© 1998 Oxford University Press
Nucleic Acids Research, 1998, Vol. 26, No. 18 **4205–4213**

Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins

L. Aravind[1,2], Detlef D. Leipe[2] and Eugene V. Koonin[2,*]

[1]Department of Biology, Texas A&M University, College Station, TX 70843, USA and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received May 26, 1998; Revised and Accepted July 23, 1998

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022

ABSTRACT

Iterative profile searches and structural modeling show that bacterial DnaG-type primases, small primase-like proteins from bacteria and archaea, type IA and type II topoisomerases, bacterial and archaeal nucleases of the OLD family and bacterial DNA repair proteins of the RecR/M family contain a common domain, designated Toprim (topoisomerase-primase) domain. The domain consists of ~100 amino acids and has two conserved motifs, one of which centers at a conserved glutamate and the other one at two conserved aspartates (DxD). Examination of the structure of Topo IA and Topo II and modeling of the Toprim domains of the primases reveal a compact β/α fold, with the conserved negatively charged residues juxtaposed, and inserts seen in Topo IA and Topo II. The conserved glutamate may act as a general base in nucleotide polymerization by primases and in strand rejoining by topoisomerases and as a general acid in strand cleavage by topoisomerases and nucleases. The role of this glutamate in catalysis is supported by site-directed mutagenesis data on primases and Topo IA. The DxD motif may coordinate Mg^{2+} that is required for the activity of all Toprim-containing enzymes. The common ancestor of all life forms could encode a prototype Toprim enzyme that might have had both nucleotidyl transferase and polynucleotide cleaving activity.

INTRODUCTION

DNA replication is executed by a complex protein machinery in all cells (1). In addition to the replicative DNA polymerase, which is involved in catalyzing template-dependent nucleotide polymerization, there are a number of enzymes and protein assemblages that function prior to, in the course of and after polymerization. These include the protein complexes that melt the double-stranded (ds)DNA at the origin of replication in an ATP-dependent reaction and helicases that unwind DNA during replication. In spite of their diversity in the three domains of life and in viruses, all DNA polymerases are characterized by their inability to initiate polymerization *de novo*. They all require a primer molecule to provide a 3′-OH onto which the first nucleotide is added (1). There are three basic strategies for obtaining the primer: (i) by using a protein hydroxyl group, as in the case of viral and plasmid protein-primed DNA polymerases (2); (ii) by introducing a nick into a circular dsDNA molecule, as in the rolling circle replication model typical of many plasmids and bacteriophages, or into a covalently closed terminal hairpin, as observed in poxviruses or parvoviruses (3,4); (iii) by means of a specialized RNA polymerase, called a primase, which synthesizes a short RNA molecule in a template-dependent manner and provides the DNA polymerase with a 3′-OH to continue chain elongation (5). The latter mechanism is used predominantly, if not exclusively, in replication of the genomic DNA in prokaryotic and eukaryotic cells. Genes for primases have been identified in the genomes of bacteria, archaea, eukaryotes and several viruses. The primase family that has been studied in most detail includes bacterial proteins typified by *Escherichia coli* DnaG and their bacteriophage homologs (5–7). Recently, homologs of the DnaG family primases have been identified by computer analysis of proteins encoded in archaeal genomes (8). In addition, archaea encode apparent orthologs of the two subunits of the eukaryotic primase that are also conserved in baculoviruses (9; L.Aravind, unpublished observations). Finally, herpesviruses encode a unique primase whose sequence seems to be unrelated to those of the other primase families (10).

The closed circular chromosomes of prokaryotes and the loops of the linear chromosomes of eukaryotes pose fundamental topological problems of changing linking and writhing numbers in the course of DNA replication, repair and transcription. This is performed by topoisomerases that catalyze the complex reaction of breakage of one or both strands of DNA, passage of another strand or two through this break and finally sealing of the original break (11). Type I topoisomerases are ATP-independent and catalyze relaxation of negatively supercoiled DNA, interconversion of knotted and unknotted DNA and DNA ring concatenation. Bacterial type I topoisomerases (Topo IA), typified by *E.coli* ω protein, the first topoisomerase to be discovered (12), form a family with the eukaryotic topoisomerase III (Topo III) and the reverse gyrases found in archaea and thermophilic bacteria (10).

*To whom correspondence should be addressed. Tel: +1 301 435 5913; Fax: +1 301 480 9241; Email: koonin@ncbi.nlm.nih.gov

The authors wish it to be known that in their opinion, the first two authors should be regarded as joint First Authors

In contrast, eukaryotic and some viral type I topoisomerases (Topo IB), while catalyzing the same topological modifications of DNA as Topo I and Topo III, are unrelated to them in sequence (13). Reverse gyrases contain a helicase domain, in addition to the Topo IA domain, and catalyze the formation of positive supercoils in DNA (14,15).

Type II topoisomerases (Topo II), including bacterial gyrases, introduce negative supercoils in DNA (16). They require ATP for strand translocation and contain an ATPase domain or subunit related to MutL, Hsp90 and histidine kinases and a distinct domain or subunit for DNA binding, cleavage and rejoining (14,17). Traditionally, Topo II enzymes have been considered to be unrelated to Topo IA or Topo IB. However, the annotation of the respective structures in the SCOP database indicates that Topo I and Topo II have structurally similar catalytic domains (18,19). The similarity between these domains was also confirmed by a structural alignment included in the FSSP database (20). To our knowledge, however, the functional and evolutionary implications of this relationship have so far not been considered.

Here we show that Topo IA and Topo II share a structurally conserved domain involved in DNA strand breakage and rejoining not only with one another, but also with the DnaG-type primases, a family of ATP-dependent nucleases and a family of DNA repair proteins. These observations suggest a previously unsuspected, deep mechanistic analogy between such superficially different processes as primer formation, DNA breakage and rejoining by topoisomerase and DNA cleavage by certain nucleases. We hypothesize that at a very early stage of evolution, topoisomerases and primases could have evolved from a single ancestral enzyme that might have had multiple functions in replication and repair.

MATERIALS AND METHODS

The databases used in this study were the non-redundant database (NR) and the protein sequences encoded in complete genomes from the GenBank genomes division. Sequence analysis was performed using the SEALS package, which integrates database search programs and bulk sequence data handling into simple command line options (21). The basic strategy involved iterative database searches with the PSI-BLAST program (22) using multiple starting query sequences. The program constructs a position-dependent weight matrix from multiple alignments generated from the BLAST hits above a certain expectation value (*e* value) and carries out iterative database searches using the information derived from this matrix (22). The program also allows generation of 'checkpoint' matrices with fixed *e* value cut-offs and number of iterations that can be used in searches of new databases such as complete genomes or in subsequent searches with altered *e* value cut-offs (A.Schaffer, L.Aravind and E.V.Koonin, unpublished results). The statistical evaluation of the PSI-BLAST results is based on the extreme value distribution statistics originally developed by Karlin and Altschul for local alignments without gaps (23) and subsequently modified for gapped alignments (22,24). While there is no analytical proof of the applicability of the Karlin–Altschul statistics to searches that use position-dependent matrices as queries, extensive computer simulations showed an excellent fit of the distribution of score obtained in such searches to the extreme value distribution (22). Therefore, *e* values reported for each retrieved sequence at the point when its alignment with the query exceeds the cut-off for the first time appear to be robust estimates of statistical significance; evidently, once a sequence gets included in the model, *e* values reported for it (and its closely related homologs) at subsequent iterations become inflated and do not accurately represent the statistical significance. All *e* values reported here are for the first appearance of the given sequences above the cut-off.

