BMC Structural Biology

Research article

Open Access

Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases

Lakshminarayan M Iyer, Eugene V Koonin and L Aravind*

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Email: Lakshminarayan M Iyer - lakshmin@ncbi.nlm.nih.gov; Eugene V Koonin - koonin@ncbi.nlm.nih.gov;
L Aravind* - aravind@ncbi.nlm.nih.gov

* Corresponding author

Published: 28 January 2003

BMC Structural Biology 2003, **3**:1

Received: 18 January 2003
Accepted: 28 January 2003

This article is available from: http://www.biomedcentral.com/1472-6807/3/1

Abstract

Background: The eukaryotic RNA-dependent RNA polymerase (RDRP) is involved in the amplification of regulatory microRNAs during post-transcriptional gene silencing. This enzyme is highly conserved in most eukaryotes but is missing in archaea and bacteria. No evolutionary relationship between RDRP and other polymerases has been reported so far, hence the origin of this eukaryote-specific polymerase remains a mystery.

Results: Using extensive sequence profile searches, we identified bacteriophage homologs of the eukaryotic RDRP. The comparison of the eukaryotic RDRP and their homologs from bacteriophages led to the delineation of the conserved portion of these enzymes, which is predicted to harbor the catalytic site. Further, detailed sequence comparison, aided by examination of the crystal structure of the DNA-dependent RNA polymerase (DDRP), showed that the RDRP and the β' subunit of DDRP (and its orthologs in archaea and eukaryotes) contain a conserved double-psi β-barrel (DPBB) domain. This DPBB domain contains the signature motif DbDGD (b is a bulky residue), which is conserved in all RDRPs and DDRPs and contributes to catalysis via a coordinated divalent cation. Apart from the DPBB domain, no similarity was detected between RDRP and DDRP, which leaves open two scenarios for the origin of RDRP: i) RDRP evolved at the onset of the evolution of eukaryotes via a duplication of the DDRP β' subunit followed by dramatic divergence that obliterated the sequence similarity outside the core catalytic domain and ii) the primordial RDRP, which consisted primarily of the DPBB domain, evolved from a common ancestor with the DDRP at a very early stage of evolution, during the RNA world era. The latter hypothesis implies that RDRP had been subsequently eliminated from cellular life forms and might have been reintroduced into the eukaryotic genomes through a bacteriophage. Sequence and structure analysis of the DDRP led to further insights into the evolution of RNA polymerases. In addition to the β' subunit, β subunit of DDRP also contains a DPBB domain, which is, however, distorted by large inserts and does not harbor a counterpart of the DbDGD motif. The DPBB domains of the two DDRP subunits together form the catalytic cleft, with the domain from the β' subunit supplying the metal-coordinating DbDGD motif and the one from the β subunit providing two lysine residues involved in catalysis. Given that the two DPBB domains of DDRP contribute completely different sets of active residues to the catalytic center, it is hypothesized that the ultimate ancestor of RNA polymerases functioned as a homodimer of a generic, RNA-binding DPBB domain. This ancestral protein probably did not have catalytic activity and served as a cofactor for a ribozyme RNA polymerase. Subsequent evolution of DDRP and RDRP involved accretion of distinct sets of additional domains. In the DDRPs, these included a RNA-binding Zn-ribbon, an AT-hook-like module and a sandwich-barrel hybrid motif (SBHM) domain. Further, lineage-specific accretion of SBHM domains and other, DDRP-specific domains is observed in bacterial DDRPs. In contrast, the orthologs of the β' subunit in archaea and eukaryotes contains a four-stranded α + β domain that is shared with the α-subunit of bacterial DDRP, eukaryotic DDRP subunit RBP11, translation factor eIF1 and type II topoisomerases. The additional domains of the RDRPs remain to be characterized.

Conclusions: Eukaryotic RNA-dependent RNA polymerases share the catalytic double-psi β-barrel domain, containing a signature metal-coordinating motif, with the universally conserved β' subunit of DNA-dependent RNA polymerases. Beyond this core catalytic domain, the two classes of RNA polymerases do not have common domains, suggesting early divergence from a common ancestor, with subsequent independent domain accretion. The β-subunit of DDRP contains another, highly diverged DPBB domain. The presence of two distinct DPBB domains in two subunits of DDRP is compatible with the hypothesis that the ultimate ancestor of RNA polymerases was a RNA-binding DPBB domain that had no catalytic activity but rather functioned as a homodimeric cofactor for a ribozyme polymerase.

Background

Polymerization of ribonucleoside triphosphates (NTPs), which is central to a variety of crucial biological processes, including transcription, primer synthesis during DNA replication, addition of polyA tails to mRNA, addition of CCA to tRNA, uridylation in RNA editing, replication of RNA viruses, amplification of RNA in eukaryotic post-transcriptional gene silencing (PTGS), and oligoadenylate synthesis during interferon signaling, is catalyzed by a broad variety of distinct enzymes [1–4]. These polymerases belong to two major mechanistic categories, those that copy a nucleic acid template and those that are independent of a template. The DNA-dependent RNA polymerases (DDRPs) involved in the transcription of cellular and DNA viral genes, primases, and RNA-dependent RNA polymerases (RDRPs) of RNA viruses and cellular PTGS systems are template-dependent polymerases. Template-independent RNA polymerases (nucleotidyltransferases) include CCA-adding enzymes, polyA polymerases, uridylyl transferases and oligoA-synthetases. Some of the RNA polymerases function as single-subunit proteins, whereas others form large complexes that consist of multiple, distinct subunits; however, in all known cases, the basic catalytic activity maps to a single polypeptide containing a characteristic metal-chelating active site [5–7]. Typically, active sites of polymerases contain acidic or polar residues, which coordinate divalent metal cations, most often Mg^{2+}. The metal cations direct a 5' nucleoside triphosphate to form a phosphoester bond with the 3' hydroxyl of the preceding nucleotide (the 3'-terminal nucleotide of a growing polynucleotide chain), with the elimination of pyrophosphate [6].

Despite the basic biochemical similarity among all RNA polymerases, sequence and structure comparisons indicate that these enzymes belong to at least five evolutionarily unrelated folds. The RDRPs of RNA viruses define one major lineage of nucleic acid polymerases, which additionally includes reverse transcriptases, archaeo-eukaryotic DNA polymerases, and nucleotide cyclases [8–13]. The DNA-dependent RNA polymerase of certain bacteriophages, such as T7, and the archaeo-eukaryotic primase (also detected in some bacteria) are divergent derivatives of the same fold [11,14]. The core catalytic domain of all these enzymes, the so-called "palm" domain, has an RNA-recognition motif (RRM)-like fold with strategically placed metal-coordinating residues, which form the active site [11,15,16]. In contrast, bacterial DnaG-type primases (also present in archaea and some eukaryotes) contain a polymerase domain of the Rossmann-like TOPRIM fold, which is shared with topoisomerases and OLD-family nucleases [17–19]. The recently solved structures of the DDRPs from yeast and the thermophilic bacterium Thermus thermophilus indicate that the β' subunit (according to the subunit nomenclature of Escherichia coli DDRP, which we hereinafter employ to designate all orthologs of the respective E. coli subunits) of these enzymes defines another distinct catalytic scaffold, which is unrelated to any of the above template-dependent RNA polymerases [20–24]. Additionally, the structural and evolutionary affinities of two other template-dependent RNA polymerases, namely RDRPs involved in PTGS [25–27] and primases of herpesviruses [28], remain obscure. In contrast to the template-dependent RNA polymerases, which have several distinct scaffolds of the catalytic domains, all template-independent RNA polymerases have the same fold of the principal catalytic domain and belong to the pol β superfamily of nucleotidyl transferases [29,30].

Thus, RNA polymerase activity apparently has been independently "invented" on several occasions. Nevertheless, DDRP is one of the most conserved enzymes, which is represented, without exception, in all cellular life forms [31,32]. The DDRP complex from most organisms consists of 5 to 15 polypeptides. Of these, four subunits, which correspond to the bacterial α, β, β' and ω, are universally present in all cellular DDRPs and constitute the conserved DDRP core [7]. Orthologs of the β and β' subunits are also encoded in the genomes of several families of large eukaryotic DNA viruses [33–35]. Biochemical studies have shown that the catalytic site of DDRP resides in the β' subunit and contains three invariant aspartates that coordinate a Mg^{2+} cation [22–24,36]. Recent structural analyses demonstrated that the core of DDRP is assembled around the β and β' subunits, which interact with each other to form a positively charged nucleic-acid-binding cleft. The α subunit further stabilizes this cleft, whereas the ω subunit interacts with the β' subunit [7]. Given their ubiquitous presence in all cellular life forms and high level of sequence conservation, it seems evident that the core subunits of this complex RNA synthesis machine were already present in the last universal common ancestor (LUCA) of the extant cellular life forms and performed functions mechanistically similar to those of modern DDRP.

In contrast to the ubiquitous DDRP, cellular RDRPs that are involved in PTGS so far have been detected only in eukaryotes [1,26]. The PTGS phenomenon covers a variety of complementarity-dependent silencing pathways, such as RNA interference (RNAi) in animals and slime mold, co-suppression (silencing of transgene and the corresponding endogenous genes) and virus gene resistance in plants, and quelling in fungi, all of which share a common mechanism of RNA turnover [28,37–40]. Essentially, double-stranded (ds) RNA, which is formed in these processes, triggers the activation of a sequence-specific RNA degradation system, which targets homologous RNAs [41,42]. The dsRNA is broken down by the Dicer enzyme into 21–25 nucleotide (nt) fragments called small interfering RNA or siRNAs [43–45]. The siRNAs subsequently associate with a complex of proteins called RISC and target homologous RNA by serving as guides for multiple rounds of RNA cleavage [45,46].

The RDRP is a component of the PTGS system that has been initially described in plants as a cellular RNA polymerase activity induced upon viral infection that synthesized antisense RNA in a primer-dependent or independent manner [26,47–50]. Subsequent genetic studies showed that mutations in RDRP genes impaired PTGS in a variety of systems [25,38,51,52]. Experimental studies in *C. elegans* and in plants suggested that the RDRP is involved in production and amplification of dsRNA using the target RNA as template and siRNAs as primers or guides and resulting in amplification of the RNA silencing response [53–56].

The RDRP is present in one or more copies in a wide range of eukaryotes, from early-branching parabasalids, such as *Giardia*, to multicellular forms, including fungi, plants and animals [57,58]. This broad representation in eukaryotes, including organisms that are considered primitive, suggests that RDRP might have been encoded in the genome of the common ancestor of all modern eukaryotes. However, this gene has been subsequently lost in several eukaryotic lineages, such as the yeast *Saccharomyces cerevisiae*, insects and vertebrates. Although RDRPs display some diversity in domain architecture and the level of sequence conservation, they form a tight, well-conserved family of large proteins with no detectable prokaryotic or viral homologs.

Here, we investigate the evolutionary history of RDRPs and DDRPs in an attempt to unveil the origin of the RDRP. We identify the first homologs of the cellular RDRP outside the eukaryotic clade, in bacteriophages, and show that RDRPs and DDRPs share a homologous catalytic core, which comprises a six-stranded double-psi-β-barrel (DPBB) domain. Evidence is presented that the ultimate ancestor of RNA polymerases might have been a RNA-binding DPBB domain, which functioned as a cofactor for a polymerase ribozyme, and that subsequent evolution of DDRPs and RDRPs proceeded via accretion of various simple modules, such as the sandwich-barrel hybrid motif domain, around this conserved core.

Results and discussion

Bacteriophage homologs of the eukaryotic RDRPs and prediction of their active site

To investigate the evolutionary affinities of the RDRPs, we carefully examined the results of BLAST searches of the non-redundant (NR) protein sequence database (National Center for Biotechnology Information, NIH, Bethesda) for various RDRP sequences. These searches identified a large region of 700–800 residues as the conserved module shared by all RDRPs. This region was further examined by searches of NR using the PSI-BLAST program, which was iterated to convergence with a profile inclusion threshold of E = 0.01. These searches, e.g. the search initiated with the core RDRP module from *Petunia* (residues 37–775), readily retrieved RDRPs from plants, *C. elegans*, *Dictyostelium discoideum*, several fungi, and *Giardia lamblia* with statistically highly significant expectation (E) values. Interestingly, the second iteration of this search retrieved the YonO protein from the *Bacillus subtilis* phage Spβc2 with a significant E-value (7×10^{-3}). YonO was also detected with significant E-values in searches initiated with other eukaryotic RDRP modules, e.g., that of *Schizosaccha-*

romyces pombe. Reciprocal searches started with the YonO protein sequence retrieved closely related proteins encoded by prophages in the genome of *Clostridium acetobutylicum* (CAC1139) and *Clostridium perfringens* (CPE1103). Hereinafter, we refer to these bacteriophage proteins as the YonO-like RDRP homologs (YRHs).

The Gibbs sampling alignment procedure detected 9 statistically significant motifs that are conserved across the entire set of RdRps and YRHs, with a probability of occurrence by chance in this set of protein sequences estimated as $<10^{-18}$. A multiple alignment of the RDRPs and YRHs was constructed using the T-Coffee program and adjusted using the alignments reported by PSI-BLAST and secondary structure prediction. Secondary structure prediction, which was produced using the multiple alignment as the query, included 16 α-helices and 20 β-strands, which are conserved in RDRPs and YRHs (Fig. 1). Conserved sequence motifs are distributed throughout the aligned region; two motifs with the highest density of conserved residues are located between strands 8 and 11 and strands 18 and 20 (Fig. 1). Twelve residues are conserved in all sequences from the two families, of which nine are charged and one is polar (serine) (Fig. 1). The conserved charged and polar residues conceivably might be involved in catalysis and substrate-binding. The most notable signature shared by the RDRPs and the YRH proteins is the DbDGD (b is a bulky residue) motif located between strands 19 and 20 (Fig. 1). Closely spaced acidic residues that coordinate divalent cations are characteristic of the active sites of most nucleic acid polymerases, in spite of the fact that they belong to several unrelated structural folds [9,13,14,17,28,29,59]. Given that the DbDGD motif is the only set of closely spaced acidic residues shared by the RDRPs and the YRHs, it is likely to form part of the nucleotidyltransferase active site of these enzymes. Thus, these comparisons identify previously undetected prokaryotic homologs of the eukaryotic RDRPs, the YRH proteins, which probably also have RNA-dependent RNA polymerase activity.

Structural and evolutionary relationship between the catalytic domains of the RDRPs and the DDRPs

Interestingly, the PSI-BLAST searches started with the sequences of the RDRP module of the RDRPs and the YRH proteins consistently retrieved the β' subunits of DDRPs, albeit at statistically not significant E-values. The high-scoring segment pairs (HSPs) detected in these searches aligned the highly conserved region of the RDRPs between the predicted strands 18 and 20, including the DbDGD motif, with the portion of the DDRP β' subunit sequence, which contains the metal-chelating active site, with a similar conserved motif, DxDGD. Additionally, a search of the NR database with a position-specific scoring matrix (PSSM) that included all unique RDRP and YRH sequences detected, with borderline E-values, a large protein from Corynebacterium glutamicum (Cgl1702), which also contained a DXDGD motif. Reciprocal searches showed that this protein contained two regions of statistically significant similarity to the DDRPs. The N-terminal region of similarity corresponded to the conserved portion of the β subunits and the C-terminal region corresponded to the core of the β' subunit and included the DXDGD motif (see discussion below). Despite the low statistical significance of the similarity between RDRPs and DDRPs, the presence of the same signature motif in the experimentally identified or predicted active sites of the two classes of enzymes prompted us to perform a detailed comparison of the catalytic domains of these two classes of RNA polymerases.

Among nucleic acid polymerases, a DxDxD motif is conserved only in the RDRPs, the DDRP β' subunits and the euryarchaea-specific DNA polymerase subunit II. However, in the latter polymerase, the motif has the signature DGDED, as opposed to the DxDGD pattern shared by the RDRPs and DDRPs. To investigate the distribution of this motif in the entire protein database, we conducted pattern searches using the GREF program of the SEALS package, with queries corresponding to the residue conservation profile in the predicted active sites of the RDRPs. All these queries contained the DxDGD motif with additional flanking conserved residues. The only proteins with evolutionarily conserved DxDGD motifs that were detected in these searches were the RDRPs, the β' subunit of DDRPs, the integrin calcium-binding module, and the EF-hand domain (data not shown). Among these domains, the motif was embedded in the context of conserved (predicted) β-strands only in the RDRPs and DDRPs. Moreover, in searches conducted with extended queries, such as DbDGDxhxh or DGDxhxh patterns (h is a hydrophobic residue), the RDRPs and the β' subunits of DDRPs were identified as the only protein families in which these motifs were conserved. We also ran database searches using the PHI-BLAST program that combines a regular BLAST search with a pattern search [60]. In these searches, the active site motif was used as the pattern query, and various RDRP sequences as the sequence queries. All these searches retrieved the YonO proteins with significant E-values and some of the DDRP β' subunits with borderline E-values. For example, searches seeded with the RDRP (RrpB gene product) sequence from *Dictyostelium* (gi:14475571) and the pattern [GAS]-D- [FLYMQN]-D- [G]-D-X-[ACLIVMFY]-X- [ACLIVMFY] detected the rice DDRP catalytic subunit with an E-value of 0.048.

