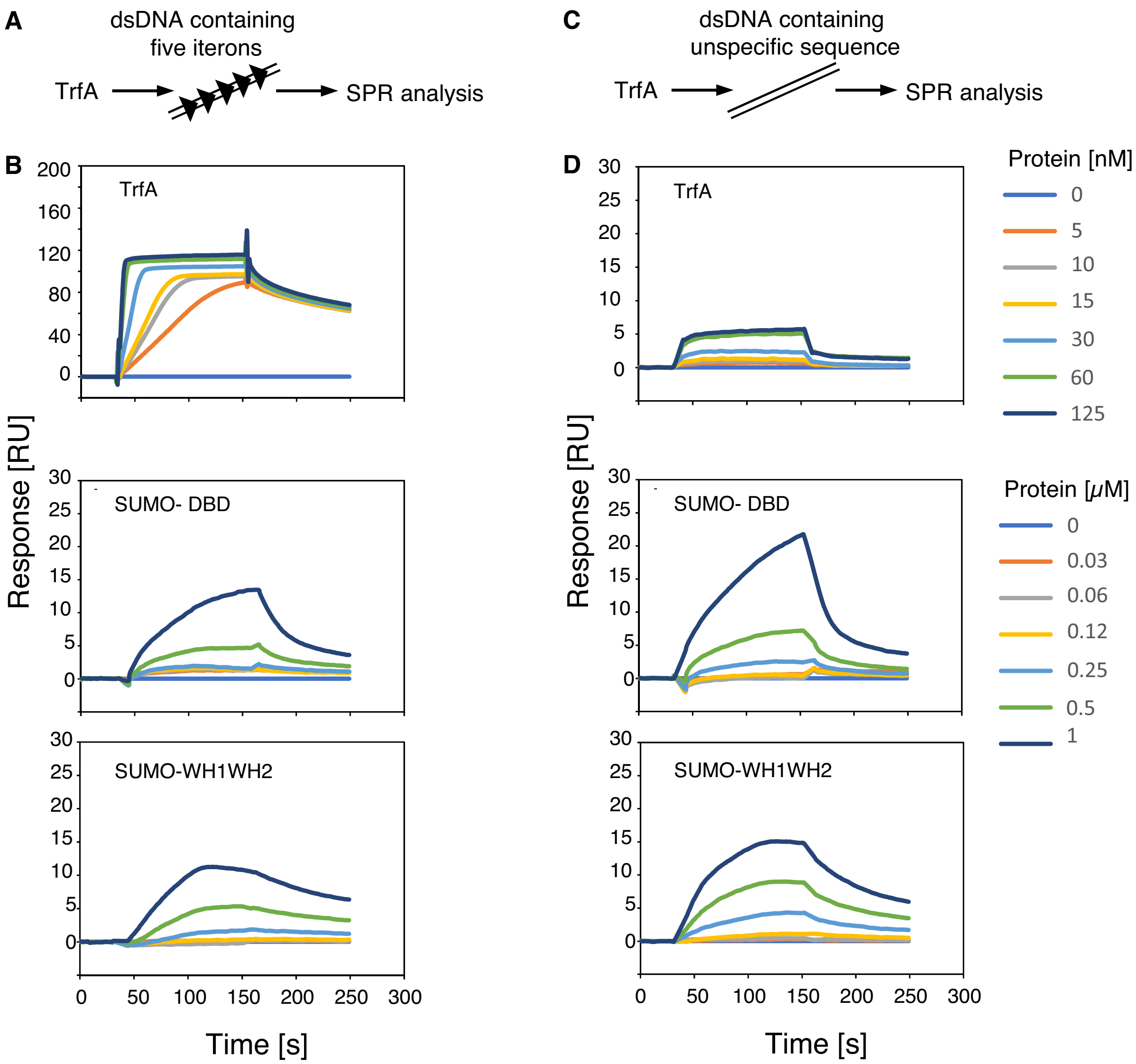

Figure 6. SPR analysis of dsDNA binding by TrfA and SUMO fusion proteins. SPR analysis with DNA fragment containing sequence of five iterons (**A** and **B**) and fragment of pUC19 vector sequence (**C** and **D**). Schematic representation of SPR analysis with DNA fragments (**A** and **C**). (**B** and **D**) The SPR analysis of dsDNA binding was performed using the following proteins: TrfA, SUMO-DBD, SUMO-WH1WH2 ('Materials and Methods' section). Sensorgrams show the SPR analyses of binding of each protein variants to a double-stranded DNA fragment. Injections contained the indicated concentrations of protein variants in HBS-EP buffer. HBS-EP was also used as a running buffer.