The motif searching program MoST (25) was used as an independent test of the results produced by PSI-BLAST. MoST performs an iterative search of the database using as the query a position-specific weight matrix derived from a multiple alignment block without gaps. The results are evaluated in terms of the ratio of the expected number of sequence segments with a given score to the actually observed number (*r* value).

The likelihood of an alignment of two sequences being indicative of a structural similarity was determined using the ZEGA program (26). Under this method, the probability that a given alignment score is observed in the absence of a structural relationship is calculated using an analytical function derived from the distribution of alignment scores for sequences with the same structural fold and those with different folds. The alignments are constructed using a modification of the Needleman–Wunsch algorithm (27) with zero end gap penalties.

The principal cause of erroneous results in database searches is the presence in many protein sequences of compositionally biased (low complexity) regions that tend to produce spurious alignments with database sequences that have a similar bias (28,29). This effect may be particularly deleterious for iterative database search methods due to the amplification of errors. In order to prevent such artifacts, low complexity regions in the query sequences were masked using the SEG program (30,31) and coiled coil regions (a special case of compositional bias) were masked using the Lupas coil detection method (32,33). The SEG program was applied with two sets of parameters, namely the standard ones used by default with the BLAST family programs {window length (*W*) 12, trigger complexity [$K_2(1)$] 2.2, extension complexity [$K_2(2)$] 2.5} and the parameters adjusted to delineate non-globular domains in proteins [$W = 45$, $K_2(1) = 3.4$, $K_2(2) = 3.75$].

Sequence alignments were constructed using the Gibbs sampling procedure as implemented in the MGIBBS (34) and MACAW (35) programs and adjusted on the basis of the PSI-BLAST results and structure prediction. The alignments were formatted using the SEAVIEW (36) and ALSCRIPT programs (37). Amino acid pattern searches were performed using the GREF program of the SEALS package (21). Phylogenetic trees were constructed using the PAUP 3 software (38). Protein secondary structure prediction on the basis of a multiple sequence alignment was carried out using the PHD program (39) and subsequent secondary structure-based threading of the PDB database was carried out using the threading option of the PHD program (40). Protein databank (PDB) files were visualized using SWISS-PDB viewer v.2.6. Homology modeling of protein structures was performed using the ProMod program (41,42).

RESULTS AND DISCUSSION

The topoisomerase-primase (Toprim) domain—delineation of the superfamily

When the *E.coli* DnaG primase sequence, with masked low complexity regions, was used as a query in a PSI-BLAST search, uncharacterized proteins from the four archaea with completely

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022

sequenced genomes and *Sulfolobus* were recovered in the second iteration with *e* values $<10^{-3}$. Thus, the DnaG homolog previously detected in *Methanococcus jannaschii* (8) belongs to a protein family that is highly conserved in all archaea. As archaea are believed to have a eukaryote-type DNA replication system (43) and encode homologs of both subunits of the eukaryotic polymerase α type primase (8,44; L.Aravind, unpublished observations), the presence of the bacterial type primases was of interest and prompted a further investigation by means of more detailed sequence analysis. This resulted in the identification of a widespread family of bacterial and archaeal proteins that share a conserved domain with the DnaG-type bacterial primases. This domain encompassed the previously described motifs IV and V (6), which are the signatures of bacterial and phage primases and are characterized, respectively, by an invariant glutamate and an aspartate dyad (DxD) motif, both preceded by conserved hydrophobic regions predicted to form β-strands. The conserved DxD in motif V resembles similar motifs in the catalytic sites of other nucleic acid polymerases, including the large subunits of eukaryotic primases and herpesvirus primases (10,45). Site-directed mutagenesis results indicate that the conserved domain that encompasses motifs IV and V is directly involved in primer synthesis by the DnaG-type primases (46).

A subset of the primase superfamily proteins from bacteria and archaea that showed highly significant sequence similarity to DnaG (*e* values $<10^{-5}$ within two to four iterations) consist of only ~120 amino acid residues and are almost entirely made up of the conserved catalytic domain (Figs 1 and 2). As these proteins do not contain additional domains, particularly non-globular ones, they seemed to be most suitable queries for PSI-BLAST searches aimed at detection of possible distant homologs of the core primase domain. Indeed, searches with two of these small primase-related proteins, namely YusF and YabF from *Bacillus subtilis*, not only confirmed their relationship with the DnaG-type primases, but also recovered, at a statistically significant level (*e* values ~10^{-3} in the third iteration), two topoisomerase families, namely Topo IA and Topo II, and the RecR family proteins. In addition, these searches retrieved, albeit with low significance (*e* values ~0.02), the ATP-dependent nucleases of a family typified by bacteriophage P2 OLD protein. Subsequent searches from other starting points recovered all members of the superfamily, including the OLD family, at statistically significant *e* values. An additional analysis performed using the MoST program showed that a block of the alignment of the DnaG-type family (including newly detected bacterial and archaeal homologs) including motif V retrieved from the database the two families of topoisomerases without any false positives when the restrictive cut-off of $r = 0.001$ was used. In order to evaluate the fold prediction for the primases and other members of the emerging superfamily, pairwise alignments of different members with the two topoisomerases of known structure were constructed and the probability that the respective proteins of unknown structure adopt the same fold was computed using the ZEGA program (26). The hypothesis that primases, RecR and OLD nucleases have the same fold as the Topo IA and Topo II domains was supported with $P < 10^{-5}$.

The relationship between the primases, the two families of topoisomerases and nucleases was surprising, since these enzymes catalyze very different reactions, albeit on the same substrate, namely dsDNA. Nevertheless, several lines of evidence, in addition to the statistical support provided by the PSI-BLAST, MoST and ZEGA results, corroborate the structural and functional relevance of the observed similarity. Examination of the multiple alignment of the conserved regions shows that the DnaG-type primase motifs IV and V and, in particular, the invariant glutamate in motif IV and the DxD in motif V are conserved in Topo IA, Topo II and the OLD nucleases, with deviations found only in the RecR protein family (Fig. 1). Furthermore, several amino acid residues are conserved outside these principal motifs and a third motif, located between motif IV and V and containing a nearly invariant glycine, was identified (Fig. 1). The boundaries of the conserved region shown in Figure 1 correspond almost precisely to the N-terminal globular domain of Topo IA (45) and to the distinct central domain in the non-ATP-binding part of Topo II (16). Multiple alignment-based secondary structure prediction (39) for the DnaG family followed by secondary structure-based PDB database threading (40) produced a top *Z* score of 2.9 with the yeast Topo II structure (1bgw); threading scores in this range strongly suggest structural similarity (40), thus supporting the prediction of a topoisomerase-like fold in the primases. Taken together, these findings indicate that DnaG-type primases, Topo IA, Topo II, RecR and OLD family nucleases contain a structurally conserved domain that we designated the Toprim domain.