Examination of the recently solved structures of the DDRP showed that, although β' subunit is a large protein, the active site containing the DxDGD motif mapped to a small, compact β-barrel domain. A search of the PDB database

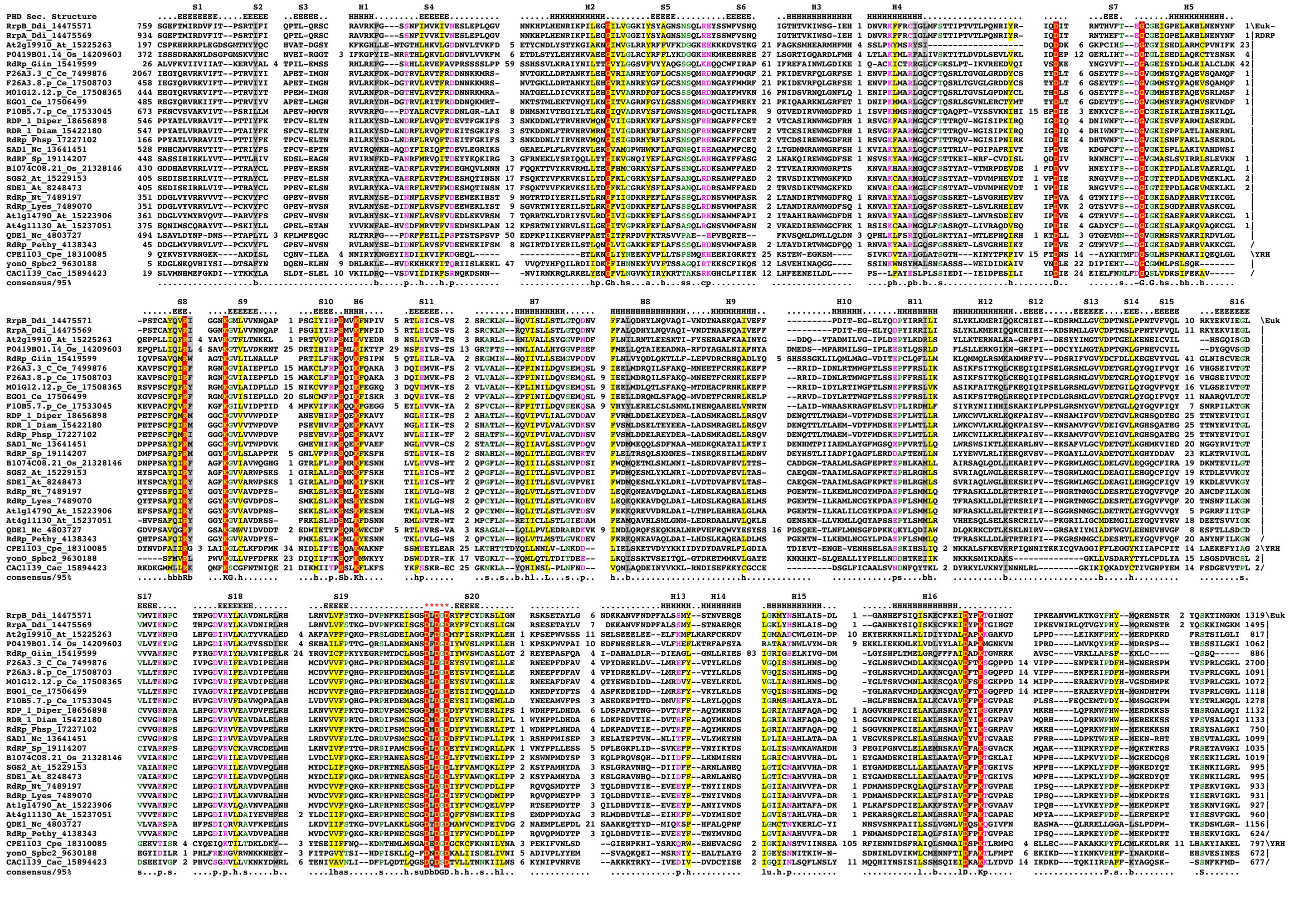


Figure 1

Multiple sequenced alignment of the conserved RDRP module from eukaryotic RNA-dependent RNA polymerases and their bacteriophage homologs. The sequences are denoted by gene names, abbreviated species names and Gene Identification (GI) numbers from the GenBank database. Species name abbreviations: At: *Arabidopsis thaliana*, Cac: *Clostridium acetobutylicum*, Ce: *Caenorhabditis elegans*, Cpe: *Clostridium perfringens*, Ddi: *Dictyostelium discoideum*, Diam: *Diaporthe ambigua*, Diper: *Diaporthe perjuncta*, Giin: *Giardia intestinalis*, Lyes: *Lycopersicon esculentum*, Nc: *Neurospora crassa*, Nt: *Nicotiana tabacum*, Os: *Oryza sativa*, Pethy: *Petunia hybrida*, Phsp: *Phomopsis* sp., Sp: *Schizosaccharomyces pombe*, Spbc2: Bacteriophage Spβc2. The positions of the first and the last residue of the aligned region in the corresponding protein are indicated before and after each sequence, respectively. The numbers between aligned blocks represent poorly conserved inserts that are not shown. The coloring is based on the 95% consensus shown underneath the alignment; h indicates hydrophobic residues (ACFILMVWY), a indicates aromatic residues (FYW), l indicates aliphatic residues (ILVA), p indicates polar residues (STEDKRNQHC), c indicates charged residues (DEKR), bi indicates bulky residue (ILFYWMKREQ), **s** indicates small residues (AGSVCDN). The predicted secondary structure elements are shown below the alignment; H indicates α-helix and E indicate extended conformation (β-strand). The predicted helices are marked H1-16 and the predicted strands S1-20. The signature motif DbDGD, which is predicted to form part of the catalytic site, is shown in reverse shading. The three sequences at the bottom are those of bacteriophage homologs of the RDRPs (the YRH proteins); the remaining are RDRP sequences.

using the DALI program [61], with this β-barrel from *Thermus thermophilus* β' subunit submitted as a query (pdb: 1iw7, chain D region: 625–749), retrieved the N-terminal domain of the CDC48-like AAA ATPases, formate dehydrogenases, aspartate decarboxylases and Barwin. The reported structural alignments span ~70 residues with significant Z-scores (8.8–6.6) and root mean square deviations (RMSD) between the C-α backbones in the range of 1.8–2.5 Å, which is characteristic of homologous domains. The region of structural similarity shared by the catalytic domain of the β' subunit of the DDRPs and the above proteins corresponds to a distinct globular fold known as the double-psi beta-barrel (DPBB) [62–64]. The SCOP database included the RNA polymerase subunits in the multidomain protein class but mentions the presence of a DPBB domain in each of the β and β' subunits ([65]; see below).

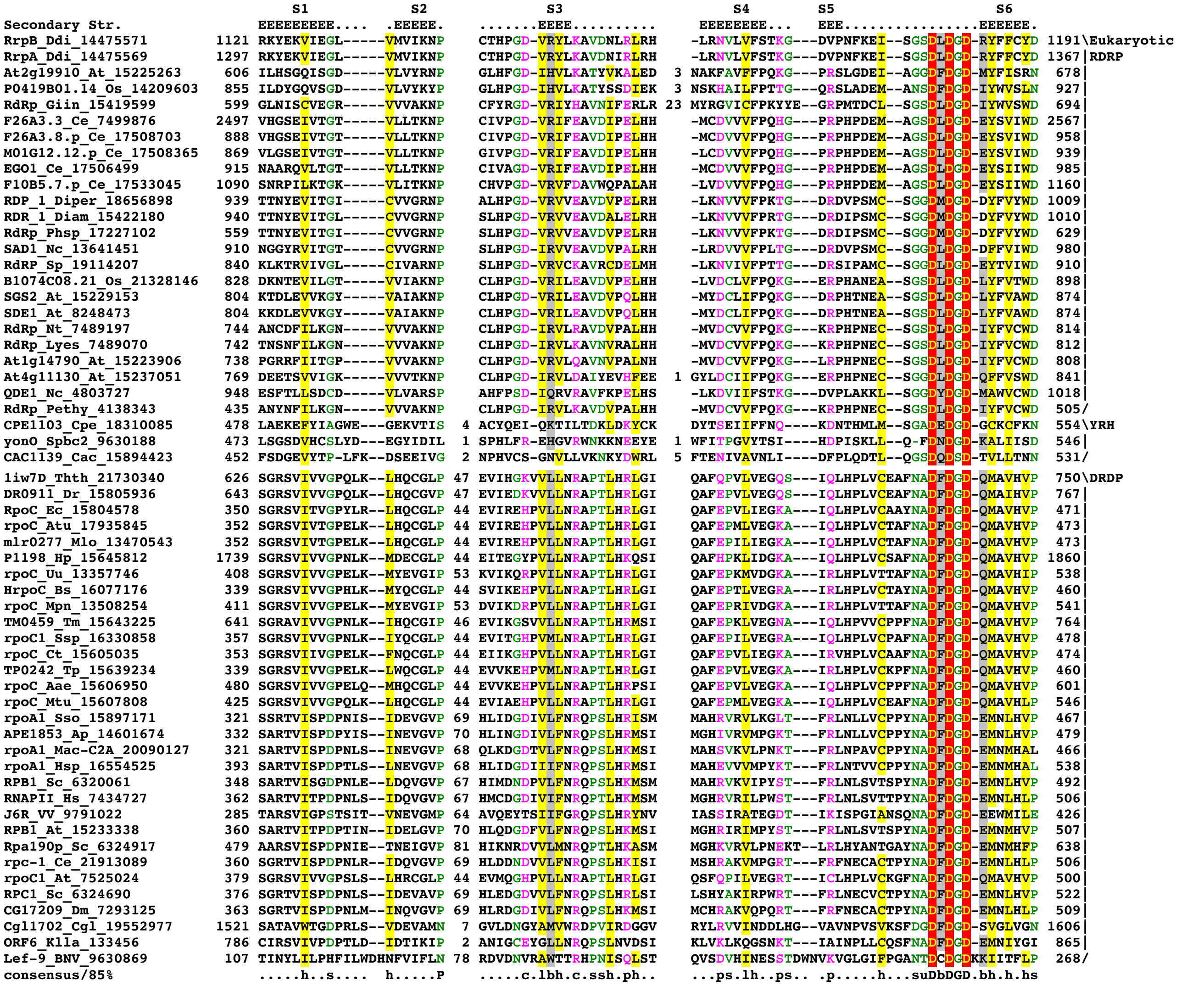
The region of sequence similarity between the RDRPs and the β' subunits of DDRPs, that was detected in the searches described above, almost precisely mapped to the DPBB-fold catalytic domain of DDRP. The multiple-alignment-based secondary structure prediction for the RDRPs and the YRHs indicated that, like the similar region of the β' subunit, this region was enriched in β-strands (Fig. 1). Multiple alignment constructed on the basis of sequence alignments produced in the PSI-BLAST and PHI-BLAST searches and superposition of the predicted secondary structure elements of RDRP and the experimentally determined structure of DDRP revealed considerable concordance between the RDRP and DDRP families. In addition to the DbDGD motif, which is located in the loop between strands 5 and 6, several residues are conserved between these polymerase families (Fig. 2). These include hydrophobic residues in the β-strands, a conserved proline after strand 2, a small (typically, glycine) and polar (typically, aspartate) residues before strand 3, small residues that typically mark the boundaries of secondary structure elements and several charged and polar residues (Fig. 2). A comparison of the RDRP and DDRP sequences using the Gibbs sampling method identified two conserved blocks, which were highly statistically significant ($E < 10^{-16}$) within the analyzed sequence set. One of these blocks centered at the DbDGD motif, whereas the other one encompassed strands 3 and 4, which contain several partially conserved hydrophobic and polar positions (Fig. 2).

Importantly, approximately 10 conserved residues that are detected in the structure-based alignments of the DPBB domains are also conserved in the RDRPs (Fig. 3). These include a doublet comprised of a small and a polar residues preceding each of strands 3 and 6, a small residue after strand 2 (proline in the DDRP β' subunit and RDRP), bulky and polar residues in strand 3, a small residue immediately downstream of strand 5, a bulky residue in the loop between strands 5 and 6, and two hydrophobic residues in strand 6 (Fig. 3). The predicted DPBB domains of the RDRPs differ from most of the DPBB domains of DDRP β' subunits in that the former lack a long and variable, in both length and sequence, insert between strands 2 and 3 that is characteristic of the latter (Fig. 2). However, the DPBB from the RNA polymerases of yeast killer plasmids and the divergent predicted corynebacterial polymerase Cgl1702 also lack this insert (Fig. 2). Taken together, these observations suggest that, although RDRPs and the β' subunits of DDRPs show limited sequence similarity, these two RNA polymerase families share a homologous catalytic domain, which consists of a double-psi-β-barrel containing the metal-coordinating DbDGD motif.

The DPBB domains and early evolution of the catalytic domain of RNA polymerases

The SCOP database, while classifying the DDRP subunits in the multidomain category, recognizes two DPBBs in the DDRP complex. One of these corresponds to the catalytic, metal-coordinating domain, containing the DbDGD motif, in the β' subunit (residues: 626–750 of pdb id:1iw7 chain D), whereas the other one corresponds to the conserved core domain of the β subunit (residues:673–994 of pdb id: 1iw7 chain C). The DPBB domains of the β and β' subunits interact to form the catalytic cleft of the RNA polymerases, with two conserved lysines in the β subunit protruding into the cleft and interacting with the substrate (Fig. 4). However, the DPBB domain in the β subunit is distorted by two large inserts (Fig. 4 and see below) and shows no detectable sequence similarity to the β' subunit DPBB. Because of these inserts and the resulting distortion, the DPBB of the β subunit does not show similarity to any domains in structure database searches. Nevertheless, a structural alignment based on visual inspection confirms the presence of all the *bona fide* features of the DPBB domain in the β subunit (Fig. 3). The Cgl1702 protein from *Corynebacterium*, which we identified as one of the most divergent members of the DDRP clade, showed similarity to the other DDRPs only in two regions, which corresponded to the DPBBs from the β and the β' subunit. This observation, taken together with the spatial proximity of the two DPBBs in the catalytic cleft of the DDRP (Fig. 4), indicates that these domains comprise the ancestral conserved core of the RNA polymerases. The arrangement of the two DPBBs in the β and β' subunits suggests that, in the primordial ancestor of the DDRPs, the two DPBBs formed a "head-to-tail" homodimer that bound RNA at the domain interface.

Although the cores of the two DDRP subunits are homologous and are likely to have evolved from a common ancestor (see discussion below), they have completely different sets of conserved residues and contribute distinct moieties to the catalytic cleft (Figs. 3, 4). Thus, these specific active residues apparently emerged after the divergence of the β and β' subunit DPBBs, whereas the ancestral DPBB homodimer probably did not have catalytic activity and merely bound a ribozyme that originally catalyzed the RNA polymerization. Subsequently, the two subunits diverged from each other, with one (β') acquiring a divalent cation-binding site in the insert between the two terminal strands of the DPBB, whereas the DPBB in the other (β) subunit acquired two basic residues that interacted with the nucleic acid. At this stage, which might be considered the emergence of modern-type RNA