affect the variability in the plasmid host range. Interestingly, genes encoding TrfA-like proteins were annotated both to plasmid DNA and chromosomes. However, it is possible that these annotations were an artifact of metagenomics sequencing and data analysis. Alternatively, recombination of plasmid DNA with bacterial chromosomes is also a possibility. Notably, the identified TrfA-like protein sequences shared some degree of similarity, and all of these sequences had an N-terminal region that was identified by us as an additional domain important for interactions with DNA. This N-terminal region of TrfA protein is also important for replisome assembly of *E. coli*, what was shown previously. The QLSLF motif (residues 134–138) was identified as required for TrfA interaction with the β-clamp (57,58). It cannot be excluded that in this N-termini there are also residues required for interaction with other replication initiation proteins e.g. DnaB, that is known to form a complex with TrfA (22,59). The presence of more than two DBDs in the structure of a replication initiator is not restricted to TrfA-like proteins. It was found that RctB, the replication initiator of the second chromosome of *Vibrio cholerae*, also has a multidomain organization. In addition to domains 2 and 3 with a structure similar to the WH domains of plasmid Rep proteins, RctB also contains domain 1 that interacts with DNA, and domain 4 with an undefined function (60). However, our analysis indicated that the DBD of TrfA proteins is distinct from that identified in RctB. Due to the low sequence similarity between TrfA and RctB, this protein was not included in the phylogenetic tree. In contrast, eukaryotic ORC proteins, the WH domain of which is similar to the WH1 domain of plasmid Rep proteins, were included in our analysis (Figure 1 and Supplementary Figure S1). Similarities between the WH1 domain of plasmid Rep proteins and the WH domains of archaeal Orc and yeast Orc4p proteins were pointed out previously (61).

Sequence comparison revealed that the N-terminal region of TrfA-like proteins is conserved within the group. Most importantly, we proposed that the region of TrfA analyzed in this work (residues 98–192) forms a distinctive DBD containing characteristic motifs and features (Figure 4). Interestingly, in contrast to WH1 and WH2, we did not identify the canonical WH fold (α1-β1-α2-α3-β2-β3) or the typical HTH motif in the DBD (Figure 4). Based on our structure prediction, this region forms a helical bundle (α1, α2, α3, α4, α5). With using fold recognition software MADOKA (62), we identified similar helical assembly in a few other proteins (e.g. transcriptional activator GCN4, PDB_id: 5APX; 30S ribosomal protein S20, PDB_id: 5XYU). In TrfA this helical bundle is followed by α6 and a β-hairpin. The interaction of α5 and neighboring amino acid residues with DNA has been clearly supported by our MS analysis, structural predictions, MD simulations and mutant analysis. Four TrfA mutants, with either decreased or increased DNA binding affinities, have been identified within this region (P151S, K165E, R169E and A171T) (Figure 4). According to our model, substitutions K165E, R169E and A171T are located in the proximity to DNA. The residue P151 is located between α4 and α5 and might be important for stabilization of the helical bundle. The predicted binding interface matches the sequence of the peptides crosslinked with DNA identified in MS experiments. We noticed that two substitutions (K165E, R169E) that affected TrfA interaction with DNA are located in a specific basic patch RnKKnnR (163–169aa) (Figure 4B). Similarly, the substitution R103E resulting in reduction of TrfA–DNA interaction is located in another basic patch KKRK (101–104aa) in DBD (Figure 4B). Beside these basic patches, we recognized the presence of the hydrophobic region in the DBD domain of the TrfA protein (176–192 aa) that was also identified in all TrfA-like proteins, as well as in the other plasmid-encoded Rep (Figure 1). In our model, this hydrophobic patch forms the link between DBD and WH1, where two β-strands connecting the two domains can be identified (Figure 4AB). We believe that this region is also important for the protein–DNA interaction, as shown by mutation of the H179 residue that disrupts the binding of TrfA to DNA (19).