The sequences of other known primases and topoisomerases were analyzed for possible distant similarity to the Toprim domain. While no indication of such a relationship could be detected in eukaryotic and herpesvirus primases or in Topo IB, the non-ATPase subunits of the recently identified archaeal and eukaryotic Topo VI (14) contain possible counterparts of the three conserved motifs (Fig. 1). Secondary structure prediction and evaluation of the probability of fold similarity to 1ecl or 1bgw using the ZEGA program (P ~10^{-3}–10^{-4}) suggested structural similarity to the Toprim domain. Thus, Topo VI may contain a highly diverged version of the Toprim domain.

Structure of the Toprim domain and its function in catalysis

A model of the Toprim domain structure in the DnaG-type primase was constructed using multiple alignment of the Toprim domain (Fig. 1) and the experimentally determined structures of the respective domains from Topo IA (1ecl) and Topo II (1bgw) as templates. The Toprim domain has an α/β fold with four conserved strands and three helices (Fig. 3); with the exception of the second helix and the C-terminal strand, each of these elements contains positions that are highly conserved in the Toprim domain alignment (Fig. 1). The Toprim domain contains three regions that can accommodate variable sized inserts, which are particularly prominent in the topoisomerases (Fig. 3A and B). In Topo IA, the long insert 2, which possibly interacts with the helical connective region linking the Toprim domain to the saddle-shaped β-sheet, is partially disordered in the crystal structure and in the phage T4 topoisomerase, the Toprim domain is partitioned into two separate gene products in the region of this insert. Signifcant size variations are also observed in the loop connecting the C-terminal strand to the preceding helix (Fig. 3A). The invariant glutamate is located in a sharp turn that connects the first strand to the first helix. A structurally similar β/α element in the C-terminal region of the Toprim domain contains the DxD motif (Figs 1 and 3). The three conserved acidic residues are juxtaposed in space (Fig. 3). In the case of Topo IA, it has been noticed that the DxD motif resembles the Mg^{2+}-binding site of the