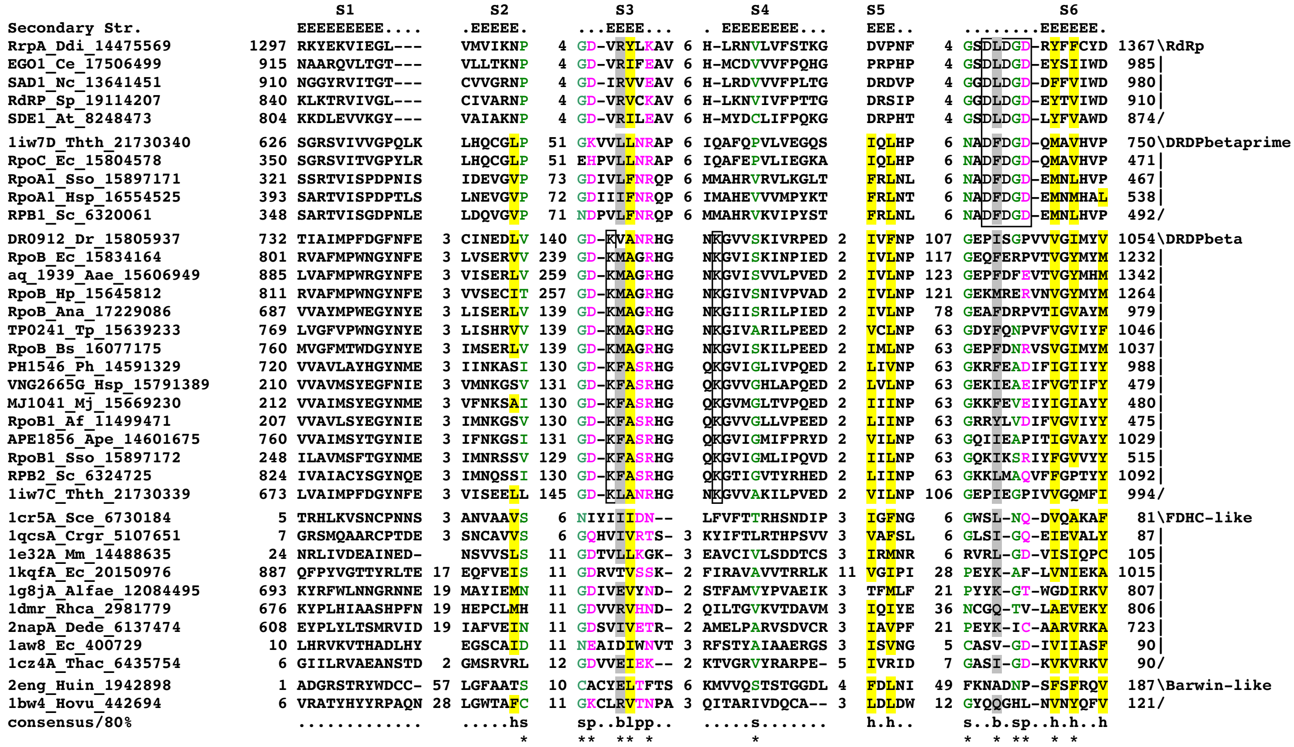
```
                                     S1         S2                S3                       S4     S5                   S6
Secondary Str.                  EEEEEEEEE....  .EEEEE.    .......EEEE..........     EEEEEEEE...   EEE................EEEEEE.
RrpB_Ddi_14475571          1121 RKYEKVIEGL-----VMVIKNP    CTHPGD-VRYLKAVDNLRLRH    -LRNVLVFSTKG---DVPNFKEI--SGSDLDGD-RYFFCYD 1191\Eukaryotic
RrpA_Ddi_14475569          1297 RKYEKVIEGL-----VMVIKNP    CTHPGD-VRYLKAVDNIRLRH    -LRNVLVFSTKG---DVPNFKEI--SGSDLDGD-RYFFCYD 1367|RDRP
At2g19910_At_15225263       606 ILHSGQISGD-----VLVYRNP    GLHFGD-IHVLKATYVKALED  3 NAKFAVFFPQKG---PRSLGDEI--AGGDFDGD-MYFISRN  678|
P0419B01.14_Os_14209603     855 ILDYGQVSGD-----VLVYKYP    GLHPGD-IHVLKATYSSDIEK  3 NSKHAILFPTTG---QRSLADEM--ANSDFDGD-IYWVSLN  927|
RdRp_Giin_15419599          599 GLNISCVEGR-----VVVAKNP    CFYRGD-VRIYHAVNIFERLR 23 MYRGVICFPKYYE--GRPMTDCL--SGSDLDGD-IYWVSWD  694|
F26A3.3_Ce_7499876         2497 VHGSEIVTGT-----VLLTKNP    CIVPGD-VRIFEAVDIPELHH    -MCDVVVFPQHG---PRPHPDEM--AGSDLDGD-EYSVIWD 2567|
F26A3.8.p_Ce_17508703       888 VHGSEIVTGT-----VLLTKNP    CIVPGD-VRIFEAVDIPELHH    -MCDVVVFPQHG---PRPHPDEM--AGSDLDGD-EYSVIWD  958|
M01G12.12.p_Ce_17508365     869 VLGSEIVTGT-----VLLTKNP    GIVPGD-VRIFEAVDIPELHH    -LCDVVVFPQHG---PRPHPDEM--AGSDLDGD-EYSVIWD  939|
EGO1_Ce_17506499            915 NAARQVLTGT-----VLLTKNP    CIVAGD-VRIFEAVDIPELHH    -MCDVVVFPQHG---PRPHPDEM--AGSDLDGD-EYSIIWD  985|
F10B5.7.p_Ce_17533045      1090 SNRPILKTGK-----VLITKNP    CHVPGD-VRVFDAVWQPALAH    -LVDVVVFPQHG---PRPHPDEM--AGSDLDGD-EYSIIWD 1160|
RDP_1_Diper_18656898        939 TTNYEVITGI-----CVVGRNP    ALHPGD-VRVVEAVDVPELRH    -LKNVVVFPKTG---DRDVPSMC--SGGDMDGD-DYFVYWD 1009|
RDR_1_Diam_15422180         940 TTNYEVITGL-----CVVGRNP    SLHPGD-VRVVEAVDALELRH    -LKNVVVFPRTG---DRDIPSMC--SGGDMDGD-DYFVYWD 1010|
RdRp_Phsp_17227102          559 TTNYEVITGI-----CVVGRNP    SLHPGD-VRVVEAVDVPELRH    -LKNVVVFPKTG---DRDIPSMC--SGGDMDGD-DYFVYWD  629|
SAD1_Nc_13641451            910 NGGYRVITGT-----CVVGRNP    SLHPGD-IRVVEAVDVPALRH    -LRDVVVFPLTG---DRDVPSMC--SGGDLDGD-DFFVIWD  980|
RdRP_Sp_19114207            840 KLKTRVIVGL-----CIVARNP    SLHPGD-VRVCKAVRCDELMH    -LKNVIVFPTTG---DRSIPAMC--SGGDLDGD-EYTVIWD  910|
B1074C08.21_Os_21328146     828 DKNTEVILGT-----VVIAKNP    CLHPGD-VRILEAVDVPELHH    -LVDCLVFPQKG---ERPHANEA--SGSDLDGD-LYFVTWD  898|
SGS2_At_15229153            804 KTDLEVVKGY-----VAIAKNP    CLHPGD-VRILEAVDVPQLHH    -MYDCLIFPQKG---DRPHTNEA--SGSDLDGD-LYFVAWD  874|
SDE1_At_8248473             804 KKDLEVVKGY-----VAIAKNP    CLHPGD-VRILEAVDVPQLHH    -MYDCLIFPQKG---DRPHTNEA--SGSDLDGD-LYFVAWD  874|
RdRp_Nt_7489197             744 ANCDFILKGN-----VVVAKNP    CLHPGD-IRVLRAVDVPALHH    -MVDCVVFPQKG---KRPHPNEC--SGSDLDGD-IYFVCWD  814|
RdRp_Lyes_7489070           742 TNSNFILKGN-----VVVAKNP    CLHPGD-IRVLKAVNVRALHH    -MVDCVVFPQKG---KRPHPNEC--SGSDLDGD-IYFVCWD  812|
At1g14790_At_15223906       738 PGRRFIITGP-----VVVAKNP    CLHPGD-VRVLQAVNVPALNH    -MVDCVVFPQKG---LRPHPNEC--SGSDLDGD-IYFVCWD  808|
At4g11130_At_15237051       769 DEETSVVIGK-----VVVTKNP    CLHPGD-IRVLDAIYEVHFEE  1 GYLDCIIFPQKG---ERPHPNEC--SGGDLDGD-QFFVSWD  841|
QDE1_Nc_4803727             948 ESFTLLSDCD-----VLVARSP    AHFPSD-IQRVRAVFKPELHS    -LKDVIIFSTKG---DVPLAKKL--SGGDYDGD-MAWVCWD 1018|
RdRp_Pethy_4138343          435 ANYNFILKGN-----VVVAKNP    CLHPGD-IRVLKAVDVPALHH    -MVDCVVFPQKG---KRPHPNEC--SGSDLDGD-IYFVCWD  505/
CPE1103_Cpe_18310085        478 LAEKEFYIAGWE---GEKVTIS  4 ACYQEI-QKTILTDKLDKYCK    DYTSEIIFFNQ----KDNTHMLM--SGADEDGD-GCKCFKN  554\YRH
yonO_Spbc2_9630188          473 LSGSDVHCSLYD---EGYIDIL  1 SPHLFR-EHGVRWNKKNEEYE  1 WFITPGVYTSI----HDPISKLL--Q-FDNDGD-KALIISD  546|
CAC1139_Cac_15894423        452 FSDGEVYTP-LFK--DSEEIVG  2 NPHVCS-GNVLLVKNKYDWRL  5 FTENIVAVNLI----DFPLQDTL--QGSDQDSD-TVLLTNN  531/
1iw7D_Thth_21730340         626 SGRSVIVVGPQLK--LHQCGLP 47 EVIHGKVVLLNRAPTLHRLGI    QAFQPVLVEGQS---IQLHPLVCEAFNADFDGD-QMAVHVP  750\DRDP
DR0911_Dr_15805936          643 SGRSVIVVGPQLK--LHQCGVP 47 EVIEDKVVLLNRAPTLHRLGI    QAFEPVLVEGQS---IQLHPLVCEAFNADFDGD-QMAIHVP  767|
RpoC_Ec_15804578            350 SGRSVITVGPYLR--LHQCGLP 44 EVIREHPVLLNRAPTLHRLGI    QAFEPVLIEGKA---IQLHPLVCAAYNADFDGD-QMAVHVP  471|
rpoC_Atu_17935845           352 SGRSVIVTGPELK--LHQCGLP 44 EVIREHPVLLNRAPTLHRLGI    QAFEPMLVEGKA---IQLHPLVCTAFNADFDGD-QMAVHVP  473|
mlr0277_Mlo_13470543        352 SGRSVIVTGPELK--LHQCGLP 44 EVIREHPVLLNRAPTLHRLGI    QAFEPILIEGKA---IQLHPLVCTAFNADFDGD-QMAVHVP  473|
P1198_Hp_15645812          1739 SGRSVIVVGPNLK--MDECGLP 44 EITEGYPVLLNRAPTLHKQSI    QAFHPKLIDGKA---IQLHPLVCSAFNADFDGD-QMAVHVP 1860|
rpoC_Uu_13357746            408 SGRSVIVVGPELK--MYEVGIP 53 KVIKQRPVILNRAPTLHRLGI    QAFEPKMVDGKA---IRLHPLVTTAFNADFDGD-QMAVHIP  538|
HrpoC_Bs_16077176           339 SGRSVIVVGPHLK--MYQCGLP 44 SVIKEHPVLLNRAPTLHRLGI    QAFEPTLVEGRA---IRLHPLVCTAYNADFDGD-QMAVHVP  460|
rpoC_Mpn_13508254           411 SGRSVIVVGPELK--MYEVGIP 53 DVIKDRPVLLNRAPTLHRLGI    QAFEPRIVDGKA---IRLHPLVTTAFNADFDGD-QMAVHVP  541|
TM0459_Tm_15643225          641 SGRAVIVVGPNLK--IHQCGIP 46 EVIKGSVVLLNRAPTLHRMSI    QAFEPKLVEGNA---IQLHPVVCPPFNADFDGD-QMAVHVP  764|
rpoC1_Ssp_16330858          357 SGRSVIVVGPNLK--IYQCGLP 44 EVITGHPVMLNRAPTLHRLGI    QAFEPILVEGRA---IQLHPLVCPAFNADFDGD-QMAVHVP  478|
rpoC_Ct_15605035            353 SGRSVIIVGPELK--FNQCGLP 44 EIIKGHPVLLNRAPTLHRLGI    QAFEPVLIEGKA---IRVHPLVCAAFNADFDGD-QMAVHVP  474|
TP0242_Tp_15639234          339 SGRSVIVVGPELK--LWQCGLP 44 EVVKEHPVMLNRAPTLHRLGI    QAFEPVLVEGKA---IRLHPLVCKPFNADFDGD-QMAVHVP  460|
rpoC_Aae_15606950           480 SGRSVIVVGPELQ--MHQCGLP 44 EVVKEHPVLLNRAPTLHRPSI    QAFEPVLVEGKA---IQLHPLVCPPFNADFDGD-QMAVHVP  601|
rpoC_Mtu_15607808           425 SGRSVIVVGPQLK--LHQCGLP 44 EVIAEHPVLLNRAPTLHRLGI    QAFEPMLVEGKA---IQLHPLVCEAFNADFDGD-QMAVHLP  546|
rpoA1_Sso_15897171          321 SSRTVISPDPNIS--IDEVGVP 69 HLIDGDIVLFNRQPSLHRISM    MAHRVRVLKGLT---FRLNLLVCPPYNADFDGD-EMNLHVP  467|
APE1853_Ap_14601674         332 SARTVISPDPYIS--INEVGVP 70 HLINGDIVLFNRQPSLHRMSI    MGHIVRVMPGKT---FRLNLLVCPPYNADFDGD-EMNLHVP  479|
rpoA1_Mac-C2A_20090127      321 SARTVISPDPNLS--INEVGVP 68 QLKDGDTVLFNRQPSLHKMSI    MAHSVKVLPNKT---FRLNPAVCPPYNADFDGD-EMNMHAL  466|
rpoA1_Hsp_16554525          393 SARTVISPDPTLS--LNEVGVP 68 HLIDGDIIIFNRQPSLHRMSI    MAHEVVVMPYKT---FRLNTVVCPPYNADFDGD-EMNMHAL  538|
RPB1_Sc_6320061             348 SARTVISGDPNLE--LDQVGVP 67 HIMDNDPVLFNRQPSLHKMSM    MAHRVKVIPYST---FRLNLSVTSPYNADFDGD-EMNLHVP  492|
RNAPII_Hs_7434727           362 SARTVITPDPNLS--IDQVGVP 67 HMCDGDIVIFNRQPTLHKMSM    MGHRVRILPWST---FRLNLSVTTPYNADFDGD-EMNLHLP  506|
J6R_VV_9791022              285 TARSVIGPSTSIT--VNEVGMP 64 AVQEYTSIIFGRQPSLHRYNV    IASSIRATEGDT---IKISPGIANSQNADFDGD-EEWMILE  426|
RPB1_At_15233338            360 SARTVITPDPTIN--IDELGVP 70 HLQDGDFVLFNRQPSLHKMSI    MGHRIRIMPYST---FRLNLSVTSPYNADFDGD-EMNMHVP  507|
Rpa190p_Sc_6324917          479 AARSVISPDPNIE--TNEIGVP 81 HIKNRDVVLMNRQPTLHKASM    MGHKVRVLPNEKT--LRLHYANTGAYNADFDGD-EMNMHFP  638|
rpc-1_Ce_21913089           360 SGRTVISPDPNLR--IDQVGVP 69 HLDDNDVVLFNRQPSLHKISI    MSHRAKVMPGRT---FRFNECACTPYNADFDGD-EMNLHLP  506|
rpoC1_At_7525024            379 SGRSVIVVGPSLS--LHRCGLP 44 EVMQGHPVLLNRAPTLHRLGI    QSFQPILVEGRT---ICLHPLVCKGFNADFDGD-QMAVHVP  500|
RPC1_Sc_6324690             376 SGRTVISPDPNLS--IDEVAVP 69 HLEDGDVVLFNRQPSLHRLSI    LSHYAKIRPWRT---FRLNECVCTPYNADFDGD-EMNLHVP  522|
CG17209_Dm_7293125          363 SGRTVISPDPNLM--INQVGVP 69 HLRDGDIVLFNRQPSLHKMSI    MCHRAKVQPQRT---FRFNECACTPYNADFDGD-EMNLHLP  509|
Cgl1702_Cgl_19552977       1521 SATAVWTGDPRLS--VDEVAMN  7 GVLDNGYAMVWRDPVIRDGGV    RYLRVVINDDLHG--VAVNPVSVKSFDGDFDGD-SVGLVGN 1606|
ORF6_Klla_133456            786 CIRSVIVPDPTLD--IDTIKIP  2 ANIGCEYGLLNRQPSLNVDSI    KLVKLKQGSNKT---IAINPLLCQSFNADFDGD-EMNIYGI  865|
Lef-9_BNV_9630869           107 TINYLILPHFILWDHNFVIFLN 78 RDVDNVRAWTTRHPNISQLST    QVSDVHINESSTDWNVKVGLGIFPGANTDCDGDKKIITFLP  268/
consensus/85%                   .....h..s....  h.....P    .....c.lbh.c.ssh.ph..    ...ps.lh..ps.. .p.....h...suDbDGD.bh.h.hs
```

Figure 2

Multiple alignment of the double-psi β-barrel (DPBB) domains from the β' subunit of DNA-dependent RNA polymerases with the predicted DPBB domains of RNA-dependent RNA polymerases. The conventions for naming sequences and coloring conserved residues are as described in the legend to Figure 1. The shared secondary structure elements are shown above the alignment with E denote the β-strand (extended) conformation. Non-conserved regions are depicted as numbers with inserts. The species abbreviations are as in figure Fig. 1. The species abbreviations that are not listed in the legend to Figure 1 are: Thth: *Thermus thermophilus*, Dr: *Deinococcus radiodurans*, Ec: *Escherischia coli*, Atu: *Agrobacterium tumefaciens*, Mlo: *Mesorhizobium loti*, Hp: *Helicobacter pylori*, Uu:*Ureaplasma ureolyticum*, Bs: *Bacillus subtilis*, Tm: *Thermotoga maritima*, Ssp: *Synechocystis* species, Aae: *Aquifex aeolicus*, Ct *Chlamydia trachomatis*, Mpn *Mycoplasma pneumoniae*, Mtu: *Mycobacterium tuberculosis*, Tp: *Treponema pallidum*, Sso: *Sulfolbus solfataricus*, Ap: *Aeropyrum pernix*, Mac: *Methanosarcina acetivorans*, Hsp: *Halobacterium* species, Sc: *Saccharomyces cerevisiae*, Hs: *Homo sapiens*, VV: Vaccinia virus, BNV: *Bombyx mori* Nuclear Polyhedrosis virus, Cgl: *Corynebacterium glutamicum*, Klla: *Kluyveromyces lactis.*

polymerases, the catalytic activity probably was taken over by the protein.

The basic DPBB fold consists of six β-strands and two variable regions, which are located after strand 2 and strand 5 and, at least in some case, adopt a helical conformation [62–64]. The domain can be further split into two



```
                                 S1             S2           S3             S4         S5                 S6
Secondary Str.              EEEEEEEEE....    .EEEEE.     ...EEEE...   . EEEEEEEE...    EEE..     ........EEEEEE.
RrpA_Ddi_14475569      1297 RKYEKVIEGL---    VMVIKNP   4 GD-VRYLKAV 6 H-LRNVLVFSTKG    DVPNF   4 GSDLDGD-RYFFCYD 1367\RdRp
EGO1_Ce_17506499        915 NAARQVLTGT---    VLLTKNP   4 GD-VRIFEAV 6 H-MCDVVVFPQHG    PRPHP   4 GSDLDGD-EYSIIWD  985|
SAD1_Nc_13641451        910 NGGYRVITGT---    CVVGRNP   4 GD-IRVVEAV 6 H-LRDVVVFPLTG    DRDVP   4 GGDLDGD-DFFVIWD  980|
RdRP_Sp_19114207        840 KLKTRVIVGL---    CIVARNP   4 GD-VRVCKAV 6 H-LKNVIVFPTTG    DRSIP   4 GGDLDGD-EYTVIWD  910|
SDE1_At_8248473         804 KKDLEVVKGY---    VAIAKNP   4 GD-VRILEAV 6 H-MYDCLIFPQKG    DRPHT   4 GSDLDGD-LYFVAWD  874/

1iw7D_Thth_21730340     626 SGRSVIVVGPQLK    LHQCGLP  51 GKVVLLNRAP 6 IQAFQPVLVEGQS    IQLHP   6 NADFDGD-QMAVHVP  750\DRDPbetaprime
RpoC_Ec_15804578        350 SGRSVITVGPYLR    LHQCGLP  51 EHPVLLNRAP 6 IQAFEPVLIEGKA    IQLHP   6 NADFDGD-QMAVHVP  471|
RpoA1_Sso_15897171      321 SSRTVISPDPNIS    IDEVGVP  73 GDIVLFNRQP 6 MMAHRVRVLKGLT    FRLNL   6 NADFDGD-EMNLHVP  467|
RpoA1_Hsp_16554525      393 SARTVISPDPTLS    LNEVGVP  72 GDIIIFNRQP 6 IMAHEVVVMPYKT    FRLNT   6 NADFDGD-EMNMHAL  538|
RPB1_Sc_6320061         348 SARTVISGDPNLE    LDQVGVP  71 NDPVLFNRQP 6 MMAHRVKVIPYST    FRLNL   6 NADFDGD-EMNLHVP  492/

DR0912_Dr_15805937      732 TIAIMPFDGFNFE  3 CINEDLV 140 GD-KVANRHG   NKGVVSKIVRPED  2 IVFNP 107 GEPISGPVVVGIMYV 1054\DRDPbeta
RpoB_Ec_15834164        801 RVAFMPWNGYNFE  3 LVSERVV 239 GD-KMAGRHG   NKGVISKINPIED  2 IVLNP 117 GEQFERPVTVGYMYM 1232|
aq_1939_Aae_15606949    885 LVAFMPWRGYNFE  3 VISERLV 259 GD-KMAGRHG   NKGVISVVLPVED  2 IVLNP 123 GEPFDFEVTVGYMHM 1342|
RpoB_Hp_15645812        811 RVAFMPWNGYNFE  3 VVSECIT 257 GD-KMAGRHG   NKGIVSNIVPVAD  2 IVLNP 121 GEKMRERVNVGYMYM 1264|
RpoB_Ana_17229086       687 VVAYMPWEGYNYE  3 LISERLV 139 GD-KMAGRHG   NKGIISRILPIED  2 IVLNP  78 GEAFDRPVTIGVAYM  979|
TP0241_Tp_15639233      769 LVGFVPWNGYNYE  3 LISHRVV 139 GD-KMAGRHG   NKGIVARILPEED  2 VCLNP  63 GDYFQNPVFVGVIYF 1046|
RpoB_Bs_16077175        760 MVGFMTWDGYNYE  3 IMSERLV 139 GD-KMAGRHG   NKGVISKILPEED  2 IMLNP  63 GEPFDNRVSVGIMYM 1037|
PH1546_Ph_14591329      720 VVAVLAYHGYNME  3 IINKASI 130 GD-KFASRHG   QKGVIGLIVPQED  2 LIVNP  63 GKRFEADIFIGIIYY  988|
VNG2665G_Hsp_15791389   210 VVAVMSYEGFNIE  3 VMNKGSV 131 GD-KFASRHG   QKGVVGHLAPQED  2 LVLNP  63 GEKIEAEIFVGTIFY  479|
MJ1041_Mj_15669230      212 VVAIMSYEGYNME  3 VFNKSAI 130 GD-KFASRHG   QKGVMGLTVPQED  2 IIINP  63 GKKFEVEIYIGIAYY  480|
RpoB1_Af_11499471       207 VVAVLSYEGYNIE  3 IMNKGSV 130 GD-KFASRHG   QKGVVGLLVPEED  2 LIINP  63 GRRYLVDIFVGVIYY  475|
APE1856_Ape_14601675    760 VVAIMSYTGYNIE  3 IFNKGSI 131 GD-KFASRHG   QKGVIGMIFPRYD  2 VILNP  63 GQIIEAPITIGVAYY 1029|
RpoB1_Sso_15897172      248 ILAVMSFTGYNME  3 IMNRSSV 129 GD-KFASRHG   QKGVIGMLIPQVD  2 IILNP  63 GQKIKSRIYFGVVYY  515|
RPB2_Sc_6324725         824 IVAIACYSGYNQE  3 IMNQSSI 130 GD-KFASRHG   QKGTIGVTYRHED  2 LIINP  63 GKKLMAQVFFGPTYY 1092|
1iw7C_Thth_21730339     673 LVAIMPFDGYNFE  3 VISEELL 145 GD-KLANRHG   NKGVVAKILPVED  2 VILNP 106 GEPIEGPIVVGQMFI  994/

1cr5A_Sce_6730184         5 TRHLKVSNCPNNS  3 ANVAAVS   6 NIYIIIDN--   LFVFTTRHSNDIP  3 IGFNG   6 GWSL-NQ-DVQAKAF   81\FDHC-like
1qcsA_Crgr_5107651        7 GRSMQAARCPTDE  3 SNCAVVS   6 GQHVIVRTS- 3 KYIFTLRTHPSVV  3 VAFSL   6 GLSI-GQ-EIEVALY   87|
1e32A_Mm_14488635        24 NRLIVDEAINED-    NSVVSLS  11 GDTVLLKGK- 3 EAVCIVLSDDTCS  3 IRMNR   6 RVRL-GD-VISIQPC  105|
1kqfA_Ec_20150976       887 QFPYVGTTYRLTE 17 EQFVEIS  11 GDRVTVSSK- 2 FIRAVAVVTRRLK 11 VGIPI  28 PEYK-AF-LVNIEKA 1015|
1g8jA_Alfae_12084495    693 KYRFWLNNGRNNE 19 MAYIEMN  11 GDIVEVYND- 2 STFAMVYPVAEIK  3 TFMLF  21 PYYK-GT-WGDIRKV  807|
1dmr_Rhca_2981779       676 KYPLHIAASHPFN 19 HEPCLMH  11 GDVVRVHND- 2 QILTGVKVTDAVM  3 IQIYE  36 NCGQ-TV-LAEVEKY  806|
2napA_Dede_6137474      608 EYPLYLTSMRVID 19 IAFVEIN  11 GDSVIVETR- 2 AMELPARVSDVCR  3 IAVPF  21 PEYK-IC-AARVRKA  723|
1aw8_Ec_400729           10 LHRVKVTHADLHY    EGSCAID  11 NEAIDIWNVT 3 RFSTYAIAAERGS  3 ISVNG   5 CASV-GD-IVIIASF   90|
1cz4A_Thac_6435754        6 GIILRVAEANSTD  2 GMSRVRL  12 GDVVEIEK-- 2 KTVGRVYRARPE-  5 IVRID   7 GASI-GD-KVKVRKV   90/

2eng_Huin_1942898         1 ADGRSTRYWDCC- 57 LGFAATS  10 CACYELTFTS 6 KMVVQSTSTGGDL  4 FDLNI  49 FKNADNP-SFSFRQV  187\Barwin-like
1bw4_Hovu_442694          6 VRATYHYYRPAQN 28 LGWTAFC  11 GKCLRVTNPA 3 QITARIVDQCA--  3 LDLDW  12 GYQQGHL-NVNYQFV  121/
consensus/80%               .............    .....hs     sp..blpp..   .....s.......    h.h..     s..b.sp..h.h..h
                                                   *     **  ** *          *                     *  * **  * *
```