Substitutions affecting the interaction of TrfA with DNA were also identified within the WH1 and WH2 domains (N234E, R236E, K280E, P314S, R347E and E361K) (Figure 4). Based on structure prediction, all these substitutions, except for E361K, are located in close proximity to DNA. It is likely that E361 affects TrfA interaction with DNA by stabilizing WH2 structure. The importance of residue R347 was confirmed by MS experiment. Interestingly, some of the identified residues were located within or in the immediate vicinity of another basic patches KnRK (309–312aa) and RnKKnR (347–352aa). It is very likely that the clusters of basic amino acids are important contributors to the interactions of replication initiation proteins with DNA, as suggested by the identification of similar patches of basic residues in eukaryotic Orc1 protein homologs from different species (e.g., *S. cerevisiae*, *Mus musculus*, *Homo sapiens*) (13). Both *in vivo* and *in vitro* studies have shown that two basic residues from the patch identified in the yeast Orc1 protein are important for DNA binding (13). The RnKK and KnKK sequences are also considered to be conserved acetylation motifs in nonhistone proteins (63,64), and in bacteria, acetylation of proteins can regulate processes such as RNA and DNA metabolism, enzyme activity, motility and cell shape (65); therefore, it can be speculated that lysines within these basic patches found in the TrfA protein could be subjected to such post-translational modification. It is hence worth exploring further if the replication activity of TrfA proteins can be regulated via acetylation as in case of the bacterial replication initiator DnaA (66). The level of DnaA acetylation that inhibits replication activity depends on cell growth, and this inhibition can be reversed when DnaA activity is required. To date, there are no data regarding if the plasmid Rep proteins are acetylated and if this modification might regulate Rep protein activity.

Our structure prediction of the TrfA protein revealed that the protein domains DBD, WH1 and WH2 bind to one side of a DNA molecule (Figure 4A, left and right). It was previously shown that WH1 contains a dimerization interface and is involved in the stabilization of the Rep complex with DNA (16,17,67). Our data are consistent with these previous observations. Although we were unable to crosslink TrfA with DNA via WH1, in our model the WH1 domain remains in proximity to DNA (Figure 4) and we identified two residues (N234 and R236) the substitutions of which affect the interaction with DNA (see Supplementary Figure S9 B and Supplementary Table S3). In the WH1 domain of TrfA, two substitutions (G254D and S267L) resulting in the constitutive formation of a protein monomer have also been previously described (26). Similarly, in the π protein from the R6K plasmid, the WH1 domain has been shown to be responsible for protein dimerization, and it interacts with DNA nonspecifically through phosphate groups, whereas the WH2 domain interacts with iteron DNA via base-amino acid contacts (18). The WH1 domain is also responsible for dimerization of the RepE protein from plasmid F (16). In the crystal structure of the nucleoprotein complex of the RepE dimer, only the WH2 domain contacts DNA, but that changes in case of a monomer protein, where also the WH1 contributes to the interaction with DNA (16). The superimposition of constructed by us model and crystal structure of nucleoprotein complex of RepE (PDB id: 1REP) revealed that helix α9′ located within WH2 of TrfA corresponds to and is similarly oriented to DNA as critical for DNA interaction helix α4′ of RepE (Supplementary Figure S14) (16).