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022

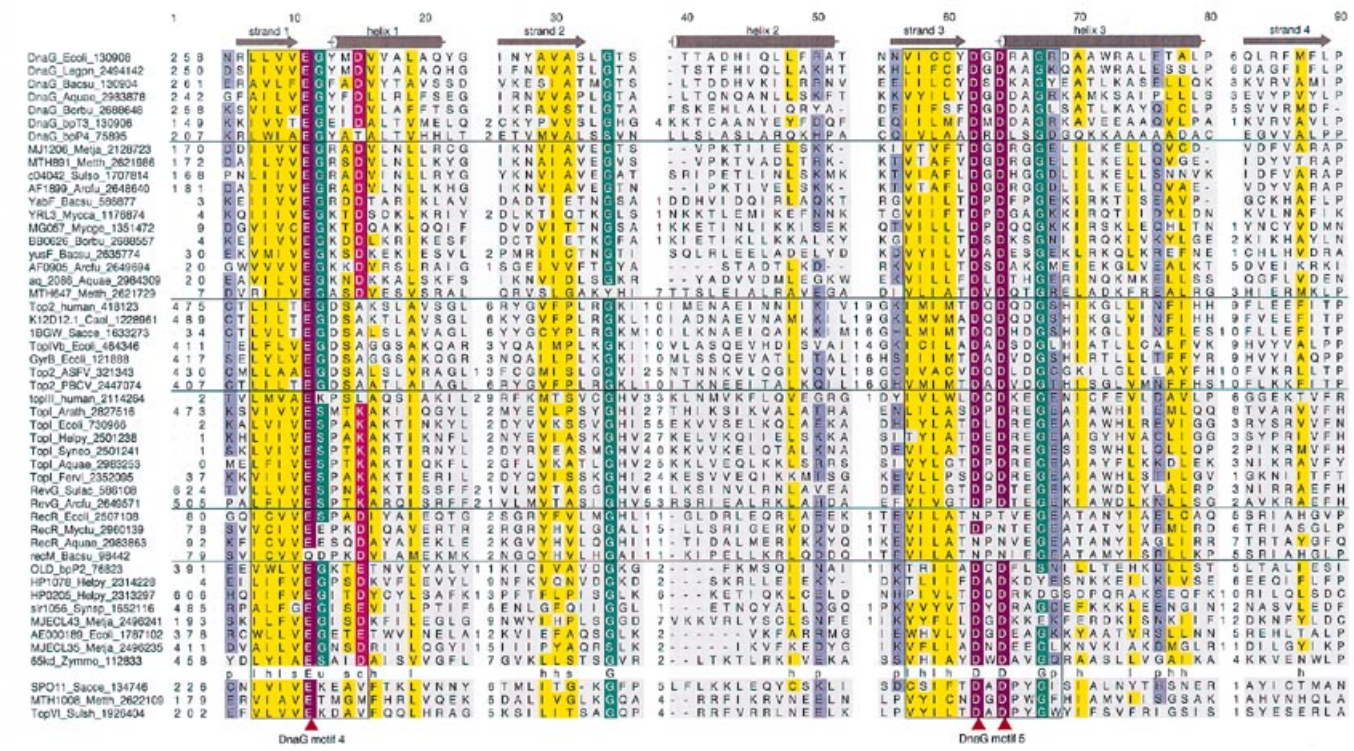

Figure 1. Multiple alignment of the Toprim domain in DnaG-type primases, topoisomerases and related proteins. From top to bottom (separated by horizontal lines) the alignment contains sequences from: (1) bacterial and phage DnaG primases; (2) DnaG-related proteins from bacteria and archaea; (3) Topo II family; (4) Topo IA family; (5) RecR family of repair and recombination proteins; (6) OLD-related proteins. The 80% consensus for these proteins is shown below the aligned sequences. The Toprim-like domains from archaeal Topo VI A subunits and the eukaryotic SPO11-like proteins are shown below the consensus line. The conserved acidic residues (see text) are indicated by red arrowheads. Numbers indicate the distance to the N-terminal methionine and the residues between alignment blocks. The color coding for residues that are conserved in at least 80% of the aligned sequences is: purple for negatively charged (D and E); pink for charged (D, E, H, K and R); green for tiny (G, A and S); yellow for hydrophobic (A, C, F, I, L, M, V, W and Y) or aliphatic (I, L and V); turquoise for small (A, C, D, G, N, P, S, T and V); blue-grey for polar (C, D, E, H, K, N, Q, R, S and T) residues. The secondary structure elements derived from the X-ray structures of Topo IA and Topo II and accordingly predicted for the rest of the superfamily are shown above the alignment. The protein identifiers are composed of a systematic gene name (e.g. DnaG) or a PDB abbreviation (1BGW), followed by the five letter source organism name abbreviation, followed by the Gene Identification no. Source organism abbreviations: Ecoli, *Escherichia coli*; Legpn, *Legionella pneumophila*; Bacsu, *Bacillus subtilis*; Aquae, *Aquifex aeolicus*; Borbu, *Borrelia burgdorferi*; bpT3, bacteriophage T3; Metja, *Methanococcus jannaschii*; Metth, *Methanobacterium thermoautotrophicum*; Sulso, *Sulfolobus solfataricus*; Arcfu, *Archaeoglobus fulgidus*; Mycca, *Mycoplasma capricolum*, Mycge, *Mycoplasma genitalium*; human, *Homo sapiens*; Caeel, *Caenorhabditis elegans*; Sacce, *Saccharomyces cerevisiae*; ASFV, African swine fever virus; PBCV-1, *Paramecium bursaria* Chlorella virus 1; Helpy, *Helicobacter pylori*; Syneo, *Synechocystis* PCC6803; Feris, *Fervidobacterium islandicum*; Arath, *Arabidopsis thaliana*; Sulac, *Sulfolobus acidocaldarius*; Myctu, *Mycobacterium tuberculosis*; bpP2, bacteriophage P2; Zymmo, *Zymomonas mobilis*; Sulsh, *Sulfolobus shibatae*.

Klenow fragment of DNA polymerase I (47). Given the present identification of the relationship with DnaG-type primases, this general similarity to the active sites of many nucleic acid polymerases, including eukaryotic, archaeal and herpesvirus primases (6,10), may be functionally relevant. Consistent with this, primases, Topo IA, reverse gyrases and Topo II have all been shown to require Mg^{2+} for their activity (11,48).

Mutations of the conserved glutamate completely abolished the activity of DnaG-type primases in the polymerization reaction, even in the case of a E→Q substitution (46). The topoisomerase reaction is distinct from polymerization and involves serial breakage and rejoining of the polynucleotide chain(s), with an intermediate step in which the DNA fragment is covalently bound to the active tyrosine in the enzyme (49). This reaction requires a general acid for breakage, to donate a proton to the sugar hydroxyl, and a general base to abstract the proton during the rejoining reaction (49). Substitution of alanine for the conserved glutamate (E9) of the Toprim domain in *E.coli* Topo IA abolished both strand cleavage and rejoining, suggesting that this residue is indeed critical to the reaction mechanism, possibly playing the dual role of a general acid and a general base (49). This strikingly similar result in terms of the necessity of the conserved glutamate for activity of both the primase and the topoisomerase suggests a common reaction mechanism. In the primases, the conserved glutamate most likely functions as a general base which abstracts the proton from the 3′-OH of the growing chain, similarly to its proposed role in the strand rejoining step of the topoisomerase reaction. The resulting 3′-O^- may then attack the incoming 5′-NTP, resulting in chain elongation. Subtle differences in the primase and topoisomerase reaction mechanisms are suggested, however, by the finding that a E9Q mutant of Topo IA is active in both the cleavage and the rejoining assays (49).

Topo IA requires Mg^{2+} for relaxation of supercoils but not for strand cleavage (11). Mutation of each of the three conserved acidic residues (E and DxD) partially inhibited the relaxation activity, Mg^{2+} binding and DNA binding by *E.coli* Topo IA, though D111 (the proximal aspartate of the DXD motif) mutations had only a mild effect (48). In the crystal structure, the