Figure 3

A structure-based multiple alignment of the DPBB domains from the β and β' subunits of DDRPs with a selection of other structurally characterized DPBB domains and the predicted DPBB domains of RDRP. The alignments were generated by structural superposition of representatives of the DPBB domains followed by addition of sequence neighbors. The predicted DPBB domain of the RDRPs was included on the basis of the alignment with the DDRP β' subunit's DPBB domain. The functionally important lysines in the β subunits and the metal-coordinating aspartates in the β' subunits are boxed. The conserved positions shared by the RDRPs with the structurally characterized DPBBs are indicated below the alignment by asterisks. The other conventions and abbreviations are as in the legends to Fig. 1 and Fig. 2.

symmetrical structural units, each with three core β-strands. These two units are combined into a barrel with a complex topology, with strand 5 wedged between strands 1 and 2 and strand 2 wedged between strands 4 and 5, yielding two 'psi' loop structures (Fig. 4). The psi-loops between strands 1 and 2 and between strands 4 and 5 often harbor the active site or substrate-interaction residues of the respective DPBBs. Sequence comparisons and identification of specific conserved motifs showed that the DPBB fold includes five major superfamilies. These superfamilies are typified by the formate dehydrogenase C-terminal (FDHC) domain, Barwin endoglucanase, aspartyl protease, and the DPBB domains from the β and β' subunits of DDRP. Two of these superfamilies show a restricted phyletic distribution. Barwin-like endoglucanases are present only in plants and fungi, whereas aspartyl proteases are found in eukaryotes, retroviruses and a minority of bacteria. The latter superfamily also has a circular permutation, which suggests secondary derivation. Thus, none of these superfamilies are likely to have been present in the last universal common ancestor (LUCA) of extant life forms. In contrast, the DPBBs detected in DDRP subunits are universally present in all modern life forms and hence are traceable to LUCA. The FDHC superfamily includes several distinct families. The N-terminal domain of the CDC48-like AAA ATPases is present largely in eukaryotes and archaea and the N-terminal domain of thymidine phosphorylases is archaea-specific. In contrast, the FDHC-like DPBB domain associated with oxidoreductases, such as formate dehydrogenase and molybdopterin-containing dehydrogenases, is widespread in archaea and all major lineages of free-living bacteria. Thus, at least one DPBB domain of the FHDC superfamily probably was present in LUCA.

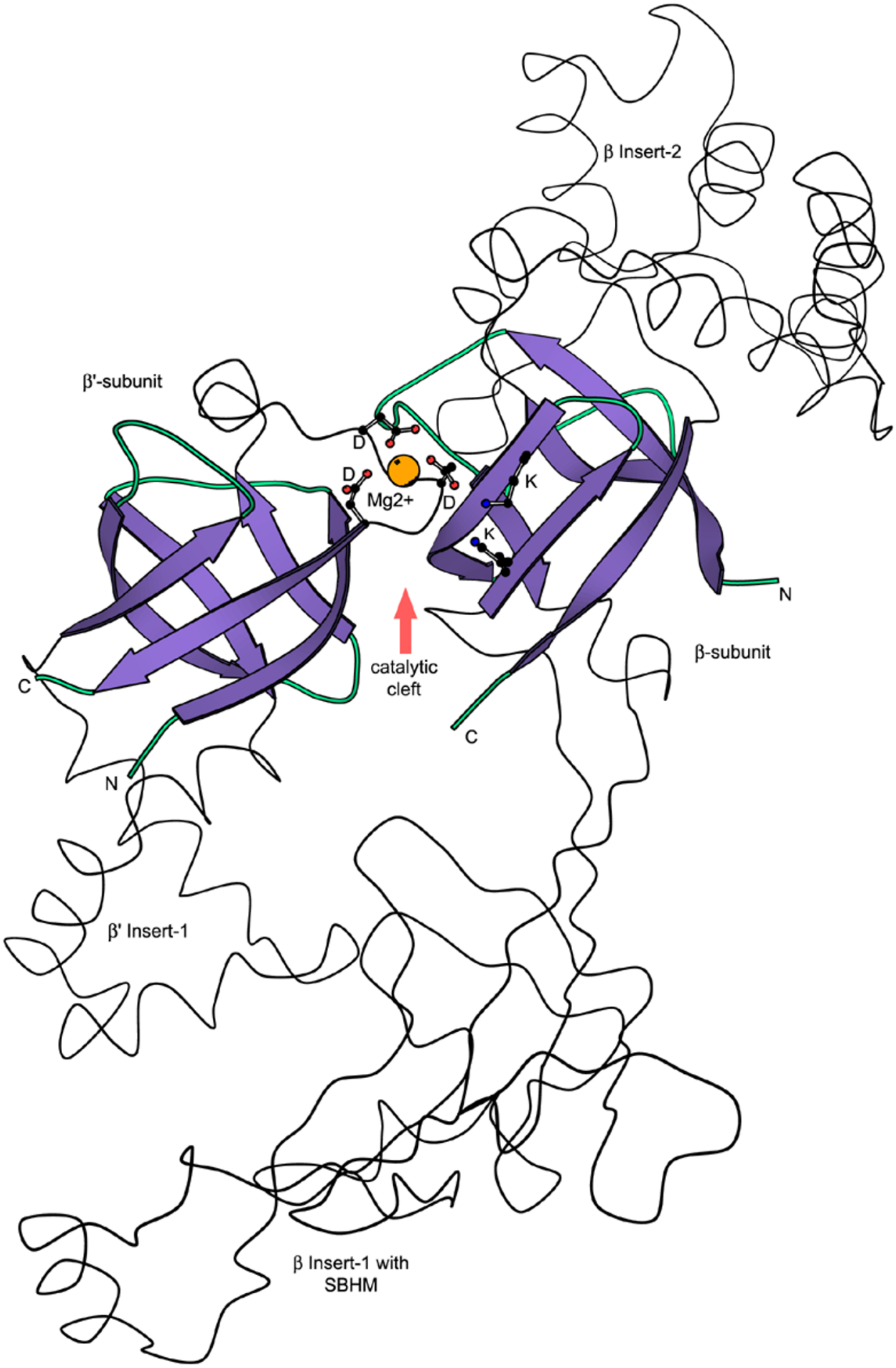


Figure 4

Structure of the catalytic cleft of DDRP formed by interacting DPBB domains of the β and β' subunits. The metal-coordinating DbDGD motif of the β' subunit and the functionally important lysines projecting into the catalytic cleft of the β subunit are shown in ball and stick representation. The two double psi-barrels are juxtaposed in an asymmetric head to tail configuration.

The above analysis shows that LUCA probably encoded three distinct DPBB domains. Although the DPBB domain in the RNA polymerase β subunit is distorted by inserts and shows extreme sequence divergence, its role in the formation of the catalytic cleft and the spatial arrangement with respect to the more canonical DPBB of the β' subunit (Fig. 4) are compatible with the hypothesis that the two DPBB domains of DDRP shared a common ancestor to the exclusion of the other ancient, FDHC-like DPBB. A structure-based sequence alignment shows that the DPBB domains of the RNA polymerases share several conserved residues with other DPBB domains. These include 7 hydrophobic residues, two sets of symmetrically positioned small (mostly glycine) and polar (mostly aspartate) residues preceding strands 3 and 6, respectively, and at least three other small residues, two polar residues and two bulky residues (Fig. 3). Of these, the small residues preceding strands 3 and 6 contribute a positive ϕ angle, which results in turning of the C-α backbone to stabilize the β-barrel [62]. These shared features notwithstanding, the ancient DPBBs have no conserved residues in their catalytic or substrate-binding sites. This is best compatible with the notion that the common ancestor of the DPBB domains was a generic binding domain devoid of high specificity or catalytic activity (see also discussion above).

The same three-stranded structural units that interlock to give rise to DPBB also form another type of β-barrel via terminal dimerization. This is the Elongation factor-Isomerase (EI)-barrel found in ribosomal proteins, such as L3, GTPase translation factors, ferredoxin reductase, and L-fucose isomerase [62,66]. It seems likely that the EI-barrel and the DPBB evolved from ancient three-stranded ancestral units, which were stabilized via the formation of different barrel structures through homo-dimerization.

Several versions of the DPBB, e.g., those in RNA polymerases, and of the EIB, such as those in translation elongation factors and ribosomal protein L3, are parts of ribonucleoprotein complexes. Other ancient forms, especially those that occur as accessory domains in enzymes, appear to bind small molecules. Hence, the primordial three-strand unit might have formed barrel-shaped dimers that non-specifically associated with RNA or with small molecules. These generic barrel domains subsequently diverged to occupy different functional niches in ribonucleoproteins and metabolic enzymes. These considerations, together with the inferred relationships between the different versions of the DPBB domain, lead to the scenario of early evolution outlined in Figure 5. Only the last bifurcation of the DPBB domains in the proposed tree yielded the distinct forms present in the β and β' subunits of DDRP, which apparently acquired catalytic activity. Substantial divergence of the β-barrels seems to have occurred at a stage of evolution when the actual RNA polymerase activity resided in a ribozyme whose functioning was facilitated by a non-specific, RNA-binding protein cofactor.

Ancient conserved domains and evolution of DNA-dependent RNA polymerases

While all bacteria, archaea and eukaryotes encode orthologs of at least four distinct subunits of DDRP, the divergent RNA polymerases of baculoviruses and yeast killer plasmids and the *Corynebacterium* Cgl1702 protein only contain counterparts to the β and β' subunits. The minimal RNA polymerase seems to be represented by the predicted catalytic core of Cgl1702, which appears to consist entirely of the β-type and β'-type DPBBs, in this case juxtaposed in the same polypeptide. These observations are not particularly surprising given that β and β' subunits form the catalytic cleft of the DDRP (Fig. 3), whereas the other subunits occupy peripheral positions in the complex. In an attempt to gain further insight into the early evolution of the catalytic core of RNA polymerases, we sought to identify additional globular domains in β and β' subunits.

Using the experimentally determined structure as a guide, we split the β and β' subunits of the *Thermus thermophilus* DDRP and the largest two subunits of the yeast DDRP into individual globular domains and investigated them by visual inspection of the topology, structural searches of the PDB database using the DALI program and iterative searches of the NR database using PSI-BLAST. An obvious ancient conserved domain thus detected was a rubredoxin-like Zn-ribbon. The Zn-ribbons are widespread, small domains that are comprised of two β-hairpins bounded by consecutive extended regions stabilized by metal-coordinating residues in the hairpin loops [67–70]. Zn-ribbons are present in a variety of nucleic-acid-binding proteins, including several ribosomal proteins, translation factors, aminoacyl-tRNA synthetases, RNA polymerase cofactors in archaea and eukaryotes, and several transcription factors. These domains are also present in other contexts where they function as structural scaffolds and as regulators of redox reactions. The β' subunit of bacterial DDRPs contains a single Zn-ribbon at the N-terminus (region 56–81 of *Thermus* β' subunit; Fig. 6). The orthologous largest subunits of the archaeal and eukaryotic DDRPs have two Zn-ribbons in the N-terminal portion of the protein (regions 59–88 and 103–173 in the yeast protein). The distal ribbon is distorted by a large insert between the two core halves of the domain and a substitution of asparagine for one of the metal-coordinating cysteines.

Although the sequences of the Zn-ribbons of bacterial and archaeo-eukaryotic DDRP subunits have diverged

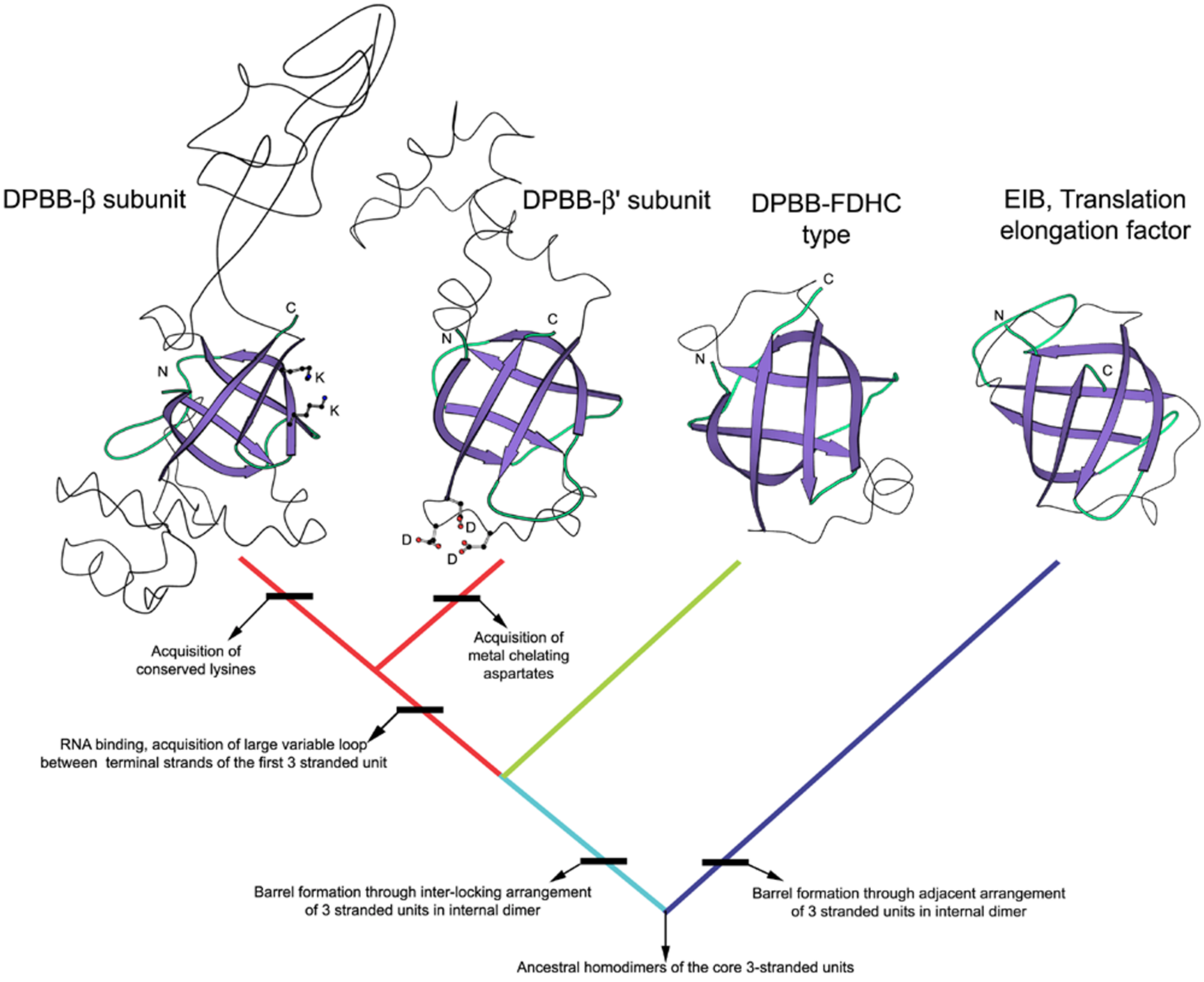

DPBB-β subunit
DPBB-β' subunit
DPBB-FDHC type
EIB, Translation elongation factor
N
C
K
K
D
D
D
Acquisition of conserved lysines
Acquisition of metal chelating aspartates
RNA binding, acquisition of large variable loop between terminal strands of the first 3 stranded unit
Barrel formation through inter-locking arrangement of 3 stranded units in internal dimer
Barrel formation through adjacent arrangement of 3 stranded units in internal dimer
Ancestral homodimers of the core 3-stranded units

Figure 5
A hypothetical scheme of evolution of two types of 6-stranded β-barrels from 3-stranded units. The scheme was derived as the most parsimonious explanation for the phyletic patterns and structural peculiarities of each lineage of 6-stranded barrels. The emergence of particular properties or characters characteristic of a given clade is indicated by horizontal bars.

considerably, their similar location in the orthologous polypeptides from different kingdoms suggests that β' subunit of LUCA already contained a Zn-ribbon. Based on the crystal structure, it has been proposed that the region containing the Zn ribbon may interact with the promoter during transcription initiation in bacteria [23]. Hence, it is plausible that a nucleic-acid-binding Zn-ribbon had been recruited to the ancestral polymerase core, which consisted of the DPBB domains, and contributed to non-specific interactions with the template. Consistent with this, the Zn-ribbon had been subsequently reused in several basal archaeo-eukaryotic transcription factors, such as TFIIB, TFIIE and TFIIS [67,68,71–73].