In our experiments, we demonstrate that the DBD is capable of interacting with DNA. Most likely due to the intrinsic instability of DBD, we were unable to purify this domain as a separate peptide, necessitating its purification as a fusion protein with a stabilizing tag. Although both the His_6-tagged Rep and native Rep proteins behave similarly (68,69) and plasmid Rep proteins are usually purified as His_6-tagged versions (17,24,58,70), we purified DBD chimera with SUMO to stabilize DBD and excluded any influence of the basic histidine residues on the DNA binding affinity. Although the DBD can bind to DNA on its own, as tested in experiments with SUMO-DBD, both the DBD and the WH1 and WH2 domains are essential for TrfA nucleoprotein complex formation because a single substitution in one of these domains results in a protein defective in DNA binding and replication activity. The analysis of the constructed mutants showed that in many cases a single substitution caused complete inability of mutant TrfA to interact with DNA. This observation indicates that most likely the TrfA avidity required for DNA complex formation depends altogether on interactions located in all three domains. This assumption is also supported by the fact that the response signal, obtained when TrfA protein was analyzed for interaction with DNA, was much higher compared to the response obtained with SUMO fusion proteins containing only DBD or WH1, WH2 domains. The difference was even more noticeable when binding to DNA fragment containing single iteron sequence was analyzed. It is possible that due to low binding affinity to single iteron sequence and under applied conditions with limited DNA molecules immobilized on the sensor, the response signal was near or below detectable level in case of SUMO fusion proteins. It could also be speculated that SUMO-DBD complex with DNA is somehow stabilized on the longer DNA molecules. Also, we cannot exclude that SUMO might negatively affect the interaction of the chimera with DNA. Nevertheless, both *in vitro* and *in vivo* tests showed that the effects of substitutions located in DBD, resulting in mutant defective in DNA binding, can be compensated by the introduction of a substitution enhancing DNA interaction located in WH2 and vice versa. All these data point out essentiality of each one domain of TrfA for the protein binding to DNA. The functional advantage empowered by this three-domain structural arrangement was emphasized by the sequence specificity of TrfA interaction with DNA. The interaction with DNA fragments containing iterons has much higher affinity compared to the protein interaction with DNA containing unspecific sequence. Neither the DBD nor WH1 or WH2 domains are capable of interacting with DNA in a sequence specific mode individually, but when joined together in the full-length TrfA protein, the combination of binding sites located in those domains brings about an additive or all-or-nothing effect, resulting in a sequence specific interaction with DNA. Based on TrfA structural model, each domain directly interacts a few nucleotides (2–3bp) that creates the critical contacts, but all three domains only together can anchor protein to DNA via those a few nucleotides that are correctly spaced within entire 23 bp iteron sequence. This most likely provides a sequence specificity of TrfA interaction with iterons. Although our data indicate the involvement of all three domains in TrfA complex formation with DNA, the experiment with a DNA fragment containing a single iteron shows the importance of the cooperative interaction with DNA containing multiple binding sites previously postulated not only for TrfA (23), but also for other Rep proteins (68,71,72). The interactions involved in the formation of multimolecular complexes of Rep on DNA remain to be described. Our data show the complexity of structural requirements for a functional, sequence specific Rep interaction with DNA. Protein regions considered as disordered and having intrinsic instability might form domains and/or structures that play a crucial role in the formation of a protein complex with DNA. It is very likely that additional, not yet identified, interactions or functions could also be located within those regions.

DATA AVAILABILITY

The MS data obtained in this work have been deposited to the ProteomeXchange Consortium via the PRIDE (73) partner repository with the dataset identifier PXD013286. The model of nucleoprotein complex of TrfA protein has been deposited in ModelArchive (https://modelarchive.org/doi/10.5452/ma-jisol) and PDB-Dev (PDBDEV_00000068).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Elzbieta Purta for help during initial TrfA model construction and Magdalena Rajewska for corrections of the manuscript. We thank for the technical support of Vivian de los Ríos (Proteomics facility, CIB-CSIC), and Paulina Czaplewska and Katarzyna Macur (Proteomics facility, IFB) with the MS experiments.