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022

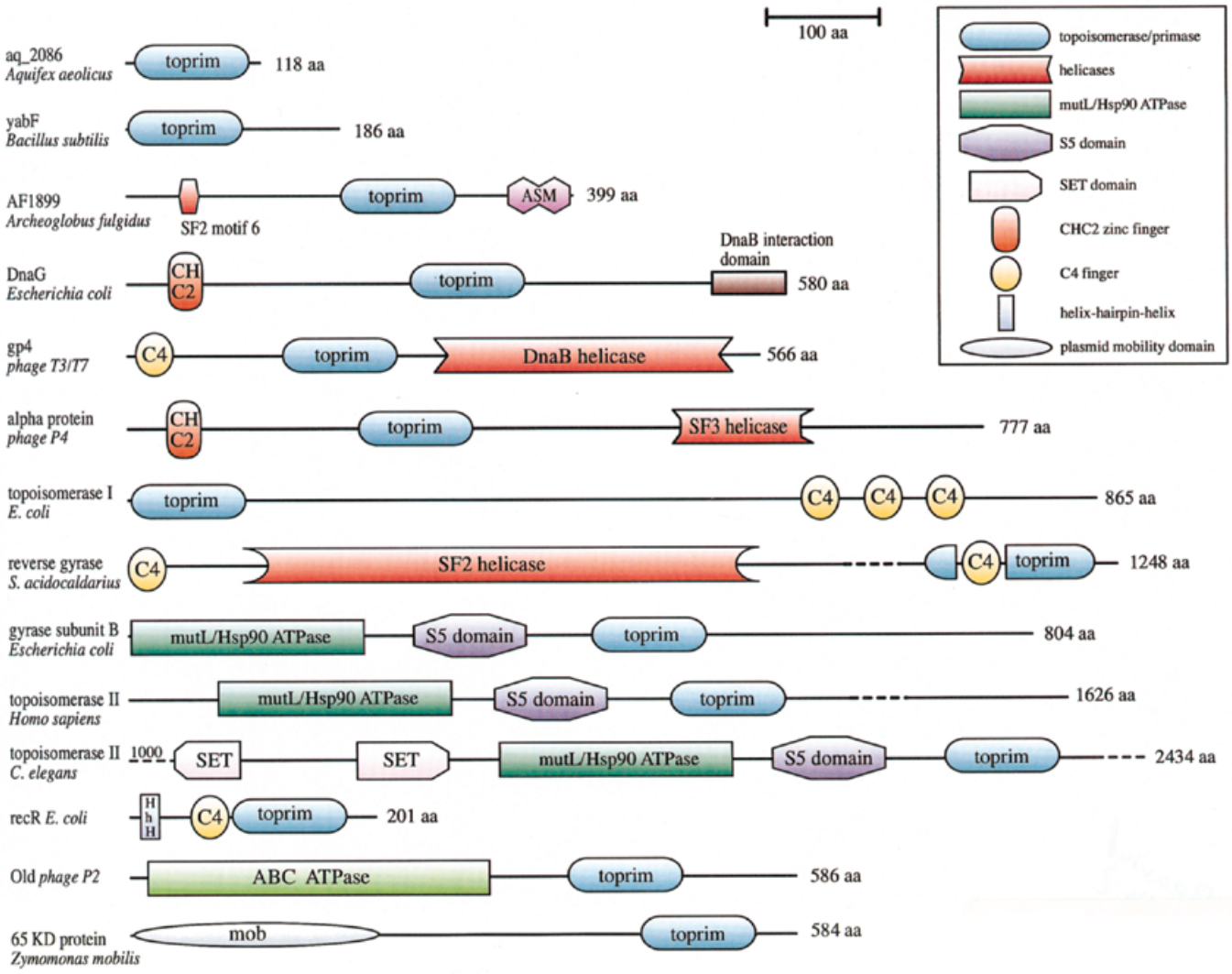

Figure 2. Domain architectures of the Toprim domain-containing proteins. Domain designations: C4, the 'little finger' domain, a small, widespread nucleic acid-binding finger that appears to be structurally distinct from other C4 fingers (L.Aravind, unpublished observations); CHC2, a distinct zinc chelating domain found in the phage P4 and cellular primases; ASM, a highly conserved motif found only in the large archaeal DnaG-like proteins; SF2, superfamily 2 (II) helicase domain; SF3, superfamily 3 (AAA ATPase-like) helicase domain; MutL/Hsp90, ATPase domain of the HSP90–gyrase–histidine kinase superfamily; ABC ATPase, ATPase domain of the ABC transporter/SMC superfamily; DnaB, DnaB family helicase; HhH, helix–hairpin–helix DNA-binding motif; MOB, plasmid mobility protein, DNA strand-nicking domain; S5, a putative nucleic acid-binding domain shared with EF-G and ribosomal protein S5 in Topo II; SET, a domain found in chromatin-associated proteins (the SET domain in *C.elegans* Topo II has a long insert of ~100 residues).

conserved acidic residues of Topo IA belong to a network of hydrogen bonds and salt bridges around the active site, supporting the notion emerging from the mutagenesis studies, namely that in addition to the catalytic role proposed for the conserved glutamate, these residues are involved in DNA association, metal coordination and DNA-dependent conformational changes (45,49). This is consistent with the detection of conformational changes in the primases in the presence of Mg^{2+} (50).

Given the role of the conserved glutamate in both strand breakage and nucleotidyl transferase activity, it seemed likely that proteins containing the Toprim domain may act both as nucleotidyl transferases and as nucleases. In accord with this notion, the Toprim domain was indeed identified in the OLD family of nucleases (Fig. 1) and it can be predicted that the catalytic site of these nucleases includes the conserved residues of the Toprim domain.

Identification of the Toprim domain further helps in understanding the reaction mechanisms of Topo II. Traditionally, ATP-dependent and -independent topoisomerases have been considered to be very different and determination of the crystal structures did not fully clarify their relationships. However, in accord with the structure–structure comparison results reported in the SCOP (18,19) and FSSP (20) databases, we found that these two types of topoisomerases have a conserved catalytic domain. Furthermore, conservation of the catalytic glutamate and DxD motifs suggests that, the role of ATP hydrolysis in Topo II activity notwithstanding, the basic reaction mechanism is likely to be the same. Specifically, the position of the Toprim domain in Topo II

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022

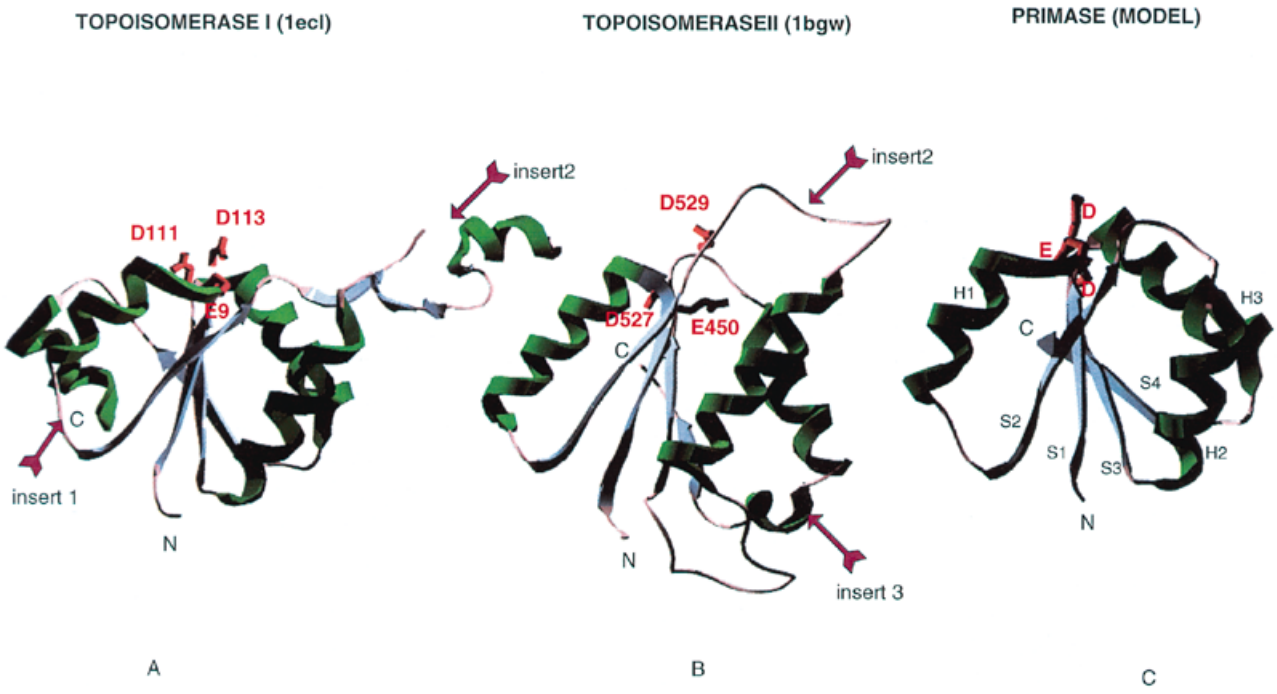

Figure 3. Structure of the Toprim domain in Topo IA and Topo II and a structural model of the core primase domain. The conserved acidic residues (see text) are indicated in each structure. (**A**) Structure of the Toprim domain in Topo IA from *E.coli* (1ecl) with the large, partially disordered insert in region 2. (**B**) Structure of the Toprim domain of Topo II from yeast (1bgw) with a relatively short (compare with A) loop inserted in region 2 and a large insert in region 3. (**C**) Structural model for the Toprim domain in the DnaG-type primases. Swiss PDB Viewer was used to thread the raw sequence through the structural alignment of 1ecl and 1bgw. The alignment was then manually adjusted to globally minimize the energy and remove clashes with the backbone. This structural alignment with the templates was submitted for homology modeling using the ProMod II program. Note the absence of large inserts typical of the topoisomerases in the primase-type Toprim domain. The predicted structural elements are indicated by letter: S, strand; H, helix.

is compatible with a direct interaction with the cleaved strand (16), making it likely that the two enzyme families use the same cleavage–rejoining cycle dependent on the conserved glutamate displaying alternately the properties of a general acid and a general base. Thus, taking into account the putative Toprim domain in Topo VI, the sequence and structural comparisons reported here seem to unify all known topoisomerases in terms of the origin, structure and catalytic mechanism of their cleavage–rejoining domains, with the single exception of Topo IB. There are notable differences in the reaction mechanisms of Topo IA and Topo II, on the one hand, and Topo IB, on the other, in that the former are covalently linked to the 5′-phosphoryl group of the cleaved DNA, whereas the latter are linked to the 3′-phosphoryl group. Furthermore, unlike Topo IA and Topo II, relaxation of supercoils by Topo IB does not require Mg^{2+} (11). For Topo IB, a distinct structural, functional and evolutionary relationship with site-specific recombinases has been recently demonstrated (13), confirming that this family is indeed unrelated to the Toprim domain-containing topoisomerases.