A potential additional nucleic-acid-binding moiety that is conserved in all β'-subunits is the AT-hook like module (Fig. 6). The AT-hook is a simple, small protein module that consists of a flap-like structure with a positively charged surface that is inserted into the minor groove of DNA [74]. The two AT-hook-like modules (residues 583–603 of 1IW7 chain D) in the DDRP β' subunit are inserted within the α-helix that is located immediately upstream of the DPBB domain (Fig. 6). Previously, AT-hooks have

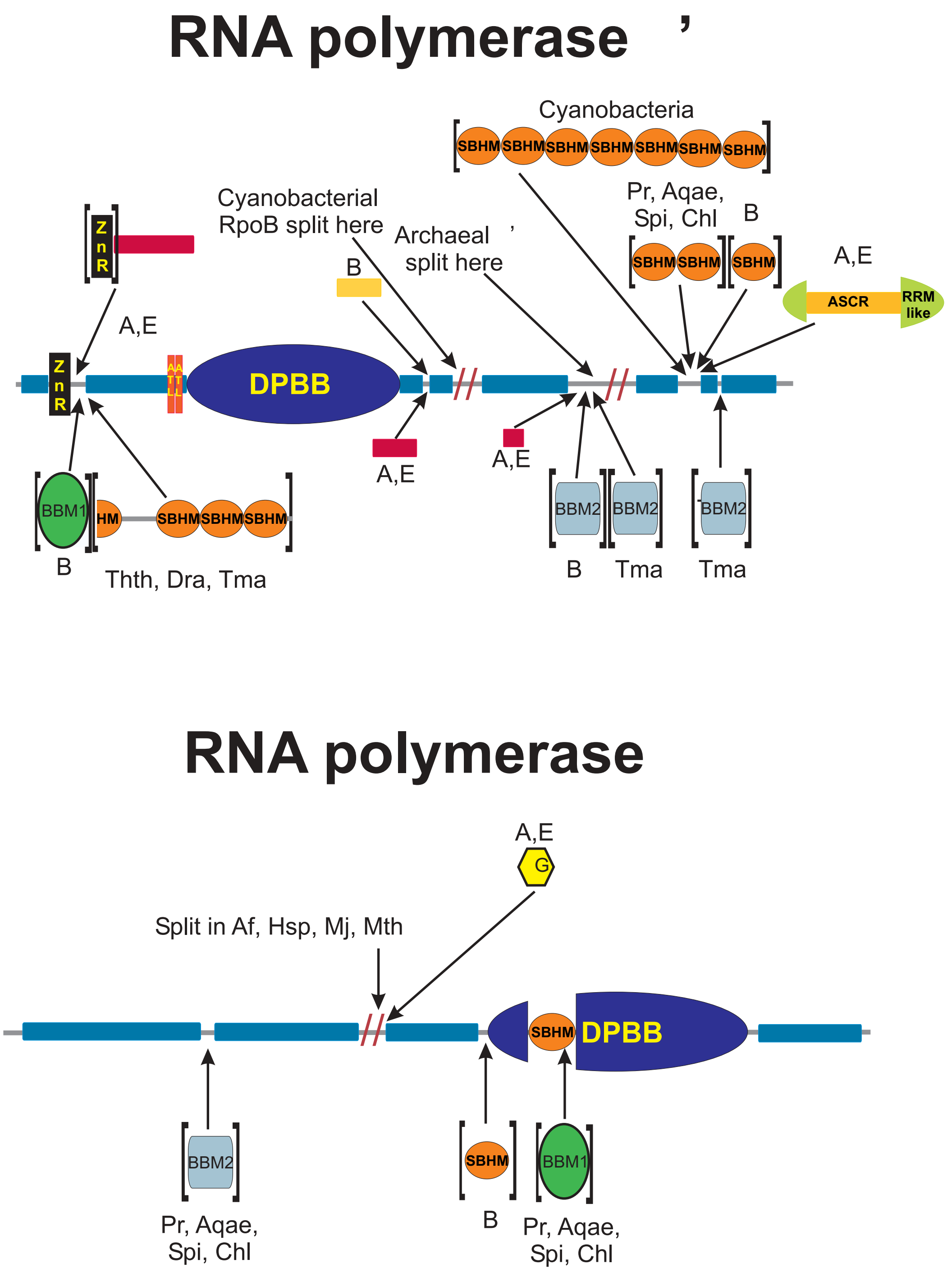


Figure 6

Domain architectures of the β and β' subunits of DDRP. Domain designations: DPBB, double-psi β-barrel, SBHM, sandwich-barrel hybrid motif, ZnR, Zn-ribbon, ATL, AT-hook like, BBM, β, β'-(specific) module. βG is a domain containing a minimal version of the β grasp fold. Other designations: A, archaea, B, bacteria, E, eukaryota, Pr, Proteobacteria, Aqae, Aquifex, Spi, spirochetes, Chl, Chlamydia, Tth, *Thermus thermophilus*, Dra, *Deinococcus radiodurans*, Tma, *Thermotoga maritima*; Af, *Archaeoglobus fulgidus*, Hsp, *Halobacterium sp*., Mj, *Methanocaldococcus jannaschii*, Mth, *Methanothermobacter thermoautotrophicus*. Other globular regions that are conserved between the cellular DDRPs are shown by blue rectangles and non-conserved regions are shown as gray lines. Globular regions conserved in the archaeo-eukaryotic lineage are shown by red rectangles, whereas those conserved in all bacteria are shown by yellow rectangles. The proteins are not shown to scale. Splits indicated by arrows and slashes are instances when different portions of the respective subunits are encoded in separate genes.

been detected primarily in eukaryotic chromatin-associated proteins, particularly transcription factors, where they function as accessory DNA-binding domains [74]. The ubiquity of the AT-hook-like modules in the β' subunits of DDRPs suggests that they were recruited for nucleic acid recognition in LUCA if not earlier.

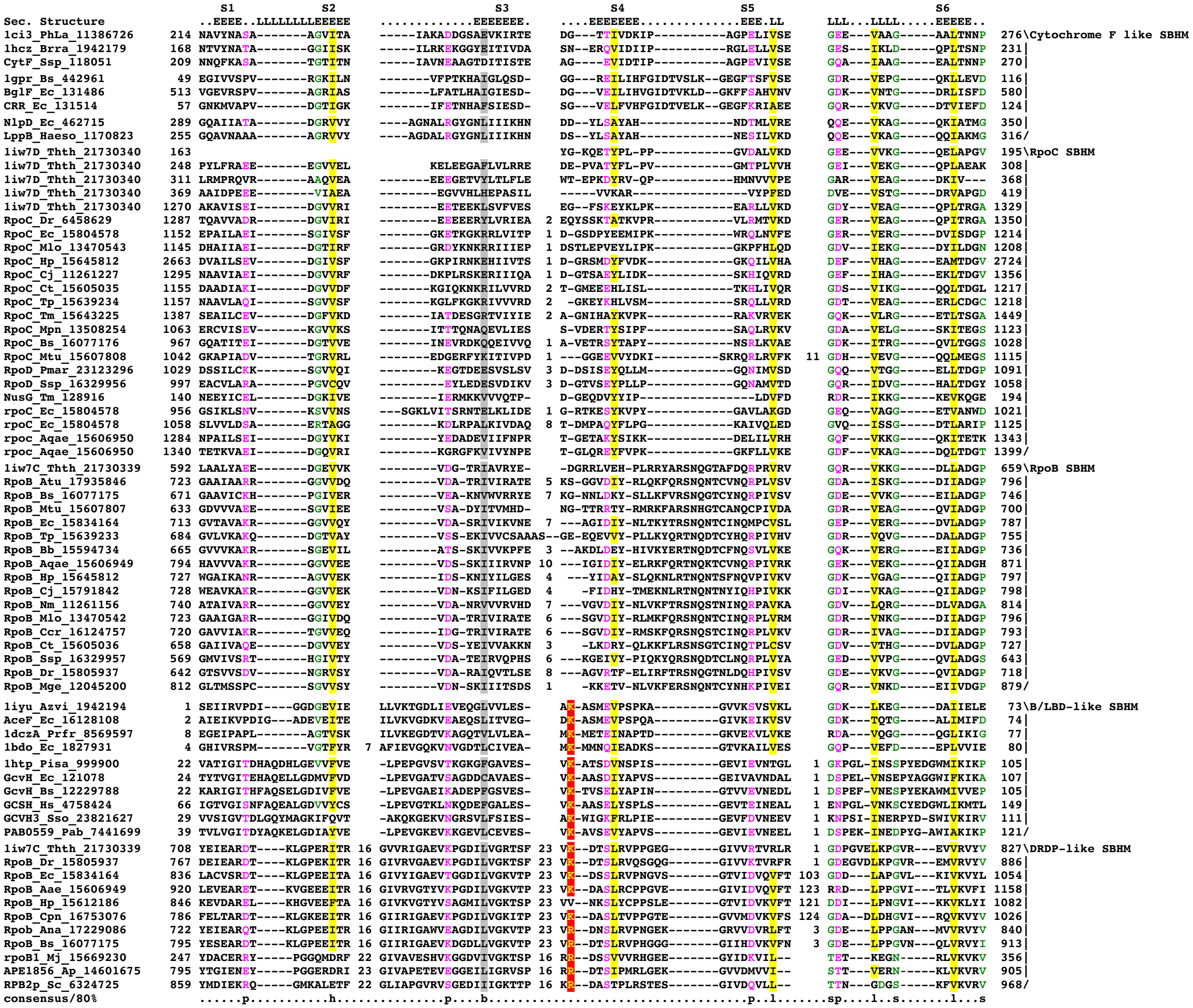
Analysis of a β-strand-rich insert between strands 2 and 3 of the β subunit DPBB (Figs. 4, 6) led to the identification of another ancient conserved domain that could be traced back to ancestral RNA polymerases. Visual examination of the structure of this insert domain revealed a topology with elements of both a sandwich and a barrel, and two "waist"-like structures that are characteristic of the sandwich-barrel hybrid motif (SBHM) fold [75,76] (see also the SCOP database [65]). The SBHM fold is present in biotin/lipoate carrier domains and their homologs found in a variety of enzymes and transporters [76]. A DALI search with this insert domain from the β subunit of *Thermus* DDRP showed similarity to the SBHMs from various biotin/lipoate-binding enzymes, such as acetyl-CoA carboxylase (1bdo) or dihydrolipoamide acetyltransferase (1iyu) with Z scores of ~6 and RMSD ~2.5 Å in the aligned regions spanning approximately 70 residues. An alignment of the SBHM domains from the DDRP β subunit with the previously identified SBHMs showed conservation of most residues characteristic of the SBHM superfamily (Fig. 7). The defining structural feature of this domain is seen in the loop between strands 5 and 6 (Figs. 7,8). This loop is bounded by small residues at each end (mostly valine at the N-terminus and glycine at the C-terminus) and has a distinct signature, with small (mostly glycine), charged (mostly aspartate) and hydrophobic residues occurring in succession, in the middle of the loop (Figs. 7,8). These observations reveal an evolutionary relationship between the SBHM domain of DDRP and the SBHMs found in other, functionally distinct proteins.

The core SBHM domain consists of a repeat of two three-stranded units with the characteristic "waist"-like structure occurring between the second and third strands of each repeat (Fig. 8). In the complete SBHM, strand 1 packs with strand 6 and the loops bounded by the two "waist"-like regions adopt an extended conformation resulting in the formation of a barrel. In the crystal structure of DDRP, the SBHM domain inserted into the β-subunit DPBB is in contact with the outward-projecting insert between strands 2 and 3 of the β'-subunit DPBB. Furthermore, the SBHM also contributes to the stabilization of the catalytic site by forming an interface between the two DPBBs (Fig. 4). The universal presence of the SBHM inserted in the DPBB domain of the DDRP β-subunit suggests that this domain was already present in the common ancestor of all extant DDRPs. The wide spread in different phylogenetic lineages of two other SBHM domains, namely those found in dihydrolipoamide acetyltransferase and the glycine-cleaving enzyme, is indicative of their presence in LUCA [76]. The SBHM of the DDRPs is specifically related to the biotin/lipoate-binding SBHMs. The two share a conserved basic residue that is present immediately prior to strand 4. In all biotin/lipoate-binding SBHMs, this position is occupied by a lysine, which covalently binds organic radical ligands (Fig. 7), but there is no evidence for such a role in the DDRP β subunits. These observations suggest that the ancestral SBHM interacted with various ligands with a low specificity. From such a precursor, which existed prior to LUCA, the SBHM domain apparently diversified into two forms, one that covalently bound organic radicals, and another that specialized in non-covalent interactions with proteins or nucleic acids. This latter form was recruited into the RNA polymerase catalytic core where it provided additional surfaces for the interaction between the DPBB domains and formed part of the interface between the catalytic domains. Thus, it appears likely that the ancestral DDRP evolved from the coalescence of at least 3 distinct domains: i) the DPBB that probably originally bound RNA and, subsequently, segregated into the cores of the β' subunit, where it acquired the metal-coordinating active site, and of the β subunit, with the two lysines projecting into the catalytic cleft (Fig. 4), ii) the Zn-ribbon, which was probably involved in interactions with the template, and iii) the SBHM domain that stabilized the interactions between the DPBB domains forming the catalytic cleft (Fig. 6). Additionally, two AT-hook like modules, which probably had an accessory DNA-binding function, were also inserted into a conserved α-helix directly N-terminal of the DPBB domain in the β' subunit.

The remaining conserved regions in the β and β' subunits of the DDRPs do not show detectable relationship with any other ancient conserved domains. Examination of these structures shows that they are principally composed of large α-helical hairpins that form coiled coil structures or stretches with successive shorter α-helices. Some of the other conserved regions consist of β-meanders with unusually long β-strands. These regions probably emerged, respectively, through duplication and divergence of a simple α-helical unit or through stabilization of long loops by selection for hydrophobic residues, resulting in the formation of the core units of the β-meanders.

Proliferation of the Sandwich-Barrel Hybrid Motif and lineage-specific innovations of the large subunits of bacterial DDRP