Authors contribution: K.W. constructed the TrfA point mutants, purified the proteins, performed *in vivo* and *in vitro* activity tests of TrfA variants. E.Z. performed MS experiments, constructed strains, purified proteins and performed *in vivo* and *in vitro* TrfA variants activity tests. K.B. constructed strains and with KW purified and performed SPR analysis with SUMO fusion proteins. J.C., M.W. and I.G. performed the bioinformatic analysis, prepared the TrfA-33 model and were involved in figure preparation. U.U. and J.M.B. were involved in initial TrfA model construction. M.M. was involved in MS analysis with single iteron DNA. B.T. performed phylogenetic analysis. U.W. performed initial experiments with TrfA double mutants. R.D. performed CD analysis of purified proteins. K.B., K.W., B.T., J.C. and I.K. prepared figures and their legends. R.G. advised MS experiments, supervised MS analysis with single iteron DNA, discussed results and the manuscript. I.K. conceived research and supervised the project. I.K. and K.W. wrote the manuscript with the help of K.B., B.T., M.W. and J.C. The computational resources were provided by the Polish Grid Infrastructure (PL-GRID) and the Academic Computer Centre in Gdansk (TASK). E.Z. was supported by the MPD/2010/5 project operated within the Foundation for Polish Science International PhD Projects (MPD) Programme co-financed by the EU European Regional Development Fund.

FUNDING

National Science Centre [2012/04/A/NZ1/00048 to I.K.; 2017/26/D/NZ1/00239 to K.W.]; Foundation for Polish Science [TEAM, POIR.04.04.00-00-5C75/17-00 to I.K.]; International Institute of Molecular and Cell Biology in Warsaw (to J.M.B.); Ministerio de Economía, Industria y Competitividad (MINECO/AEI) [BIO2012-30852, RTI2018-094549-B-I00 to R.G.]. Funding for open access charge: Foundation for Polish Science [TEAM, POIR.04.04.00-00-5C75/17-00].

Conflict of interest statement. Janusz M. Bujnicki is an Executive Editor of *Nucleic Acids Research*.