The remaining members of the Toprim domain superfamily are the RecR/M proteins involved in recombinational repair in bacteria (51). Interestingly, these proteins show substitutions in the DxD motif (e.g. D→N) and, in some cases, in the N-terminal motif (E→Q; Fig. 1), which indicates that they most likely lack catalytic activity. RecR indeed is a DNA-binding protein without any demonstrated enzymatic activity (51,52). Recently, it has been shown that RecR forms distinct complexes with RecF and RecO proteins, which limit the extension of the RecA filament to a single-strand gap in the process of recombinational repair (51,53). This suggests that RecR/M proteins may use their Toprim domain to associate with specific DNA structures that arise as a result of damage-induced replication fork stalling and resemble DNA features recognized by topoisomerases.

Domain architectures of Toprim domain-containing proteins and evolutionary implications

The Toprim domain is present in a wide range of proteins and combines with several other functional domains (Fig. 2 and Table 1). Proteins that consist of the Toprim domain alone are found in a number of bacteria and archaea (Table 1). These proteins invariably show greater similarity to the DnaG-type primases than to other members of the Toprim domain superfamily, but their function remains uncertain. As the small proteins lack the additional domains that are involved in numerous interactions typical of the larger proteins of the Toprim superfamily (primases and topoisomerases), they may represent a novel class of nucleotidyl transferases or nucleases. Experimental determination of their activity will be of major interest.

Bacterial DnaG primases clearly form a monophyletic assemblage (strongly supported by neighbor-joining and maximal parsimony trees; data not shown), with a distinctive N-terminal Zn-chelating domain involved in DNA binding and possibly conferring weak sequence specificity to recognition of the primer initiation sites

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022

Table 1. Phyletic distribution of proteins containing the Toprim domain

	DnaG- type TOPRIM proteins			DNA Topoisomerase TOPRIM domains				Others	
	DnaG primase	**Small DnaG family proteins**	**Large DnaG family proteins**	**Topoisomerase I A Family**	**Reverse gyrase**	**TopoII/gyrase family**	**Topo VI**	**RecR family**	**OLD family**
Bacteria	All bacteria - 1 Mycoplasmas - 2, (1 degenerate copy)	*B. subtilis* - 2 *A. aeolicus* - 1 Mycoplasmas- 1 *B. burgdorferi* - 1 Proteobacteria - none	-	At least 1 copy in all bacteria Bacillus -2 Some plasmids encode their own Topo Is	*A. aeolicus* - 1 *T.maritima* - 1 (so far)	Bacteria-2 (Gyrases and TopoIV)	-	Several bacterial genera -1, Gram positives -1	*E. coli* - 1 *H. pylori* - 2 *Synechocystis* - 1
Archaea	-	*A. fulgidus* - 1 *M. jannaschii* -1 *M. thermoautotrophicum* - 2	*A. fulgidus* - 1 *M. jannaschii* -1 *M. therm.* -1 *Sulfolobus* - 1	All archaeae - 1	All archaea - 1	Archaeoglobus and Haloferax - 1 each (Independent horizontal transfer from bacteria)	All archaea - 1	-	*M. jannaschii* - 2 On extrachromosomal Element
Eukarya	-	-	-	Topo III *S. cerevisiae* -1 *C. elegans* -1 (so far)	-	Topo II *S. cerevisiae* - 1 *C. elegans* - 2 (so far)	*S. cerevisiae* - 1 *C. elegans* -1 (meiotic recombination endonuclease spo11 in *S. pombe*	-	-

For each species, the number of detected members of the given protein family is indicated; a dash indicates that no members were detected.

(54). The C-terminal domain of DnaG interacts with the DnaB helicase (55,56), whereas in bacteriophages T3, T7 and P22, the DnaG-type primase domain is fused to a C-terminal DnaB-like domain and, finally, in bacteriophage P4, there is a DnaG–superfamily III helicase fusion (Fig. 3). Interestingly, phylogenetic analysis with both neighbor-joining and maximal parsimony methods for tree construction showed that one of the small Toprim proteins from mycoplasmas has been derived by degeneration of the DnaG-type primase (data not shown). In archaea, the Toprim-containing proteins with the greatest similarity to bacterial DnaG form a highly conserved orthologous family (the criteria for the identification of orthologs in genome comparisons have been described previously; 57,58) whose unique feature is the presence of an N-terminal domain with a conserved motif that is similar to motif VI of superfamily II helicases (Fig. 2; 59). There was no statistically significant sequence similarity to helicases, but pattern searches (using the pattern QxxGRxGR) showed that this motif is unique for a large subset of the helicases and this particular family of archaeal DnaG-like proteins. In the helicases, motif VI has been shown to be involved in DNA or RNA binding (60). It may have a similar nucleic acid-binding function in the archaeal proteins and may be functionally equivalent to the Zn-binding domain in DnaG. Given the eukaryotic layout of the basic replication machinery in archaea and, in particular, the presence of genes for both subunits of the eukaryotic type primase (8,44; L.Aravind, unpublished observations), it appears most likely that the archaeal DnaG homologs are involved in repair rather than in replication.

The topoisomerases also show diverse domain architectures (Fig. 3). In Topo IA, the Toprim domain seems to be associated with two structural repeats, as shown by the crystal structure (19). These repeats form a saddle-shaped β-sheet structure with α-helical connectors which is thought to wrap around dsDNA (45). The reverse gyrases contain a unique N-terminal fusion to a superfamily II helicase domain and, in some cases, also the insertion of the 'little finger' domain, a small, widespread and mobile nucleic acid-binding, finger-like module that appears to be structurally distinct from other C4 fingers (L.Aravind, unpublished observations), into the Toprim domain (Fig. 2). In Topo II, the Toprim domain is embedded in the midst of other structurally well-defined domains that form a toroidal ring around the DNA (61); these domains are fused to the mutL/hsp90-type ATPase domain (Fig. 2; 14,17). Interestingly, one of the domains located between the ATPase and Toprim has been reported to show structural similarity to ribosomal protein S5 and domain IV of translation elongation factor G, suggesting a conserved mode of nucleic acid binding between Topo II and these RNA-binding proteins (62). In addition, the *Caenorhabditis elegans* Topo II has an inserted SET domain, which may correlate with the association of these enzymes with chromatin remodeling complexes in eukaryotes (63,64).