Further visual examination and structural alignments revealed the presence of one additional SBHM domain in the β subunit (1iw7 chain C: region 592–659) and several repeats of this domain in the β' subunit (1iw7 chain D, regions: 163–195, 248–308, 311–368, 369–419 and 1270–



```
                                S1           S2                       S3               S4                  S5                S6
Sec. Structure              ..EEEE..LLLLLLLLEEEEE    .............EEEEEEE.     ....EEEEEEE.............EEE.LL      LLL...LLLL.....EEEEE..
1ci3_PhLa_11386726      214 NAVYNASA-------AGVITA    -----IAKADDGSAEVKIRTE     DG---TTIVDKIP----------AGPELIVSE    GEE---VAAG-----AALTNNP  276\Cytochrome F like SBHM
1hcz_Brra_1942179       168 NTVYNATA-------GGIISK    -----ILRKEKGGYEITIVDA     SN--ERQVIDIIP----------RGLELLVSE    GES---IKLD-----QPLTSNP  231|
CytF_Ssp_118051         209 NNQFKASA-------TGTITN    -----IAVNEAAGTDITISTE     AG----EVIDTIP----------AGPEVIVSE    GQA---IAAG-----EALTNNP  270|
1gpr_Bs_442961           49 EGIVVSPV-------RGKILN    -------VFPTKHAIGLQSD-     GG---REILIHFGIDTVSLK--GEGFTSFVSE    GDR---VEPG-----QKLLEVD  116|
BglF_Ec_131486          513 VGEVRSPV-------AGRIAS    -------LFATLHAIGIESD-     DG---VEILIHVGIDTVKLD--GKFFSAHVNV    GDK---VNTG-----DRLISFD  580|
CRR_Ec_131514            57 GNKMVAPV-------DGTIGK    -------IFETNHAFSIESD-     SG---VELFVHFGIDTVELK--GEGFKRIAEE    GQR---VKVG-----DTVIEFD  124|
NlpD_Ec_462715          289 GQAIIATA-------DGRVVY    ----AGNALRGYGNLIIIKHN     DD--YLSAYAH-------------NDTMLVRE    QQE---VKAG-----QKIATMG  350|
LppB_Haeso_1170823      255 GQAVNAAA-------AGRVVY    ----AGDALRGYGNLIIIKHN     DS--YLSAYAH-------------NESILVKD    QQE---VKAG-----QQIAKMG  316/
1iw7D_Thth_21730340     163                                                    YG-KQETYPL-PP-----------GVDALVKD    GEE---VVKG-----QELAPGV  195\RpoC SBHM
1iw7D_Thth_21730340     248 PYLFRAEE-------EGVVEL           KELEEGAFLVLRRE     DE-PVATYFL-PV-----------GMTPLVVH    GEI---VEKG-----QPLAEAK  308|
1iw7D_Thth_21730340     311 LRMPRQVR-------AAQVEA    -------EEEGETVYLTLFLE     WT-EPKDYRV-QP-----------HMNVVVPE    GAR---VEAG-----DKIV---  368|
1iw7D_Thth_21730340     369 AAIDPEEE--------VIAEA    --------EGVVHLHEPASIL     -----VVKAR----------------VYPFED    DVE---VSTG-----DRVAPGD  419|
1iw7D_Thth_21730340    1270 AKAVISEI-------DGVVRI    --------EETEEKLSVFVES     EG--FSKEYKLPK-----------EARLLVKD    GDY---VEAG-----QPLTRGA 1329|
RpoC_Dr_6458629        1287 TQAVVADR-------DGVIRI    --------EEEEERYLVRIEA   2 EQYSSKTATKVPR-----------VLRMTVKD    GER---VEAG-----QPITRGA 1350|
RpoC_Ec_15804578       1152 EPAILAEI-------SGIVSF    -------GKETKGKRRLVITP   1 D-GSDPYEEMIPK-----------WRQLNVFE    GER---VERG-----DVISDGP 1214|
RpoC_Mlo_13470543      1145 DHAIIAEI-------DGTIRF    -------GRDYKNKRRIIIEP   1 DSTLEPVEYLIPK-----------GKPFHLQD    GDV---IEKG-----DYILDGN 1208|
RpoC_Hp_15645812       2663 DVAILSEV-------DGIVSF    -------GKPIRNKEHIIVTS   1 D-GRSMDYFVDK------------GKQILVHA    DEF---VHAG-----EAMTDGV 2724|
RpoC_Cj_11261227       1295 NAAVIAEI-------DGVVRF    -------DKPLRSKERIIIQA   1 D-GTSAEYLIDK------------SKHIQVRD    GEF---IHAG-----EKLTDGV 1356|
RpoC_Ct_15605035       1155 DAADIAKI-------DGVVDF    -------KGIQKNKRILVVRD   2 T-GMEEEHLISL------------TKHLIVQR    GDS---VIKG-----QQLTDGL 1217|
RpoC_Tp_15639234       1157 NAAVLAQI-------SGVVSF    -------KGLFKGKRIVVVRD   2  -GKEYKHLVSM------------SRQLLVRD    GDT---VEAG-----ERLCDGC 1218|
RpoC_Tm_15643225       1387 SEAILCEV-------DGFVKD    -------IATDESGRTVIYIE   2 A-GNIHAYKVPK------------RAKVRVEK    GQK---VLRG-----ETLTSGA 1449|
RpoC_Mpn_13508254      1063 ERCVISEV-------KGVVKS    -------ITTTQNAQEVLIES     S-VDERTYSIPF------------SAQLRVKV    GDA---VELG-----SKITEGS 1123|
RpoC_Bs_16077176        967 GQATITEI-------DGTVVE    -------INEVRDKQQEIVVQ   1 A-VETRSYTAPY------------NSRLKVAE    GDK---ITRG-----QVLTGGS 1028|
RpoC_Mtu_15607808      1042 GKAPIADV-------TGRVRL    -------EDGERFYKITIVPD   1 ---GGEEVVYDKI---------SKRQRLRVFK 11 GDH---VEVG-----QQLMEGS 1115|
RpoD_Pmar_23123296     1029 DSSILCKK-------SGVVQI    --------KEGTDEESVSLSV   3 D-DSISEYQLLM------------GQNIMVSD    GQQ---VTGG-----ELLTDGP 1091|
RpoD_Ssp_16329956       997 EACVLARA-------PGVCQV    ---------EYLEDESVDIKV   3 D-GTVSEYPLLP------------GQNAMVTD    GQR---IDVG-----HALTDGY 1058|
NusG_Tm_128916          140 NEEYICEL-------DGKIVE    --------IERMKKVVVQTP-     D-GEQDVYYIP----------------LDVFD    RDR---IKKG-----KEVKQGE  194|
rpoC_Ec_15804578        956 GSIKLSNV-------KSVVNS    ---SGKLVITSRNTELKLIDE   1 G-RTKESYKVPY------------GAVLAKGD    GEQ---VAGG-----ETVANWD 1021|
rpoC_Ec_15804578       1058 SLVVLDSA-------ERTAGG    --------KDLRPALKIVDAQ   8 T-DMPAQYFLPG------------KAIVQLED    GVQ---ISSG-----DTLARIP 1125|
rpoc_Aqae_15606950     1284 NPAILSEI-------DGYVKI    --------YEDADEVIIFNPR     T-GETAKYSIKK------------DELILVRH    GQF---VKKG-----QKITETK 1343|
rpoc_Aqae_15606950     1340 TETKVAEI-------DGQVRI    --------KGRGFKVIVYNPE     T-GLQREYFVPK------------GKFLLVKE    GDF---VKAG-----DQLTDGT 1399/
1iw7C_Thth_21730339     592 LAALYAEE-------DGEVVK    --------VDG-TRIAVRYE-     -DGRRLVEH-PLRRYARSNQGTAFDQRPRVRV    GQR---VKKG-----DLLADGP  659\RpoB SBHM
RpoB_Atu_17935846       723 GAAIAARR-------GGVVDQ    --------VDA-TRIVIRATE   5 KS-GGVDIY-RLQKFQRSNQNTCVNQRPLVSV    GDA---ISKG-----DIIADGP  796|
RpoB_Bs_16077175        671 GAAVICKH-------PGIVER    --------VEA-KNVWVRRYE   7 KG-NNLDKY-SLLKFVRSNQGTCYNQRPIVSV    GDE---VVKG-----EILADGP  746|
RpoB_Mtu_15607807       633 GDVVVAEE-------SGVIEE    --------VSA-DYITVMHD-     NG-TTRRTY-RMRKFARSNHGTCANQCPIVDA    GDR---VEAG-----QVIADGP  700|
RpoB_Ec_15834164        713 GVTAVAKR-------GGVVQY    --------VDA-SRIVIKVNE   7 ---AGIDIY-NLTKYTRSNQNTCINQMPCVSL    GEP---VERG-----DVLADGP  787|
RpoB_Tp_15639233        684 GVLVKAKQ-------DGTVAY    --------VSS-EKIVVCSAAAS--GE-EQEVVY-PLLKYQRTNQDTCYHQRPIVHV    GDR---VQVG-----DALADGP  755|
RpoB_Bb_15594734        665 GVVVKAKR-------SGEVIL    --------ATS-SKIVVKPFE   3 --AKDLDEY-HIVKYERTNQDTCFNQSVLVKE    GQK---VERG-----EIIADGP  736|
RpoB_Aqae_15606949      794 HAVVVAKR-------GGVVEE    --------VDS-SKIIIRVNP  10 ---IGIDIY-ELRKFQRTNQKTCVNQRPIVRK    GEK---VEKG-----QIIADGH  871|
RpoB_Hp_15645812        727 WGAIKANR-------AGVVEK    --------IDS-KNIYILGES   4  ---YIDAY-SLQKNLRTNQNTSFNQVPIVKV    GDK---VGAG-----QIIADGP  797|
RpoB_Cj_15791842        728 WEAVKAKR-------GGVVEK    --------VDN-KSIFILGED   4     -FIDHY-TMEKNLRTNQNTNYIQHPIVKK    GDI---VKAG-----QIIADGP  798|
RpoB_Nm_11261156        740 ATAIVARR-------GGVVEY    --------VDA-NRVVVRVHD   7 ---VGVDIY-NLVKFTRSNQSTNINQRPAVKA    GDV---LQRG-----DLVADGA  814|
RpoB_Mlo_13470542       723 GAAIGARR-------GGIVDQ    --------VDA-TRIVIRATE   6 ---SGVDIY-RLMKFQRSNQNTCINQRPLVRM    GDR---VNKG-----DIIADGP  796|
RpoB_Ccr_16124757       720 GAVVIAKR-------TGVVEQ    --------IDG-TRIVIRATE   6 ---SGVDIY-RMSKFQRSNQSTCINQRPLVKV    GDR---IVAG-----DIIADGP  793|
RpoB_Ct_15605036        658 GAIIVAQE-------DGVVEY    --------VDS-YEIVVAKKN   3    -LKDRY-QLKKFLRSNSGTCINQTPLCSV    GDV---VTHG-----DVLADGP  727|
RpoB_Ssp_16329957       569 GMVIVSRT-------HGIVTY    --------VDA-TEIRVQPHS   6 ---KGEIVY-PIQKYQRSNQDTCLNQRPLVYA    GED---VVPG-----QVLADGS  643|
RpoB_Dr_15805937        642 GTSVVSDV-------NGRVSY    --------VDA-RAIQVTLSE   8 ---AGVRTF-ELIRFTRSNQGTNLDQHPIVSV    GDE---VKVG-----QVIADGP  718|
RpoB_Mge_12045200       812 GLTMSSPC-------SGVVSY    --------VDN-SKIIITSDS   1    -KKETV-NLVKFERSNQNTCYNHKPIVEI    GQR---VNKD-----EIIVDGP  879/
1iyu_Azvi_1942194         1 SEIIRVPDI----GGDGEVIE    LLVKTGDLIEVEQGLVVLES-     AK-ASMEVPSPKA---------GVVKSVSVKL    GDK---LKEG-----DAIIELE   73\B/LBD-like SBHM
AceF_Ec_16128108          2 AIEIKVPDIG---ADEVEITE    ILVKVGDKVEAEQSLITVEG-     DK-ASMEVPSPQA---------GIVKEIKVSV    GDK---TQTG-----ALIMIFD   74|
1dczA_Prfr_8569597        8 EGEIPAPL-------AGTVSK    ILVKEGDTVKAGQTVLVLEA-     MK-METEINAPTD---------GKVEKVLVKE    RDA---VQGG-----QGLIKIG   77|
1bdo_Ec_1827931           4 GHIVRSPM-------VGTFYR  7 AFIEVGQKVNVGDTLCIVEA-     MK-MMNQIEADKS---------GTVKAILVES    GQP---VEFD-----EPLVVIE   80|
1htp_Pisa_999900         22 VATIGITDHAQDHLGEVVFVE    -LPEPGVSVTKGKGFGAVES-     VK-ATSDVNSPIS---------GEVIEVNTGL  1 GKPGL-INSSPYEDGWMIKIKP  105|
GcvH_Ec_121078           24 TYTVGITEHAQELLGDMVFVD    -LPEVGATVSAGDDCAVAES-     VK-AASDIYAPVS---------GEIVAVNDAL  1 DSPEL-VNSEPYAGGWIFKIKA  107|
GcvH_Bs_12229788         22 KARIGITHFAQSELGDIVFVE    -LPEVGAEIKADEPFGSVES-     VK-TVSELYAPIN---------GTVVEVNEDL  1 DSPEF-VNESPYEKAWMIVVEP  105|
GCSH_Hs_4758424          66 IGTVGISNFAQEALGDVVYCS    -LPEVGTKLNKQDEFGALES-     VK-AASELYSPLS---------GEVTEINEAL  1 ENPGL-VNKSCYEDGWLIKMTL  149|
GCVH3_Sso_23821627       29 VVSIGVTDLGQYMAGKIFQVT    -AKQKGEKVNGRSVLFSIES-     AK-WIGKFRLPIE---------GEVFDVNEEV  1 KNPSI-INERPYD-SWIVKIRV  111|
PAB0559_Pab_7441699      39 TVLVGITDYAQKELGDIAYVE    -LPEVGKEVKKGEVLCEVES-     VK-AVSEVYAPVS---------GEVIEVNEEL  1 DSPEK-INEDPYG-AWIAKIKP  121/
1iw7C_Thth_21730339     708 YEIEARDT----KLGPERITR 16 GVVRIGAEVKPGDILVGRTSF 23 VK--DTSLRVPPGEG-------GIVVRTVRLR  1 GDPGVELKPGVR---EVVRVYV  827\DRDP-like SBHM
RpoB_Dr_15805937        767 DEIEARDT----KLGPEKITR 16 GIVRVGAEVKPGDILVGKTSF 23 VK--DTSLRVQSGQG-------GIVVKTVRFR  1 GDEGVDLKPGVR---EMVRVYV  886|
RpoB_Ec_15834164        836 LACVSRDT----KLGPEEITA 16 GIVYIGAEVTGGDILVGKVTP 23 VK--DSSLRVPNGVS-------GTVIDVQVFT 103 GDD---LAPGVL---KIVKVYL 1054|
RpoB_Aae_15606949       920 LEVEARET----KVGEEEITR 16 GIVRVGTYVKPGDILVGKVTP 23 VK--DASLRCPPGVE-------GIVIDVQVFT 123 RRD---LPPGVI---TLVKVFI 1158|
RpoB_Hp_15612186        846 KEVDAREL----KHGVEEFTA 16 GIVKVGTYVSAGMILVGKTSP 23 VV--NKSLYCPPSLE-------GTVIDVKVFT 121 DDI---LPNGVI---KKVKLYI 1082|
RpoB_Cpn_16753076       786 FELTARDT----KLGKEEITR 16 GIIRIGAEVKPGDILVGKITP 23 VK--DASLTVPPGTE-------GVVMDVKVFS 124 GDA--DLDHGVI---RQVKVYV 1026|
Rpob_Ana_17229086       722 YEIEARQT----KLGPEEITR 16 GIIRIGAWVEAGDILVGKVTP 23 VR--DNSLRVPNGEK-------GRVVDVRLFT   3 GDE---LPPGAN---MVVRVYV  840|
RpoB_Bs_16077175        795 YESEARDT----KLGPEEITR 16 GIIRIGAEVKDGDLLVGKVTP 23 VR--DTSLRVPHGGG-------GIIHDVKVFN   3 GDE---LPPGVN---QLVRVYI  913|
rpoB1_Mj_15669230       247 YDACERRY-----PGGQMDRF 22 GIVAVESHVKGGDVIVGKTSP 16 RR--DSSVVVRHGEE-------GYIDKVIL--     TET---KEGN-----RLVKVKV  356|
APE1856_Ap_14601675     795 YTGIENEY-----PGGERDRI 23 GIVAPETEVEGGEILIGRVSP 16 RR--DTSIPMRLGEK-------GVVDMVVI--     STT---VERN-----KLVKVRV  905|
RPB2p_Sc_6324725        859 YMDIEKRQ-----GMKALETF 22 GLIAPGVRVSGEDIIIGKTTP 16 KR--DASTPLRSTES-------GIVDQVLL--     TTN---GDGS-----KFVKVRV  968/
consensus/80%               ......p..........h.............p....b.................................p..l.......sp....l..s.......l...s
```

Figure 7

A structure-based multiple alignment of the sandwich-barrel hybrid motif (SBHM) domains from β and β' subunits of DDRPs and other proteins. The alignments were generated by structural superposition of representatives of the SBHM domains followed by addition of sequences neighbors of the representative structures. The conserved basic residue that forms a covalent linkage with the organic radical in biotin/lipoate-binding domain-type SBHM is shown in reverse shading. The individual sequence families of SBHMs are indicated to the right of the alignment. The waist-like loops are marked as 'L' in the secondary structure shown above the alignment. Species abbreviations are as in Fig. 1 and Fig. 2 Additional species abbreviations not given above are: Ana, *Anabaena* sp., Azvi, *Azotobacter vinelandii*, Bb, *Borrelia burgdorferi*, Brra, *Brassica rapa*, Ccr,*Caulobacter crescentus*, Cj, *Campylobacter jejuni*, Cpn, *Chlamydophila pneumoniae*, Haeso, *Haemophilus somnus*, Mge, *Mycoplasma genitalium*, Mj, *Methanocaldococcus jannaschii*, Nm, *Neisseria meningitidis*, Pab, *Pyrococcus abyssi*, PhLa, *Phormidium laminosum*, Pisa, *Pisum sativum*, Pmar, *Prochlorococcus marinus*, Prfr, *Propionibacterium freudenreichii*

1329) of the bacterial DDRPs (Fig. 6). A PSI-BLAST search started with the N-terminal SBHM from the *Aquifex aeolicus* DDRP β' subunit detected several SBHMs from proteins other than the β' subunit, such as acetylornithine deacetylase (E = 2 × 10^{-7}, iteration 2), the C-terminal region of bacterial cytochrome F (E = 10^{-4}, iteration 2), biotin carboxyl carrier domain of biotin transcarboxylases (E = 6 × 10^{-4}, iteration 2) and the phosphotransferase sys-

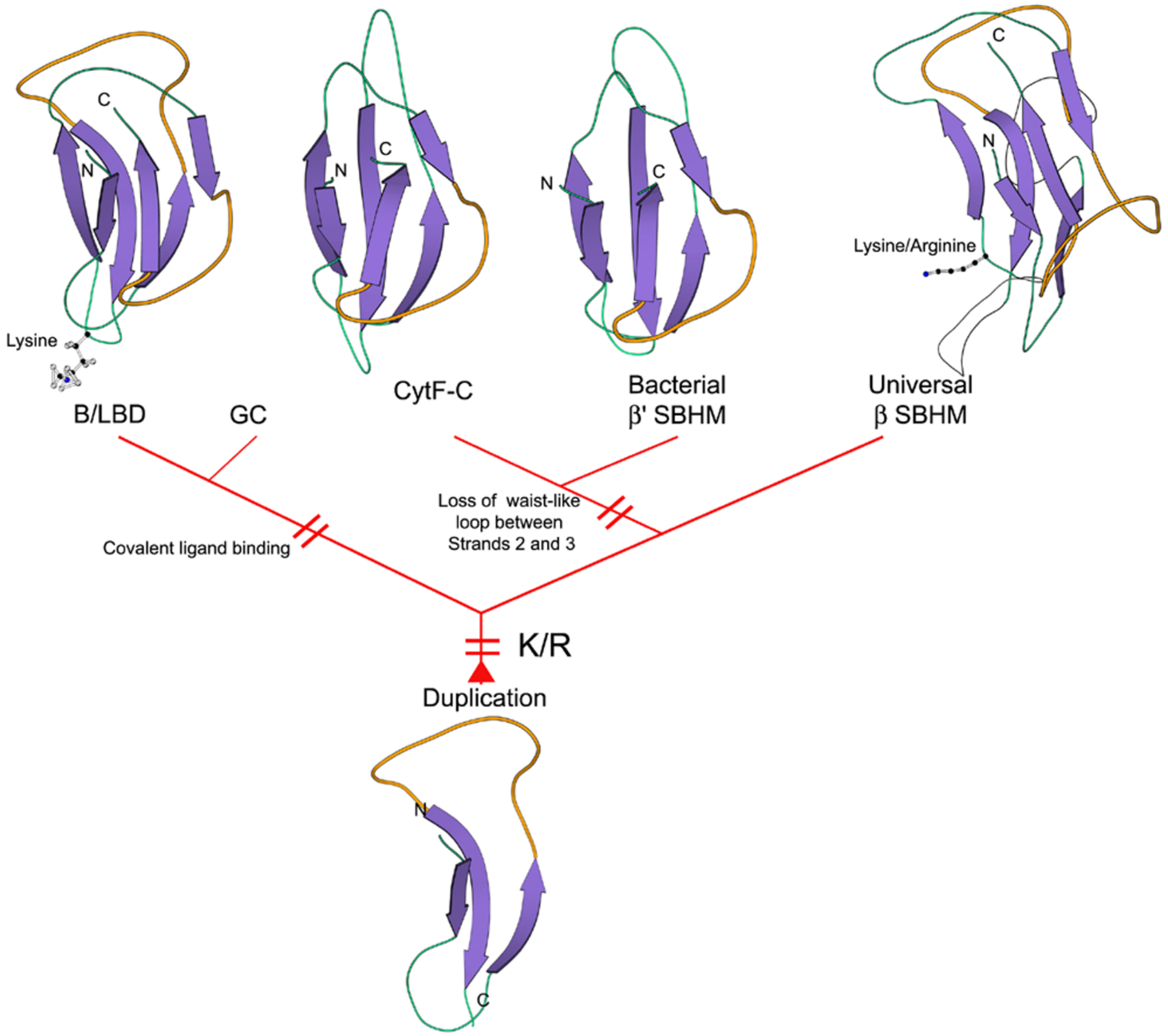


Figure 8
An evolutionary scenario for the SBHM domains from the DDRP subunits and other proteins. The scenario for the derivation of various versions of the SHBM fold from the simple, ancestral 3-stranded unit was inferred from phyletic patterns, structural features and the internal duplication. The conserved basic residue present in the biotin/lipoate-binding domain-type SBHMs and the β subunit of the DDRPs is shows as a ball-and-stick model. The emergence of different lineage-specific specializations is indicated to the side of each clade.

tem enzyme II (E = 10^{-3}, iteration 2). This search also retrieved the bacteria-specific C-terminal SBHM domain of the β subunit (e.g. *Thermotoga*, E ~10^{-3}) and, interestingly, also a SBHM in the N-terminal region of the NusG protein from *Thermotoga maritima* (E = 2 × 10^{-7}, iteration 2). Reverse searches with the C-terminal region of cytochrome F (gi: 11467387, residues 150–321) retrieved the bacterial β' sequences in the second and subsequent iterations with statistically significant E-values. None of these additional SBHMs were detectable in the archaeal or eukaryotic DDRPs, either in sequence searches or in direct comparisons of the crystal structures. Thus there appears to have been a lineage-specific proliferation and dissemination of the SBHM domains in bacteria.

Bacterial β and β' subunits show major variability in the distribution of the SBHM domains (Fig. 6). All bacterial β subunits share one additional SBHM, which is located N-terminal of the DPBB domain. Similarly, an extra SBHM, located C-terminal of the DPBB, is conserved in bacterial β' subunits. In some cyanobacteria, the portion of the β' subunit containing the SBHM occurs as a separate polypeptide. In addition to the SBHMs that are conserved among all bacterial lineages, a cassette of up to four SBHMs are present N-terminal of the DPBB domain in the β' subunit of *Thermus*, *Deinococcus* and *Thermotoga* (Fig. 6). Similarly, the β' subunit of proteobacteria, *Aquifex*, spirochaetes, cyanobacteria and chlamydiae contains an additional set of SBHMs C-terminal of the DPBB domain, preceding the C-terminal SBHM that is shared by all bacteria (Fig. 6). The bacterial-specific SBHM domains present in the β and β' subunits are typically located on the periphery of the holoenzyme, which makes them accessible for interactions with other proteins. Two of these SBHMs (regions 163–195 and 369–419 in 1iw7 chain D) are involved in the interaction with the σ 70 subunit, which is critical for transcription initiation in bacteria [23]. Thus, it appears likely that, originally, the SBHM domain of the β subunit participated in generic protein-protein interactions that allowed the ancestral DDRP to recruit accessory subunits. Subsequently, in the bacterial lineage, proliferation of SBHMs provided for several other, specific interactions involved in initiation and elongation.

Several of the SBHMs that are specific to the bacterial DDRPs show degradation of the "waist"-like loop of the first internal repeat. This is clearly a derived state because the SBHM originally evolved through the duplication of a single ancestral unit that contained an intact "waist"-like loop between strands 2 and 3. The form of the SBHM with a shortened loop in the first internal repeat is additionally seen in several bacterial-specific families of SBHMs. These include the SBHM of the phosphotransferase system enzyme II (glucose permease or PTS-EII) found in several lineages of free-living bacteria, the one at the C-termini of cytochrome F from cyanobacteria and chloroplasts, and the NlpD-like cell wall peptidases that are present in most bacteria. In addition to the specific structural relationship between some of the SBHMs of the DDRPs and those present in these bacterial proteins, they also show detectable sequence similarity. For example, a PSI-BLAST search initiated with the SBHM from NlpD retrieved the derived SBHMs of the bacterial RNA polymerase subunits (e.g. *Streptococcus* β subunit, E = 10^{-4}, iteration 2) well before the classical SBHMs (eg. *Mesorhizobium* dihydrolipoamide acetyltransferase, E = 10^{-3}, iteration 4). The β and β' subunits of the bacterial DDRP are the only proteins that contain both versions of the SBHM (Fig. 7). Hence, it appears likely that the form of the SBHM with the degraded loop in the first internal repeat initially evolved in the bacterial DDRP subunits. Subsequently, these domains appear to have been recruited for various interactions in other bacterial proteins.