REFERENCES

1. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, REVIEWS001.
2. Sutton,M.D. and Kaguni,J.M. (1997) The Escherichia coli dnaA gene: four functional domains. *J. Mol. Biol.*, **274**, 546–561.
3. Fujikawa,N., Kurumizaka,H., Nureki,O., Terada,T., Shirouzu,M., Katayama,T. and Yokoyama,S. (2003) Structural basis of replication origin recognition by the DnaA protein. *Nucleic Acids Res.*, **31**, 2077–2086.
4. Roth,A. and Messer,W. (1995) The DNA binding domain of the initiator protein DnaA. *EMBO J.*, **14**, 2106–2111.
5. Richardson,T.T., Harran,O. and Murray,H. (2016) The bacterial DnaA-trio replication origin element specifies single-stranded DNA initiator binding. *Nature*, **534**, 412–416.
6. Duderstadt,K.E., Chuang,K. and Berger,J.M. (2011) DNA stretching by bacterial initiators promotes replication origin opening. *Nature*, **478**, 209–213.
7. Ozaki,S., Kawakami,H., Nakamura,K., Fujikawa,N., Kagawa,W., Park,S.Y., Yokoyama,S., Kurumizaka,H. and Katayama,T. (2008) A common mechanism for the ATP-DnaA-dependent formation of open complexes at the replication origin. *J. Biol. Chem.*, **283**, 8351–8362.
8. Costa,A., Hood,I.V. and Berger,J.M. (2013) Mechanisms for initiating cellular DNA replication. *Annu. Rev. Biochem.*, **82**, 25–54.
9. Dueber,E.L., Corn,J.E., Bell,S.D. and Berger,J.M. (2007) Replication origin recognition and deformation by a heterodimeric archaeal Orc1 complex. *Science*, **317**, 1210–1213.
10. Bleichert,F., Botchan,M.R. and Berger,J.M. (2015) Crystal structure of the eukaryotic origin recognition complex. *Nature*, **519**, 321–326.
11. Tocilj,A., On,K.F., Yuan,Z., Sun,J., Elkayam,E., Li,H., Stillman,B. and Joshua-Tor,L. (2017) Structure of the active form of human origin recognition complex and its ATPase motor module. *Elife*, **6**, e20818.
12. Li,N., Lam,W.H., Zhai,Y., Cheng,J., Cheng,E., Zhao,Y., Gao,N. and Tye,B.K. (2018) Structure of the origin recognition complex bound to DNA replication origin. *Nature*, **559**, 217–222.
13. Kawakami,H., Ohashi,E., Kanamoto,S., Tsurimoto,T. and Katayama,T. (2015) Specific binding of eukaryotic ORC to DNA replication origins depends on highly conserved basic residues. *Sci. Rep.*, **5**, 14929.
14. Giraldo,R. (2003) Common domains in the initiators of DNA replication in Bacteria, Archaea and Eukarya: combined structural, functional and phylogenetic perspectives. *FEMS Microbiol. Rev.*, **26**, 533–554.
15. Giraldo,R., Fernandez-Tornero,C., Evans,P.R., Diaz-Orejas,R. and Romero,A. (2003) A conformational switch between transcriptional repression and replication initiation in the RepA dimerization domain. *Nat. Struct. Biol.*, **10**, 565–571.
16. Nakamura,A., Wada,C. and Miki,K. (2007) Structural basis for regulation of bifunctional roles in replication initiator protein. *Proc. Natl. Acad. Sci. USA*, **104**, 18484–18489.
17. Pierechod,M., Nowak,A., Saari,A., Purta,E., Bujnicki,J.M. and Konieczny,I. (2009) Conformation of a plasmid replication initiator protein affects its proteolysis by ClpXP system. *Protein Sci.*, **18**, 637–649.
18. Swan,M.K., Bastia,D. and Davies,C. (2006) Crystal structure of pi initiator protein-iteron complex of plasmid R6K: implications for initiation of plasmid DNA replication. *Proc. Natl. Acad. Sci. USA*, **103**, 18481–18486.