The repair proteins of the RecR/M family are the smallest representatives of the Toprim superfamily, with the exception of the stand alone primase-like proteins, and combine Toprim with two DNA-binding domains, namely the helix–hairpin–helix motif (65) and the 'little finger' domain (Fig. 2). Conceivably, these domains mediate non-specific DNA binding, whereas the inactivated version of the Toprim domain may provide specificity towards specific structural features present in damaged DNA (see above).

The bacteriophage P2 OLD (overcome lysogenization defect) protein, which has DNase as well as RNase activity (66), consists of an N-terminal ABC-type ATPase domain (67,68) and a C-terminal Toprim domain (Fig. 3); the nuclease activity of OLD is stimulated by ATP, though the ATPase activity is not DNA-dependent (66). Unfortunately, functional details on OLD are scant and further experimentation is required to define the relationship between the ATPase and Toprim nuclease domains. Apparent orthologs of OLD with the same domain organization are detectable in several bacteria and in an extrachromosomal element of the archaeon *M.jannaschii*, but, on the whole, the distribution of this family is scattered (Table 1), suggesting dissemination by horizontal gene transfer, possibly via bacteriophage and plasmid vectors.

Finally, an interesting domain architecture was observed in a *Zymomonas* plasmid-encoded protein, in which the Toprim domain is fused to a mobilization (MOB) domain (69); in this case, the Toprim domain may be involved in strand nicking and/or rejoining.

At least two lines of experimental evidence on other enzyme families indicate an intimate connection between topoisomerase, ligase and nuclease activities, which is compatible with the idea of ancient enzymes that might have had them all. Firstly, a series of recent studies has shown that Topo IB, in addition to the topoisomerase activity, has the activities of a site-specific

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022

ribonuclease (70) and a polynucleotide ligase (71). Secondly, the *Nae*I restriction endonuclease has been converted into a topoisomerase-recombinase by a single amino acid residue change (72,73). None of these protein families showed any detectable sequence or structural (as can be ascertained for Topo IB) similarity to the Toprim domain, showing that multifunctionality of topoisomerase-like enzymes is a recurrent theme in evolution.

Given the ubiquitous presence of the Toprim domain (Table 1), it is most likely that it was already encoded by the common ancestor of all life forms (the cenancestor). The ancestral protein might have resembled the extant small proteins that consist of a solo Toprim domain. Such a protein could function as a low specificity enzyme with both a nucleotidyl transferase and a polynucleotide cleaving activity. Furthermore, the ability of DnaG-type primases and OLD family nucleases to, respectively, synthesize or cleave RNA suggests that this ancestor Toprim domain might have operated even in the primeval RNA world.

In the course of further evolution of DNA replication and repair, adaptation of the ancestral Toprim domain for specific functions should have occurred primarily through duplication and fusion with other domains. In addition to the specialized enzymatic functions such as primase, topoisomerase and nuclease, the RecR/M proteins are a case of apparent recruitment of the Toprim domain for a non-enzymatic function. The observed phylogenetic distribution of Toprim-containing proteins with distinct domain architectures (Table 1 and Fig. 2) suggests that while the Toprim domain itself may trace back to the cenancestor, the domain fusions that involve it do not, with the possible exception of Topo IA. Even such fundamental functions as primase and Topo II may have evolved independently through domain accretion in the main phylogenetic lineages, with subsequent multiple horizontal gene transfer events complicating the picture. Horizontal transfer seems to account, for example, for the presence of Topo II in archaea and reverse gyrase in bacteria. Indeed, Topo II is found only in a limited subset of the archaea (Table 1) and these proteins are distinctly more similar to the bacterial homologs than to eukaryotic ones. Conversely, reverse gyrase, while universal among the archaea, is so far strictly limited to thermophilic bacteria, suggesting an archaeal origin. The most surprising aspect of the phylogenetic distribution of the Toprim domain is its absence (so far) in eukaryotes in forms other than topoisomerases (Table 1). It appears that in eukaryotes, the ancestral Toprim domain enzyme(s) might have been completely displaced by the archaeal–eukaryotic type primases and by other types of nucleases.

Using sensitive methods for sequence comparison aided by structural modeling, this study revealed the previously unsuspected structural, functional and evolutionary connection between enzymes with diverse roles in DNA metabolism. The findings presented here seem to open a window into a very early stage of cellular evolution when a single ancestral domain might have performed several distinct functions in replication, repair and nucleic acid metabolism.