Examination of the structure and multiple sequence alignments of bacterial β and β' subunits led to the identification of two additional, distinct domains, which so far are not detectable in any proteins other than the DDRP and were accordingly designated Beta-Beta' Module 1 (BBM1, region 119–165, 1iw7 chain D) and Beta-Beta' Module 2 (BBM2, region 1109–1190, 1iw7 chain D) (Fig. 6). Both these domains share an unusual, common structural core that we term the BBM-core (BBMC) of about 40–50 residues with an RMSD of 2Å over the aligned region. The BBMC consists of N- and C-terminal helices that bound a central region with 4 unusual, extended regions that fold into a curved hairpin structure (Fig. 9A). The BBM1 domain is inserted after the Zn-ribbon in all bacterial β' subunits and, additionally, into the SBHM domain of the β subunit in proteobacteria, *Aquifex* and chlamydiae (Fig. 6). Notably, this is the same set of bacteria that contain extra SBHM domains inserted in the C-terminal region of the β' subunit (Fig. 6 and see above). The BBM2 domain shows a similar pattern of insertion: the C-terminal portion of all bacterial β' subunit contains one copy, Thermotoga has two extra copies distal of the pan-bacterial copy. Proteobacteria, *Aquifex*, spirochetes and chlamydiae have a BBM2 inserted into the N-terminal region of the β subunit (Fig. 6). The functions of these previously unnoticed modules and the nature of the apparent congruence in the evolution of the two RDRP subunits in a subset of bacteria remain to be elucidated.

The RNA polymerase α subunit-core-related domain and other elaborations of the archaeo-eukaryotic RNA polymerase catalytic subunits

The bacterial DDRPs have two α-subunits that form a homodimer at the "base" of the β-β' dimer and stabilize the catalytic-binding cleft. In the archaeo-eukaryotic lineage, there is a duplication of the α-subunit, resulting in two paralogous subunits, RBP3 and RBP11. In the eukaryotic DDRP complex, these subunits occupy a spatial position very similar to that of the α-subunit homodimer of the bacterial DDRPs. All these proteins share a common α + β core domain, with two α-helices, which form the dimer interface. In the SCOP database, the α-subunit core domains have been classified together with several other domains, including the ligand-binding domain of the bacterial arginine repressor, the dimerization cofactor of the transcription factor HNF1 (DCOH), and a domain specific to the bacterial aspartyl-tRNA synthetases. To investigate this domain further, we conducted structural similarity searches of the PDB database using the DALI program, with the α-subunits of the bacterial and

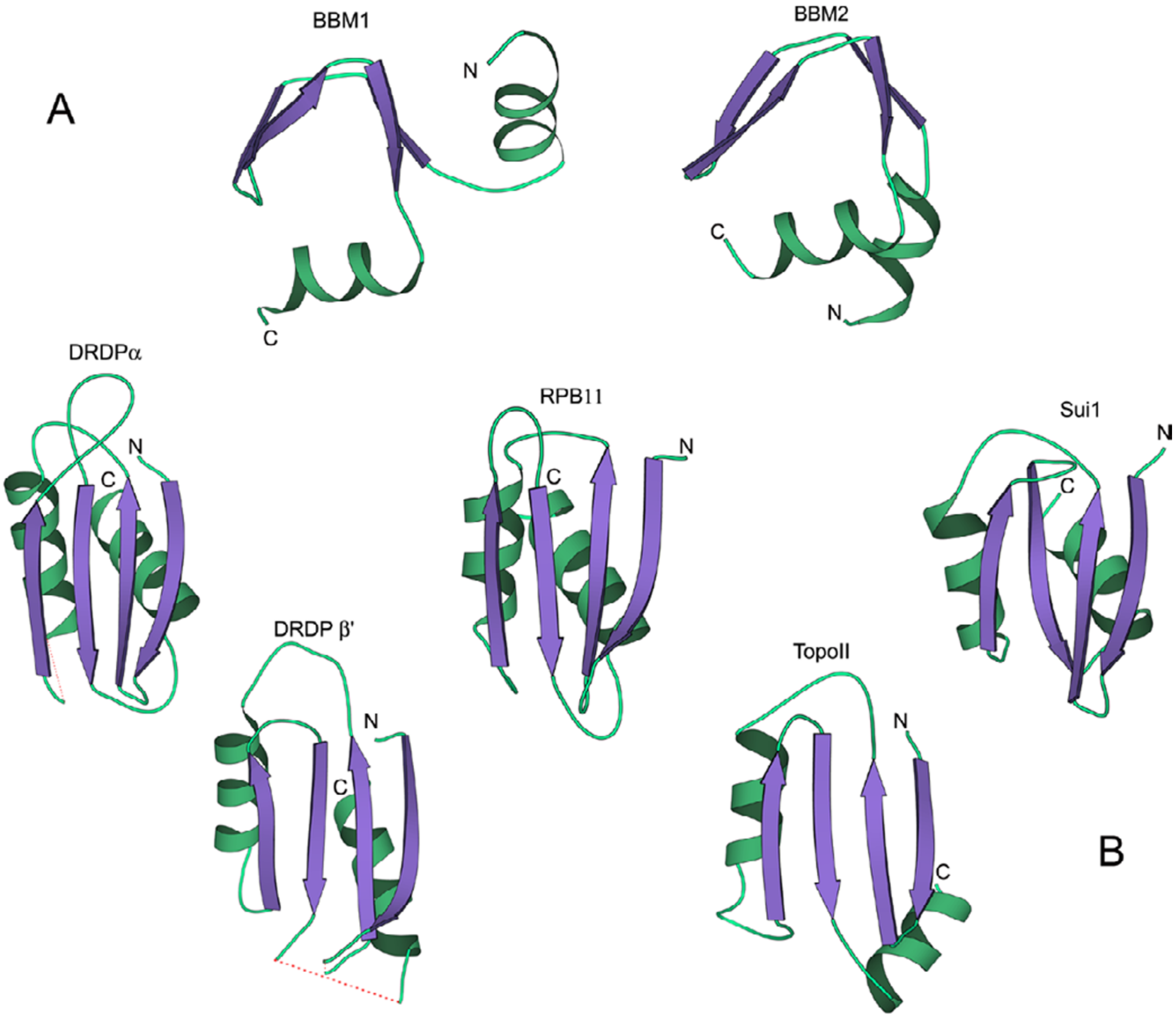


Figure 9
Structures of other conserved modules of DDRPs. (A) The core BBMC module of the BBM1 and BBM2 domains **(B)** Different versions of the α-subunit-core related (ASCR) domain ASCR domains from the bacterial DDRP α subunit, RBP11 (the eukaryotic ortholog of α), the β' subunit of the archaeo-eukaryotic DDRPs, topoisomerase II/gyrase globular domain 3, and the eukaryotic translation initiation factor 1 (Sui1) are shown. Inserts or regions of poor X-ray diffraction are shown with dotted lines. The ASCR domains from the archaeo-eukaryotic β' subunits are most similar to those from the α-subunits.

eukaryotic DDRPs used as the queries. These searches showed that the core domain of the α-subunits were most similar (Z scores 5.4–7, with RMSDs of 2.2–3 Å over the aligned C-α residues) to the arginine repressor ligand-binding domain [77] and several domains that were not previously recognized as being related (Fig. 9B). These newly detected domains included the SUI1 domain, that is found in eIF1 and related proteins [78], the third globular domain of DNA gyrase/Topoisomerase II and, most interestingly, a distinct globular domain in the yeast DDRP β' subunit (PDB: 1k83; region 1143–1271). Reciprocal DALI searches with these domains readily recovered the α-subunits of DDRP as the best hits. For example, the domain from the yeast β' subunit hit the α-subunits of bacterial DDRPs with Z-scores of 5–7 and RMSD of 2.8–3 Å. Hereinafter we refer to these domains as the α-subunit core-related (ASCR) domains. The ASCR fold has a β2-α-β2-α topology and forms a two-layered structure with a 4-stranded β-sheet and characteristic strand order of 1-2-4-3 (Fig. 9B). The ASCR domains of bacterial α-subunits and eukaryotic RBP3 share a common insert domain between helix-1 and strand 3, which apparently has been lost in the RBP11 subunit of the eukaryotic DDRPs.

The ASCR domain in the yeast β' subunit is inserted into another α + β domain. Visual examination of the topology and structural searches of the PDB database using the DALI program showed that this domain it adopts an RNA Recognition Motif (RRM)-like structure with a β-α-β2-α-β topology forming a 4-stranded β-sheet. These two domains comprise a conserved module that is specifically shared by the archaeo-eukaryotic DDRPs to the exclusion of the bacterial β' subunits (Fig. 6). This two-domain module is inserted precisely in the same region as the C-terminal SBHMs seen in the bacterial lineage (Fig. 6). The closest relatives of the ASCR domain detected in the β' subunit are those present in the α-subunits (Fig. 9B). This observation suggests that the primary event in the divergence of the archaeo-eukaryotic β' subunit was the insertion of an RRM-like domain into the ancestral core. This was followed by the insertion of the ASCR domain, which probably evolved as a result of a duplication of the RBP3/11 subunits, into this RRM-like domain. Both the spatial arrangement of the ASCR domain in the β' subunits and its extreme sequence divergence with respect to the ASCR domains of the α-subunit suggests that it has acquired a new function. The RRM-like domain and the ASCR domain are positioned in the exterior entrance to the catalytic cleft and could interact with cofactors or with nucleic acid. There is no evidence that the ASCR domain in the DDRP α-subunits binds nucleic acids and studies on SUI1 protein and the arginine repressor ligand-binding domain implicate the ASCR domain in protein-protein interactions and small molecule binding [77,79]. However, most of the ASCR domains are present in proteins that interact with nucleic acids (see above), raising the possibility of a role in non-specific nucleic acid binding.

Mirroring the insertion of the ASCR domain in the β' subunit, a distinct α + β globular domain, N-terminal to the DPBB domain (region 578–632 of 1K83, Chain B), is inserted in the β subunits of all archaeal and eukaryotic DDRPs (Fig. 6). This domain contains a 4-stranded sheet with a β2-α-β2 topology, with the strands in a 2-1-4-3 order. Visual examination of the topology and searches of the PDB database using the DALI program showed that this insert domain has the β-grasp fold [80,81]. For example, in the DALI search, the *Staphylococcus* immunoglobulin-binding protein (IGBP; PDB code 2igd) was detected with a Z score of 5.2 and a RMSD of 2.7 over ~50 residues. This insert domain in the archaeo-eukaryotic β subunits resembles the version of the β-grasp domain that is present in the bacterial translation initiation factor IF3 and the IGBPs, in contrast to the distinct version found in ubiquitins, ferredoxins and ThiS. Specifically, the β-subunit insert domain shares a 4-stranded core with the former group and, like these proteins, lacks a long insert between the two terminal strands (data not shown). Given that this domain lies in the periphery of the holoenzyme, it is likely to participate in interactions with cofactors, which is compatible with the protein-protein interaction function identified in several proteins with the β-grasp fold [80,81].

Thus, a common set of ancient conserved modules appears to have been repeatedly reused in the evolutionary diversification of the RNA polymerase subunits from their ancestral cores. The lineage-specific domain insertions in bacteria and in archaea-eukaryotes correlate with the fundamental differences between the basal transcription machinery of these two lineages [82,83]. This suggests congruent evolution of the RNA polymerase catalytic subunits and the basal transcription factors [68].

Origin of the RDRPs

The evidence of an evolutionary relationship between the core catalytic domains of the RDRPs and the DDRPs has implications for the origin of the former group of proteins. Given that DDRP is universal, whereas RDRP is eukaryote-specific, a plausible hypothesis might be that RDRP evolved at the onset of eukaryotic evolution through extensive, rapid sequence divergence following a duplication of the β' subunit of the ancestral DDRP. However, despite extensive comparisons using PSSM and careful visual examination, we failed to detect any residual sequence similarity between DDRP and RDRP, outside the core DPBB domain containing the metal-binding active site. In particular, no traces of the Zn-ribbon and the AT-hook-like modules, which are conserved in all β' subunits (Fig. 6), were detected. All RDRP sequences contain two closely spaced lysines, which are conserved in the YRH proteins (Fig. 1) and might functionally correspond to the conserved lysines of the β-type DPBB. These residues are embedded in a secondary structure context that is enriched in predicted β-strands (Fig. 1); hence, in principle, this region might correspond to the β-type DPBB of the DDRPs. However, the respective region of the RDRP sequences shows no detectable similarity to the to this DPBB domain, suggesting that the presence of two functionally important lysines in both RDRP and DDRP might be convergent. In principle, radical divergence could have eroded all similarity between RDRP and DDRP beyond the DPBB domain. An alternative, perhaps more likely explanation is that, while the core catalytic domains of the two polymerases evolved from the same ancestral DPBB, they subsequently accreted distinct sets of peripheral modules around this core. The bacteriophage YRH proteins show no specific relationship with the RDRPs from any one eukaryotic lineage, arguing against a recent lateral transfer from eukaryotes to the phages. Taken together, all these observations point to a scenario for the origin of the RDRP, under which the RDRPs and the YRHs diverged from the DDRPs at a stage of evolution that succeeded the divergence of the DPBB domains of the β and β' subunits (Fig. 5) but preceded the emergence of eukaryotes and even LUCA. One class of substrates of the extant eukaryotic RDRPs are small RNAs, such as microRNAs and stRNAs [84–90], and analogous small RNAs have been identified in bacteria [91–94]. Given that small RNAs are likely to have been abundant in the hypothetical ancient RNA world, RDRPs might have been involved in the replication of such RNAs since an early stage of evolution. Should that be the case, the YRH proteins could be late surviving derivatives of these ancient enzymes, which apparently have been shunned from cellular life forms during the evolution of prokaryotes. Eukaryotes might have acquired the YRH gene from a bacteriophage at an early stage of their evolution and retained the RDRPs thanks to the selective advantage conferred by the new level of gene expression regulation mediated, in part, by these enzymes.

The eukaryotic RDRP is part of a large ensemble of functionally linked proteins, which include the Dicer helicase-nuclease, PIWI family proteins, the Hen-2/Corymbosa-like RNA methylase, AlkB-related RNA demethylases, and Lin28-like RNA-binding proteins [1,45,46]. Since no closely related homologs of any of these proteins are encoded in the genomes of the phages that encode YRH proteins, the latter are unlikely to function in a biological context similar to that of the eukaryotic RDRPs. Nevertheless, the extended similarity between the YRH and RDRP sequences and the presence of potential nucleic-acid-binding modules, such as a C-terminal Zn-ribbon in the SpβC2 YonO protein, suggest that the YRH proteins also have a RNA polymerase activity. Possible functions of the YRH proteins include a regulatory activity via amplification of antisense RNAs or perhaps a more conventional role as a DNA-dependent RNA polymerase of the respective phages.

General discussion and conclusions

Here, we identified bacteriophage homologs of the eukaryotic RDRP, which is involved in post-transcriptional gene silencing. We also present evidence that RDRP and the ubiquitous β' subunit of DDRP contain a homologous, metal-coordinating, catalytic domain. This domain adopts the DPBB fold, in which the signature DbDGD motif, shared by RDRP and DDRP, is ensconced in the insert between the ultimate and penultimate β-strands. A second DPBB domain is present in the β subunit of the DDRPs. This version of the DPBB domain lacks the metal-coordinating motif but contributes two lysine residues, which are critical for substrate interactions in the catalytic cleft. The DPBB domain in the β subunit is distorted by large inserts and shares no detectable sequence similarity with the DPBB from the β' subunit. However, the DPBBs from the two DDRP subunits are spatially juxtaposed in the DDRP crystal, with the catalytic cleft located between them. This arrangement suggests that the primordial RNA polymerase was a head to tail homodimer of a DPBB domain; a duplication of this primordial DPBB domain probably gave rise to the two DPBBs of modern DDRPs. The presence of completely different conserved residues and the asymmetric head to tail arrangement of the monomers in the pair suggest that the ancestral DPBB dimer was not specialized, had no catalytic activity, and merely bound RNA. Originally, these DPBB domains might have functioned as protein cofactors that stabilized a ribozyme RNA polymerase and eventually displaced the ribozyme as they acquired key residues required for protein-based polymerase activity. This evolutionary scenario complements a similar model proposed for the evolution of the other major polymerase class, the palm-domain-containing RNA and DNA polymerases, whose primordial core apparently consisted of a RNA-binding domain of the RRM fold [11].

The β and β' subunits of DDRP additionally contain a Zn-ribbon and a SBHM domain respectively. The former is a widespread nucleic-acid-binding domain, whereas the latter functions in a variety of biochemical contexts as a small-molecule-binding and protein-protein interaction domain. Analysis of the conserved protein superfamilies within these folds and their phyletic distribution patterns suggests that diverse versions of these domains with distinct functions were already present in LUCA. Thus, several duplication-divergence events apparently preceded the emergence of the forms of these domains that are seen in the catalytic core of DDRPs (e.g., Fig. 5). The ancestral versions of these domains would not have the adaptations for specific roles possessed by their descendant forms, such as those in extant DDRP or RDRP, and probably functioned as generic RNA-binding and protein-protein interaction domains, with the specificity conferred by catalytic RNA molecules. Hence, considerable diversity of protein domains appears to have emerged prior to the "crystallization" [95] of a transcription machinery similar to the one that operates in modern cells and probably had been already in place in LUCA [96]. Similar conclusions regarding the early stages of protein domain evolution have been reached previously as a result of evolutionary analysis of proteins involved in translation, such as aminoacyl-tRNA synthetases and GTPases: substantial diversification of protein domains had occurred prior to the "crystallization" of elaborate, modern-type translation machinery [66,97].

Although the major events in the evolution of polymerases occurred at very early stages of evolution and LUCA already had an advanced DDRP resembling the extant forms, subsequent lineage-specific elaborations of considerable magnitude have occurred. The most notable of these is the proliferation and dissemination of SBHM and the BBM domains in the bacterial β and β' subunits. In the archaeo-eukaryotic lineage, another domain, apparently derived through duplication of the dimerization domain of the DDRP α-subunit, was inserted into the C-terminal region of the β' subunit. Additionally, the Zn-ribbon underwent duplication and elaboration in the β' subunit and was also recruited in various transcription factors. Similarly in the archaeo-eukaryotic β subunit, a minimal β-grasp domain was added to the conserved core.

While we provide evidence that the catalytic cores of the RDRPs, their bacteriophage homologs and the DDRPs evolved from a common ancestor, the provenance of the rest of the conserved RDRP module and, accordingly, the evolutionary scenario for RDRP remain less clear. No similarity to DDRPs was detected in the RDRP sequences beyond the core DPBB domain. One possibility is that the RDRP module has emerged at the onset of the evolution of eukaryotes through extreme sequence divergence following a duplication of the DDRP β' subunit. An alternative scenario holds that the RDRPs are ancient enzymes that diverged from the evolutionary precursor of DDRP at a very early, pre-LUCA stage of evolution, albeit after the duplication that led to the differentiation of the β and β' versions of the DPBB domain. In fact, it appears likely that the ancestral protein RNA polymerase, at the time of the differentiation of the β-type and β'-type DPBB domains, functioned as an RDRP because DNA probably had not evolved by that stage [96]. The switch of the ancestral enzyme to the DDRP activity and the emergence of the evolutionary precursor of RDRP might be tentatively linked to one another and associated with the advent of DNA. This view of the evolutionary history of RDRPs is compatible with the presence of RDRP homologs in phages and with their role in replicating microRNAs (potential relics of the RNA world) during PTGS in eukaryotes. Subsequently, this enzyme apparently has been expunged from the cellular RNA synthesis systems and survived only in some parasitic elements, such as bacteriophages, through which it might have been reintroduced into the genome of an ancestral eukaryote. This scenario shows parallels to the probable evolutionary history of another major, unrelated class of polymerases, the RDRPs and reverse transcriptases containing the palm domain. Among extant biological entities, these polymerases are encoded largely by RNA viruses and retroid viruses, but a reverse transcriptase had been recruited by eukaryotes as the catalytic subunit of the telomerase [98,99]. There seems to be a distinct possibility that both classes of RDRPs are vestiges of the ancient RNA world that have been largely displaced from cellular biochemical machinery at an early stage of evolution, because of the deleterious effects of RNA replication in DNA-based organisms, while surviving in selfish genetic elements.