19. Cereghino,J.L., Helinski,D.R. and Toukdarian,A.E. (1994) Isolation and characterization of DNA-binding mutants of a plasmid replication initiation protein utilizing an in vivo binding assay. *Plasmid*, **31**, 89–99.
20. Lin,J. and Helinski,D.R. (1992) Analysis of mutations in trfA, the replication initiation gene of the broad-host-range plasmid RK2. *J. Bacteriol.*, **174**, 4110–4119.
21. Kolatka,K., Kubik,S., Rajewska,M. and Konieczny,I. (2010) Replication and partitioning of the broad-host-range plasmid RK2. *Plasmid*, **64**, 119–134.
22. Konieczny,I. (2003) Strategies for helicase recruitment and loading in bacteria. *EMBO Rep.*, **4**, 37–41.
23. Perri,S. and Helinski,D.R. (1993) DNA-Sequence Requirements for Interaction of the Rk2 Replication Initiation Protein with Plasmid Origin Repeats. *J. Biol. Chem.*, **268**, 3662–3669.
24. Wegrzyn,K., Fuentes-Perez,M.E., Bury,K., Rajewska,M., Moreno-Herrero,F. and Konieczny,I. (2014) Sequence-specific interactions of Rep proteins with ssDNA in the AT-rich region of the plasmid replication origin. *Nucleic Acids Res.*, **42**, 7807–7818.
25. Durland,R.H., Toukdarian,A., Fang,F. and Helinski,D.R. (1990) Mutations in the trfA replication gene of the broad-host-range plasmid RK2 result in elevated plasmid copy numbers. *J. Bacteriol.*, **172**, 3859–3867.
26. Blasina,A., Kittell,B.L., Toukdarian,A.E. and Helinski,D.R. (1996) Copy-up mutants of the plasmid RK2 replication initiation protein are defective in coupling RK2 replication origins. *Proc. Natl. Acad. Sci. USA*, **93**, 3559–3564.
27. Gasset-Rosa,F., Diaz-Lopez,T., Lurz,R., Prieto,A., Fernandez-Tresguerres,M.E. and Giraldo,R. (2008) Negative regulation of pPS10 plasmid replication: origin pairing by zipping-up DNA-bound RepA monomers. *Mol. Microbiol.*, **68**, 560–572.
28. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
29. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W.Z., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
30. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
31. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
32. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
33. Le,S.Q. and Gascuel,O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.
34. Darriba,D., Taboada,G.L., Doallo,R. and Posada,D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**, 1164–1165.
35. Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
36. Berjanskii,M., Liang,Y.J., Zhou,J.J., Tang,P., Stothard,P., Zhou,Y., Cruz,J., MacDonell,C., Lin,G.H., Lu,P. *et al.* (2010) PROSESS: a protein structure evaluation suite and server. *Nucleic Acids Res.*, **38**, W633–W640.
37. Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**, 40.
38. Pierce,B.G., Wiehe,K., Hwang,H., Kim,B.H., Vreven,T. and Weng,Z.P. (2014) ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*, **30**, 1771–1773.
39. Li,S., Olson,W.K. and Lu,X.J. (2019) Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures. *Nucleic Acids Res.*, **47**, W26–W34.
40. Dans,P.D., Perez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
41. Eswar,N., Webb,B., Marti-Renom,M.A., Madhusudhan,M.S., Eramian,D., Shen,M.Y., Pieper,U. and Sali,A. (2006) Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinform.*, doi:10.1002/0471250953.bi0506s15.
42. Terakawa,T., Kameda,T. and Takada,S. (2011) On Easy Implementation of a Variant of the Replica Exchange with Solute Tempering in GROMACS. *J. Comput. Chem.*, **32**, 1228–1234.