REFERENCES

1 Kornberg,A. and Baker,T. (1991) *DNA Replication*, 2nd Edn. W.H.Freeman and Co., New York, NY.
2 Salas,M. (1991) *Annu. Rev. Biochem.*, **60**, 39–71.
3 Traktman,P. (1990) *Curr. Topics Microbiol. Immunol.*, **163**, 93–123.
4 Koonin,E.V. and Ilyina,T.V. (1993) *Biosystems*, **30**, 241–268.
5 Mendelman,L.V. (1995) *Methods Enzymol.*, **262**, 405–414.
6 Ilyina,T.V., Gorbalenya,A.E. and Koonin,E.V. (1992) *J. Mol. Evol.*, **34**, 351–357.
7 Griep,M.A. (1995) *Indian J. Biochem., Biophys.*, **32**, 171–178.
8 Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) *Mol. Microbiol.*, **25**, 619–637.
9 Barrett,J.W., Lauzon,H.A., Mercuri,P.S., Krell,P.J., Sohi,S.S. and Arif,B.M. (1996) *Virus Genes*, **13**, 229–237.
10 Dracheva,S., Koonin,E.V. and Crute,J.J. (1995) *J. Biol. Chem.*, **270**, 14148–14153.
11 Wang,J.C. (1996) *Annu. Rev. Biochem.*, **65**, 635–692.
12 Tse,Y. and Wang,J.C. (1980) *Cell*, **22**, 269–276.
13 Cheng,C., Kussie,P., Pavletich,N. and Shuman,S. (1998) *Cell*, **92**, 841–850.
14 Bergerat,A., de Massy,B., Gadelle,D., Varoutas,P.C., Nicolas,A. and Forterre,P. (1997) *Nature*, **386**, 414–417.
15 Guipaud,O., Marguet,E., Noll,K.M., de la Tour,C.B. and Forterre,P. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 10606–10611.
16 Berger,J.M., Gamblin,S.J., Harrison,S.C. and Wang,J.C. (1996) *Nature*, **379**, 225–232.
17 Mushegian,A.R., Bassett,D.E., Jr, Boguski,M.S., Bork,P. and Koonin,E.V. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 5831–5836.
18 Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
19 Hubbard,T.J.P., Murzin,A.G., Brenner,S.E. and Chothia,C. (1997) *Nucleic Acids Res.*, **25**, 236–239.
20 Holm,L. and Sander,C. (1998) *Nucleic Acids Res.*, **26**, 316–319.
21 Walker,D.R. and Koonin,E.V. (1997) *Ismb*, **5**, 333–339.
22 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
23 Karlin,S. and Altschul,S.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
24 Altschul,S.F. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.
25 Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
26 Abagyan,R.A. and Batalov,S. (1997) *J. Mol. Biol.*, **273**, 355–368.
27 Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
28 Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) *Nature Genet.*, **6**, 119–129.
29 Bork,P. and Koonin,E.V. (1998) *Nature Genet.*, **18**, 313–318.
30 Wootton,J.C. and Federhen,S. (1993) *Computers Chem.*, **17**, 149–163.
31 Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol.*, **266**, 554–571.
32 Lupas,A. (1996) *Methods Enzymol.*, **266**, 513–525.
33 Wilson,J.A., Hill,J.E., Kuzio,J. and Faulkner,P. (1995) *J. Gen. Virol.*, **76**, 2923–2932.
34 Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) *Protein Sci.*, **4**, 1618–1632.
35 Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) *Proteins*, **9**, 180–190.
36 Galtier,N., Gouy,M. and Gautier,C. (1996) *Comput. Appl. Biosci.*, **12**, 543–548.
37 Barton,G.J. (1993) *Protein Engng.*, **6**, 37–40.
38 Swofford,D.L. (1993) *PAUP 3.1*. Illinois Natural History Survey, Champaign, IL.
39 Rost,B., Sander,C. and Schneider,R. (1994) *Comput. Appl. Biosci.*, **10**, 53–60.
40 Rost,B., Schneider,R. and Sander,C. (1997) *J. Mol. Biol.*, **270**, 471–480.
41 Peitsch,M.C. (1996) *Biochem. Soc. Trans.*, **24**, 274–279.
42 Peitsch,M.C. (1997) *Ismb*, **5**, 234–236.
43 Edgell,D.R. and Doolittle,W.F. (1997) *Cell*, **89**, 995–998.
44 Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K., Harrison,D., Hoang,L., Keagle,P., Lumm,W., Pothier,B., Qiu,D., Spadafora,R., Vicaire,R., Wang,Y., Wierzbowski,J., Gibson,R., Jiwani,N., Caruso,A., Bush,D., Reeve,J.N. *et al.* (1997) *J. Bacteriol.*, **179**, 7135–7155.
45 Lima,C.D., Wang,J.C. and Mondragon,A. (1994) *Nature*, **367**, 138–146.
46 Strack,B., Lessl,M., Calendar,R. and Lanka,E. (1992) *J. Biol. Chem.*, **267**, 13062–13072.
47 Steitz,T.A. (1998) *Nature*, **391**, 231–232.
48 Zhu,C.X., Roche,C.J., Papanicolaou,N., DiPietrantonio,A. and Tse-Dinh,Y.C. (1998) *J. Biol. Chem.*, **273**, 8783–8789.
49 Chen,S.J. and Wang,J.C. (1998) *J. Biol. Chem.*, **273**, 6050–6056.
50 Urlacher,T.M. and Griep,M.A. (1995) *Biochemistry*, **34**, 16708–16714.
51 Webb,B.L., Cox,M.M. and Inman,R.B. (1997) *Cell*, **91**, 347–356.
52 Courcelle,J., Carswell-Crumpton,C. and Hanawalt,P.C. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 3714–3719.
53 Hegde,S.P., Qin,M.H., Li,X.H., Atkinson,M.A., Clark,A.J., Rajagopalan,M. and Madiraju,M.V. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 14468–14473.

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022

54 Kusakabe,T. and Richardson,C.C. (1996) *J. Biol. Chem.*, **271**, 19563–19570.
55 Tougu,K., Peng,H. and Marians,K.J. (1994) *J. Biol. Chem.*, **269**, 4675–4682.
56 Tougu,K. and Marians,K.J. (1996) *J. Biol. Chem.*, **271**, 21391–21397.
57 Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1996) *Curr. Biol.*, **6**, 279–291.
58 Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
59 Gorbalenya,A.E. and Koonin,E.V. (1993) *Curr. Opin. Struct. Biol.*, **3**, 419–429.
60 Hall,M.C., Ozsoy,A.Z. and Matson,S.W. (1998) *J. Mol. Biol.*, **277**, 257–271.
61 Berger,J.M. and Wang,J.C. (1996) *Curr. Opin. Struct. Biol.*, **6**, 84–90.
62 Murzin,A.G. (1995) *Nature Struct. Biol.*, **2**, 25–26.
63 Varga-Weisz,P.D., Wilm,M., Bonte,E., Dumas,K., Mann,M. and Becker,P.B. (1997) *Nature*, **388**, 598–602.
64 Orlando,V. and Paro,R. (1995) *Curr. Opin. Genet. Dev.*, **5**, 174–179.
65 Doherty,A.J., Serpell,L.C. and Ponting,C.P. (1996) *Nucleic Acids Res.*, **24**, 2488–2497.
66 Myung,H. and Calendar,R. (1995) *J. Bacteriol.*, **177**, 497–501.
67 Koonin,E.V. and Gorbalenya,A.E. (1992) *Protein Sequece Data Anal.*, **5**, 43–45.
68 Gorbalenya,A.E. and Koonin,E.V. (1990) *J. Mol. Biol.*, **213**, 583–591.
69 Guzman,L.M. and Espinosa,M. (1997) *J. Mol. Biol.*, **266**, 688–702.
70 Sekiguchi,J. and Shuman,S. (1997) *Mol. Cell*, **1**, 89–97.
71 Shuman,S. (1998) *Mol. Cell*, **1**, 741–748.
72 Jo,K. and Topal,M.D. (1995) *Science*, **267**, 1817–1820.
73 Jo,K. and Topal,M.D. (1996) *Nucleic Acids Res.*, **24**, 4171–4175.

NOTE ADDED IN PROOF

Additional iterative PSI-BLAST searches initiated with the sequences of archaeal Topo VI subunits and their eukaryotic orthologs (such as the *S.pombe* recombination nuclease SPOII) and using updated sequence databases detect a similarity to Toprim domain of the small DnaG-like proteins, albeit at a marginally significant level (e-value of ~0.06). This may support our hypothesis that Topo VI and their homologs contain a distinct version of the Toprim domain. Furthermore, additional searches resulted in the detection of cyanobacterial proteins (GI 1001384 from *Synechocystis* sp., and GI 497626 from a *Synechococcus* plasmid) that contain a fusion of the Toprim domain with a Superfamily III helicase, which is analogous to the domain architecture of the bacteriophage P4 plasmid; the *Synechocystis* protein is the first chromosomal occurrence of this domain combination. Finally, while this manuscript was being processed for publication, an independent description of the relationship between Topo IA and Topo II has been published [Berger,J.M., Fass,D., Wang,J.C. and Harrison,S.C. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 7876–7881].

Downloaded from https://academic.oup.com/nar/article/26/18/4205/2902131 by U.S. Department of Justice user on 16 August 2022