Material and Methods

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda) was searched using the BLASTP program [100]. Position-specific scoring matrix (PSSM) searches were conducted using the PSI-BLAST program, typically with a PSSM inclusion expectation (E) value threshold of 0.01, and were iterated until convergence [100,101]. Prior to PSI-BLAST searches, query sequences were evaluated for compositional bias using the SEG program [102]. If no such bias was detected, searches were run with the composition-based statistics turned off, in order to maximize sensitivity [103]. Multiple alignments of protein sequence were constructed using the T_Coffee program [104], followed by manual correction based on the PSI-BLAST results. Identification and statistical evaluation of conserved motifs in multiple protein sequences were performed using the Gibbs sampling method as implemented in the MACAW program [105,106]. Pattern searches were conducted using the GREF program of the SEALS package [107], and pattern-initiated BLAST searches were carried out using the PHI-BLAST program [60]?. Protein structure databases were searched for similar structures using the DALI program [108]. Protein secondary structure was predicted using the PHD program implemented on the PredictProtein server with a multiple alignment submitted as the [109,110]. Protein structures were visualized and manipulated using the Swiss-PDB viewer program [111] and the ribbon diagrams were made using MOLSCRIPT [112].

Authors' contributions

All three authors (LMI, EVK, LA) contributed to making the discoveries reported here. Author 1 and 3 (LMI, LA) prepared all the figures and the manuscript. All authors read and approved the final manuscript.

References

1. Anantharaman V, Koonin EV and Aravind L **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acids Res* 2002, **30:**1427-1464
2. Nelson DL and Cox MM **InLehninger Principles of Biochemistry.** *Worth Publishers Inc* 2000,
3. Goodman MF and Tippin B **The expanding polymerase universe.** *Nat Rev Mol Cell Biol* 2000, **1:**101-109
4. Burgers PM, Koonin EV, Bruford E, Blanco L, Burtis KC, Christman MF, Copeland WC, Friedberg EC, Hanaoka F and Hinkle DC **Eukaryotic DNA polymerases: proposal for a revised nomenclature.** *J Biol Chem* 2001, **276:**43487-43490

5. Cheetham GM and Steitz TA **Insights into transcription: structure and function of single-subunit DNA-dependent RNA polymerases.** *Curr Opin Struct Biol* 2000, **10:**117-123
6. Steitz TA **A mechanism for all polymerases.** *Nature* 1998, **391:**231-232
7. Cramer P **Multisubunit RNA polymerases.** *Curr Opin Struct Biol* 2002, **12:**89-97
8. Kamer G and Argos P **Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses.** *Nucleic Acids Res* 1984, **12:**7269-7282
9. Delarue M, Poch O, Tordo N, Moras D and Argos P **An attempt to unify the structure of polymerases.** *Protein Eng* 1990, **3:**461-467
10. Poch O, Sauvaget I, Delarue M and Tordo N **Identification of four conserved motifs among the RNA-dependent polymerase encoding elements.** *Embo J* 1989, **8:**3867-3874
11. Aravind L, Mazumder R, Vasudevan S and Koonin EV **Trends in protein evolution inferred from sequence and structure analysis.** *Curr Opin Struct Biol* 2002, **12:**392-399
12. Pei J and Grishin NV **GGDEF domain is homologous to adenylyl cyclase.** *Proteins* 2001, **42:**210-216
13. Makarova KS, Aravind L, Grishin NV, Rogozin IB and Koonin EV **A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30:**482-496
14. Koonin EV, Wolf YI, Kondrashov AS and Aravind L **Bacterial homologs of the small subunit of eukaryotic DNA primase.** *J Mol Microbiol Biotechnol* 2000, **2:**509-512
15. Artymiuk PJ, Poirrette AR, Rice DW and Willett P **A polymerase I palm in adenylyl cyclase?** *Nature* 1997, **388:**33-34
16. Murzin AG **How far divergent evolution goes in proteins.** *Curr Opin Struct Biol* 1998, **8:**380-387
17. Aravind L, Leipe DD and Koonin EV **Toprim – a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins.** *Nucleic Acids Res* 1998, **26:**4205-4213
18. Keck JL, Roche DD, Lynch AS and Berger JM **Structure of the RNA polymerase domain of E. coli primase.** *Science* 2000, **287:**2482-2486
19. Podobnik M, McInerney P, O'Donnell M and Kuriyan J **A TOPRIM domain in the crystal structure of the catalytic core of Escherichia coli primase confirms a structural link to DNA topoisomerases.** *J Mol Biol* 2000, **300:**353-362
20. Bushnell DA, Cramer P and Kornberg RD **Structural basis of transcription: alpha-amanitin-RNA polymerase II cocrystal at 2.8 A resolution.** *Proc Natl Acad Sci U S A* 2002, **99:**1218-1222
21. Gnatt AL, Cramer P, Fu J, Bushnell DA and Kornberg RD **Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution.** *Science* 2001, **292:**1876-1882
22. Cramer P, Bushnell DA and Kornberg RD **Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution.** *Science* 2001, **292:**1863-1876
23. Vassylyev DG, Sekine S, Laptenko O, Lee J, Vassylyeva MN, Borukhov S and Yokoyama S **Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution.** *Nature* 2002, **417:**712-719
24. Zhang G, Campbell EA, Minakhin L, Richter C, Severinov K and Darst SA **Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution.** *Cell* 1999, **98:**811-824
25. Mourrain P, Beclin C, Elmayan T, Feuerbach F, Godon C, Morel JB, Jouette D, Lacombe AM, Nikic S and Picault N **Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance.** *Cell* 2000, **101:**533-542
26. Schiebel W, Pelissier T, Riedel L, Thalmeir S, Schiebel R, Kempe D, Lottspeich F, Sanger HL and Wassenegger M **Isolation of an RNA-directed RNA polymerase-specific cDNA clone from tomato.** *Plant Cell* 1998, **10:**2087-2101
27. Vaistij FE, Jones L and Baulcombe DC **Spreading of RNA targeting and DNA methylation in RNA silencing requires transcription of the target gene and a putative RNA-dependent RNA polymerase.** *Plant Cell* 2002, **14:**857-867
28. Dracheva S, Koonin EV and Crute JJ **Identification of the primase active site of the herpes simplex virus type 1 helicase-primase.** *J Biol Chem* 1995, **270:**14148-14153
29. Holm L and Sander C **DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily.** *Trends Biochem Sci* 1995, **20:**345-347
30. Aravind L and Koonin EV **DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history.** *Nucleic Acids Res* 1999, **27:**1609-1618
31. Sweetser D, Nonet M and Young RA **Prokaryotic and eukaryotic RNA polymerases have homologous core subunits.** *Proc Natl Acad Sci U S A* 1987, **84:**1192-1196
32. Ebright RH **RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II.** *J Mol Biol* 2000, **304:**687-698
33. Passarelli AL, Todd JW and Miller LK **A baculovirus gene involved in late gene expression predicts a large polypeptide with a conserved motif of RNA polymerases.** *J Virol* 1994, **68:**4673-4678
34. Passarelli AL and Miller LK **Identification of genes encoding late expression factors located between 56.0 and 65.4 map units of the Autographa californica nuclear polyhedrosis virus genome.** *Virology* 1993, **197:**704-714
35. Iyer LM, Aravind L and Koonin EV **Common origin of four diverse families of large eukaryotic DNA viruses.** *J Virol* 2001, **75:**11720-11734
36. Dieci G, Hermann-Le Denmat S, Lukhtanov E, Thuriaux P, Werner M and Sentenac A **A universally conserved region of the largest subunit participates in the active site of RNA polymerase III.** *Embo J* 1995, **14:**3766-3776
37. Iyer LM, Kumpatla SP, Chandrasekharan MB and Hall TC **Transgene silencing in monocots.** *Plant Mol Biol* 2000, **43:**323-346
38. Cogoni C and Macino G **Post-transcriptional gene silencing across kingdoms.** *Curr Opin Genet Dev* 2000, **10:**638-643
39. Grishok A and Mello CC **RNAi (Nematodes: Caenorhabditis elegans).** *Adv Genet* 2002, **46:**339-360
40. Fire A **RNA-triggered gene silencing.** *Trends Genet* 1999, **15:**358-363
41. Montgomery MK and Fire A **Double-stranded RNA as a mediator in sequence-specific genetic silencing and co-suppression.** *Trends Genet* 1998, **14:**255-258
42. Parrish S, Fleenor J, Xu S, Mello C and Fire A **Functional anatomy of a dsRNA trigger: differential requirement for the two trigger strands in RNA interference.** *Mol Cell* 2000, **6:**1077-1087
43. Zamore PD, Tuschl T, Sharp PA and Bartel DP **RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals.** *Cell* 2000, **101:**25-33
44. Hamilton AJ and Baulcombe DC **A species of small antisense RNA in posttranscriptional gene silencing in plants.** *Science* 1999, **286:**950-952
45. Hammond SM, Bernstein E, Beach D and Hannon GJ **An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells.** *Nature* 2000, **404:**293-296
46. Hutvagner G and Zamore PD **A microRNA in a multiple-turnover RNAi enzyme complex.** *Science* 2002, **297:**2056-2060
47. Astier-Manifacier S and Cornuet P **[RNA-dependent RNA-polymerase in Brassica chinensis L].** *C R Acad Sci Hebd Seances Acad Sci D* 1970, **270:**2587-2590
48. Astier-Manifacier S and Cornuet P **RNA-dependent RNA polymerase in Chinese cabbage.** *Biochim Biophys Acta* 1971, **232:**484-493
49. Ikegami M and Fraenkel-Conrat H **Characterization of the RNA-dependent RNA polymerase of tobacco leaves.** *J Biol Chem* 1979, **254:**149-154
50. Schiebel W, Haas B, Marinkovic S, Klanner A and Sanger HL **RNA-directed RNA polymerase from tomato leaves. I. Purification and physical properties.** *J Biol Chem* 1993, **268:**11851-11857
51. Cogoni C and Macino G **Gene silencing in Neurospora crassa requires a protein homologous to RNA-dependent RNA polymerase.** *Nature* 1999, **399:**166-169
52. Pickford AS, Catalanotto C, Cogoni C and Macino G **Quelling in Neurospora crassa.** *Adv Genet* 2002, **46:**277-303
53. Litiere K, van Eldik GJ, Jacobs JJ, Van Montagu M and Cornelissen M **Posttranscriptional gene silencing of gn1 in tobacco triggers accumulation of truncated gn1-derived RNA species.** *Rna* 1999, **5:**1364-1373
54. Martens H, Novotny J, Oberstrass J, Steck TL, Postlethwait P and Nellen W **RNAi in Dictyostelium: the role of RNA-directed RNA polymerases and double-stranded RNase.** *Mol Biol Cell* 2002, **13:**445-453

55. Sijen T, Fleenor J, Simmer F, Thijssen KL, Parrish S, Timmons L, Plasterk RH and Fire A **On the role of RNA amplification in dsRNA-triggered gene silencing.** *Cell* 2001, **107:**465-476
56. Meins F Jr **RNA degradation and models for post-transcriptional gene-silencing.** *Plant Mol Biol* 2000, **43:**261-273
57. Nishikura K **A short primer on RNAi: RNA-directed RNA polymerase acts as a key catalyst.** *Cell* 2001, **107:**415-418
58. Ahlquist P **RNA-dependent RNA polymerases, viruses, and RNA silencing.** *Science* 2002, **296:**1270-1273
59. Argos P **A sequence motif in many polymerases.** *Nucleic Acids Res* 1988, **16:**9909-9916
60. Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV and Altschul SF **Protein sequence similarity searches using patterns as seeds.** *Nucleic Acids Res* 1998, **26:**3986-3990
61. Holm L and Sander C **Touring protein fold space with Dali/FSSP.** *Nucleic Acids Res* 1998, **26:**316-319
62. Castillo RM, Mizuguchi K, Dhanaraj V, Albert A, Blundell TL and Murzin AG **A six-stranded double-psi beta barrel is shared by several protein superfamilies.** *Structure Fold Des* 1999, **7:**227-236
63. Mizuguchi K, Dhanaraj V, Blundell TL and Murzin AG **N-ethylmaleimide-sensitive fusion protein (NSF) and CDC48 confirmed as members of the double-psi beta-barrel aspartate decarboxylase/formate dehydrogenase family.** *Structure Fold Des* 1999, **7:**R215-216
64. Coles M, Diercks T, Liermann J, Groger A, Rockel B, Baumeister W, Koretke KK, Lupas A, Peters J and Kessler H **The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple betaalphabetabeta element.** *Curr Biol* 1999, **9:**1158-1168
65. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C and Murzin AG **SCOP database in 2002: refinements accommodate structural genomics.** *Nucleic Acids Res* 2002, **30:**264-267
66. Leipe DD, Wolf YI, Koonin EV and Aravind L **Classification and evolution of P-loop GTPases and related ATPases.** *J Mol Biol* 2002, **317:**41-72
67. Wang B, Jones DN, Kaine BP and Weiss MA **High-resolution structure of an archaeal zinc ribbon defines a general architectural motif in eukaryotic RNA polymerases.** *Structure* 1998, **6:**555-569
68. Aravind L and Koonin EV **DNA-binding proteins and evolution of transcription regulation in the archaea.** *Nucleic Acids Res* 1999, **27:**4658-4670
69. Hard T, Rak A, Allard P, Kloo L and Garber M **The solution structure of ribosomal protein L36 from Thermus thermophilus reveals a zinc-ribbon-like fold.** *J Mol Biol* 2000, **296:**169-180
70. Chen HT, Legault P, Glushka J, Omichinski JG and Scott RA **Structure of a (Cys3His) zinc ribbon, a ubiquitous motif in archaeal and eucaryal transcription.** *Protein Sci* 2000, **9:**1743-1752
71. Ferri ML, Peyroche G, Siaut M, Lefebvre O, Carles C, Conesa C and Sentenac A **A novel subunit of yeast RNA polymerase III interacts with the TFIIB-related domain of TFIIIB70.** *Mol Cell Biol* 2000, **20:**488-495
72. Bell SD and Jackson SP **The role of transcription factor B in transcription initiation and promoter clearance in the archaeon Sulfolobus acidocaldarius.** *J Biol Chem* 2000, **275:**12934-12940
73. Hahn S and Roberts S **The zinc ribbon domains of the general transcription factors TFIIB and Brf: conserved functional surfaces but different roles in transcription initiation.** *Genes Dev* 2000, **14:**719-730
74. Aravind L and Landsman D **AT-hook motifs identified in a wide variety of DNA-binding proteins.** *Nucleic Acids Res* 1998, **26:**4413-4421
75. Neuwald AF, Liu JS, Lipman DJ and Lawrence CE **Extracting protein alignment models from the sequence database.** *Nucleic Acids Res* 1997, **25:**1665-1677
76. Anantharaman V, Koonin EV and Aravind L **Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains.** *J Mol Biol* 2001, **307:**1271-1292
77. Van Duyne GD, Ghosh G, Maas WK and Sigler PB **Structure of the oligomerization and L-arginine binding domain of the arginine repressor of Escherichia coli.** *J Mol Biol* 1996, **256:**377-391
78. Aravind L and Koonin EV **Novel predicted RNA-binding domains associated with the translation machinery.** *J Mol Evol* 1999, **48:**291-302
79. Fletcher CM, Pestova TV, Hellen CU and Wagner G **Structure and interactions of the translation initiation factor eIF1.** *Embo J* 1999, **18:**2631-2637
80. Kraulis PJ **Similarity of protein G and ubiquitin.** *Science* 1991, **254:**581-582
81. Murzin AG **Familiar strangers.** *Nature* 1992, **360:**635
82. Baumann P, Qureshi SA and Jackson SP **Transcription: new insights from studies on Archaea.** *Trends Genet* 1995, **11:**279-283
83. Bell SD and Jackson SP **Mechanism and regulation of transcription in archaea.** *Curr Opin Microbiol* 2001, **4:**208-213
84. Lagos-Quintana M, Rauhut R, Lendeckel W and Tuschl T **Identification of novel genes coding for small expressed RNAs.** *Science* 2001, **294:**853-858
85. Dennis C **The brave new world of RNA.** *Nature* 2002, **418:**122-124
86. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B and Bartel DP **MicroRNAs in plants.** *Genes Dev* 2002, **16:**1616-1626
87. Pasquinelli AE **MicroRNAs: deviants no longer.** *Trends Genet* 2002, **18:**171-173
88. Grosshans H and Slack FJ **Micro-RNAs: small is plentiful.** *J Cell Biol* 2002, **156:**17-21
89. Llave C, Kasschau KD, Rector MA and Carrington JC **Endogenous and silencing-associated small RNAs in plants.** *Plant Cell* 2002, **14:**1605-1619
90. Llave C, Xie Z, Kasschau KD and Carrington JC **Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA.** *Science* 2002, **297:**2053-2056
91. Wassarman KM, Repoila F, Rosenow C, Storz G and Gottesman S **Identification of novel small RNAs using comparative genomics and microarrays.** *Genes Dev* 2001, **15:**1637-1651
92. Davids W, Amiri H and Andersson SG **Small RNAs in Rickettsia: are they functional?** *Trends Genet* 2002, **18:**331-334
93. Eddy SR **Non-coding RNA genes and the modern RNA world.** *Nat Rev Genet* 2001, **2:**919-929
94. Rivas E, Klein RJ, Jones TA and Eddy SR **Computational identification of noncoding RNAs in E. coli by comparative genomics.** *Curr Biol* 2001, **11:**1369-1373
95. Woese C **The universal ancestor.** *Proc Natl Acad Sci U S A* 1998, **95:**6854-6859
96. Leipe DD, Aravind L and Koonin EV **Did DNA replication evolve twice independently?** *Nucleic Acids Res* 1999, **27:**3389-3401
97. Aravind L, Anantharaman V and Koonin EV **Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA.** *Proteins* 2002, **48:**1-14
98. Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V and Cech TR **Reverse transcriptase motifs in the catalytic subunit of telomerase.** *Science* 1997, **276:**561-567
99. Nakamura TM, Morin GB, Chapman KB, Weinrich SL, Andrews WH, Lingner J, Harley CB and Cech TR **Telomerase catalytic subunit homologs from fission yeast and human.** *Science* 1997, **277:**955-959
100. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402
101. Aravind L and Koonin EV **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.** *J Mol Biol* 1999, **287:**1023-1040
102. Wootton JC **Non-globular domains in protein sequences: automated segmentation using complexity measures.** *Comput Chem* 1994, **18:**269-285
103. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV and Altschul SF **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.** *Nucleic Acids Res* 2001, **29:**2994-3005
104. Notredame C, Higgins DG and Heringa J **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-217
105. Neuwald AF, Liu JS and Lawrence CE **Gibbs motif sampling: detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4:**1618-1632
106. Schuler GD, Altschul SF and Lipman DJ **A workbench for multiple alignment construction and analysis.** *Proteins* 1991, **9:**180-190

107. Walker DR and Koonin EV **SEALS: a system for easy analysis of lots of sequences.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**333-339
108. Holm L and Sander C **Dali/FSSP classification of three-dimensional protein folds.** *Nucleic Acids Res* 1997, **25:**231-234
109. Rost B and Sander C **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232:**584-599
110. Rost B, Sander C and Schneider R **PHD – an automatic mail server for protein secondary structure prediction.** *Comput Appl Biosci* 1994, **10:**53-60
111. Guex N and Peitsch MC **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723
112. Kraulis PJ **MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures.** *Journal of App Crystallography* 1991, **24:**946-950