43. Domene,C., Barbini,P. and Furini,S. (2015) Bias-Exchange Metadynamics Simulations: An Efficient Strategy for the Analysis of Conduction and Selectivity in Ion Channels. *J. Chem. Theory Comput.*, **11**, 1896–1906.
44. Chen,V.B., Arendall,W.B., Headd,J.J., Keedy,D.A., Immormino,R.M., Kapral,G.J., Murray,L.W., Richardson,J.S. and Richardson,D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D*, **66**, 12–21.
45. Abraham,M.S., Murtola,T., Schulz,R., Palle,S., Smith.,J.C., Hess,B. and Lindahl,E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1**, 19–25.
46. Hornak,V., Abel,R., Okur,A., Strockbine,B., Roitberg,A. and Simmerling,C. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, **65**, 712–725.
47. Lindorff-Larsen,K., Piana,S., Palmo,K., Maragakis,P., Klepeis,J.L., Dror,R.O. and Shaw,D.E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, **78**, 1950–1958.
48. Ivani,I., Dans,P.D., Noy,A., Perez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goni,R., Balaceanu,A. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
49. Tribello,G.A., Bonomi,M., Branduardi,D., Camilloni,C. and Bussi,G. (2014) PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.*, **185**, 604–613.
50. Pires,D.E.V. and Ascher,D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic Acids Res.*, **45**, W241–W246.
51. Meyer,R. (2009) Replication and conjugative mobilization of broad host-range IncQ plasmids. *Plasmid*, **62**, 57–70.
52. Rawlings,D.E. and Tietze,E. (2001) Comparative biology of IncQ and IncQ-like plasmids. *Microbiol. Mol. Biol. Rev.*, **65**, 481–496.
53. Carlsson,J., Drevin,H. and Axen,R. (1978) Protein thiolation and reversible protein-protein conjugation. N-Succinimidyl 3-(2-pyridyldithio)propionate, a new heterobifunctional reagent. *Biochem. J.*, **173**, 723–737.
54. Chodosh,L.A. (2001) UV crosslinking of proteins to nucleic acids. *Curr. Protoc. Mol. Biol.*, **Chapter 12**, Unit 12.5.
55. Meisenheimer,K.M. and Koch,T.H. (1997) Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. Biol.*, **32**, 101–140.
56. Yano,H., Wegrzyn,K., Loftie-Eaton,W., Johnson,J., Deckert,G.E., Rogers,L.M., Konieczny,I. and Top,E.M. (2016) Evolved plasmid-host interactions reduce plasmid interference cost. *Mol. Microbiol.*, **101**, 743–756.
57. Kongsuwan,K., Josh,P., Picault,M.J., Wijffels,G. and Dalrymple,B. (2006) The plasmid RK2 replication initiator protein (TrfA) binds to the sliding clamp beta subunit of DNA polymerase III: implication for the toxicity of a peptide derived from the amino-terminal portion of 33-kilodalton TrfA. *J. Bacteriol.*, **188**, 5501–5509.
58. Wawrzycka,A., Gross,M., Wasaznik,A. and Konieczny,I. (2015) Plasmid replication initiator interactions with origin 13-mers and polymerase subunits contribute to strand-specific replisome assembly. *Proc. Natl. Acad. Sci. USA*, **112**, E4188–E4196.
59. Jiang,Y., Pacek,M., Helinski,D.R., Konieczny,I. and Toukdarian,A. (2003) A multifunctional plasmid-encoded replication initiation protein both recruits and positions an active helicase at the replication origin. *Proc. Natl. Acad. Sci. USA*, **100**, 8692–8697.
60. Orlova,N., Gerding,M., Ivashkiv,O., Olinares,P.D.B., Chait,B.T., Waldor,M.K. and Jeruzalmi,D. (2017) The replication initiator of the cholera pathogen's second chromosome shows structural similarity to plasmid initiators. *Nucleic Acids Res.*, **45**, 3724–3737.
61. Giraldo,R. and Diaz-Orejas,R. (2001) Similarities between the DNA replication initiators of Gram-negative bacteria plasmids (RepA) and eukaryotes (Orc4p)/archaea (Cdc6p). *Proc. Natl. Acad. Sci. USA*, **98**, 4938–4943.
62. Deng,L., Zhong,G., Liu,C., Luo,J. and Liu,H. (2019) MADOKA: an ultra-fast approach for large-scale protein structure similarity searching. *BMC Bioinformatics*, **20**, 662.
63. Fu,M., Wang,C., Zhang,X. and Pestell,R.G. (2004) Acetylation of nuclear receptors in cellular growth and apoptosis. *Biochem. Pharmacol.*, **68**, 1199–1208.
64. Wang,J., Liu,N., Liu,Z., Li,Y., Song,C., Yuan,H., Li,Y.Y., Zhao,X. and Lu,H. (2008) The orphan nuclear receptor Rev-erbbeta recruits Tip60 and HDAC1 to regulate apolipoprotein CIII promoter. *Biochim. Biophys. Acta*, **1783**, 224–236.
65. Carabetta,V.J. and Cristea,I.M. (2017) Regulation, function, and detection of protein acetylation in bacteria. *J. Bacteriol.*, **199**, e00107-17.
66. Zhang,Q., Zhou,A., Li,S., Ni,J., Tao,J., Lu,J., Wan,B., Li,S., Zhang,J., Zhao,S. *et al.* (2016) Reversible lysine acetylation is involved in DNA replication initiation by regulating activities of initiator DnaA in Escherichia coli. *Sci. Rep.*, **6**, 30837.
67. Giraldo,R., Andreu,J.M. and Diaz-Orejas,R. (1998) Protein domains and conformational changes in the activation of RepA, a DNA replication initiator. *EMBO J.*, **17**, 4511–4526.
68. Diaz-Lopez,T., Lages-Gonzalo,M., Serrano-Lopez,A., Alfonso,C., Rivas,G., Diaz-Orejas,R. and Giraldo,R. (2003) Structural changes in RepA, a plasmid replication initiator, upon binding to origin DNA. *J. Biol. Chem.*, **278**, 18606–18616.
69. Toukdarian,A.E., Helinski,D.R. and Perri,S. (1996) The plasmid RK2 initiation protein binds to the origin of replication as a monomer. *J. Biol. Chem.*, **271**, 7072–7078.
70. Zhong,Z., Helinski,D. and Toukdarian,A. (2005) Plasmid host-range: restrictions to F replication in Pseudomonas. *Plasmid*, **54**, 48–56.
71. Bowers,L.M., Kruger,R. and Filutowicz,M. (2007) Mechanism of origin activation by monomers of R6K-encoded pi protein. *J. Mol. Biol.*, **368**, 928–938.
72. Xia,G.X., Manen,D., Yu,Y.Y. and Caro,L. (1993) In-Vivo and in-Vitro Studies of a Copy Number Mutation of the Repa Replication Protein of Plasmid-Psc101. *J. Bacteriol.*, **175**, 4165–4175.
73. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.