

Bacterial Homologs of the Small Subunit of Eukaryotic DNA Primase

Eugene V. Koonin*, Yuri I. Wolf, Alexy S. Kondrashov and L. Aravind

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda MD 20894, USA

Primases are RNA polymerases that synthesize the primer RNA that provides the 3' OH for strand elongation by DNA polymerases (Kornberg and Baker, 1992). Currently, three independent classes of primases are recognized. The best characterized of these are the DnaG-like proteins, which function as replicative primases in bacteria and some bacteriophages, and also have highly conserved homologs, whose function remains unknown, in all archaeal genomes sequenced to date (Aravind *et al.*, 1998). In bacteriophages, the DnaG-type primase domain is fused to the DnaB-like helicases, and this helicase-primase apparently has been acquired by eukaryotes via horizontal gene transfer (Leipe *et al.*, 2000). The catalytic domain of the DnaG-family proteins, the Toprim domain, is shared with topoisomerases (excluding topo IB from eukaryotes), OLD family nucleases and recR/M proteins (Aravind *et al.*, 1998; Keck *et al.*, 2000).

Almost complementary to the DnaG-like primases, in terms of phyletic distribution, are the eukaryote-type primases (EPs) that are comprised of two subunits (Santocanale *et al.*, 1993; Schneider *et al.*, 1998). The small subunit is a divalent cation-dependent enzyme that shows no detectable relationship with the Toprim domain. Nevertheless, it contains a highly conserved DxD dyad that resembles the equivalent dyad of the Toprim domain. The large, non-catalytic subunit of these primases is poorly conserved and appears to be required for DNA-binding and association of the newly synthesized primer with DNA polymerase α (Arezi *et al.*, 1999). Both EP subunits are conserved in all archaea and in baculoviruses (Leipe *et al.*, 1999). A distinct third type of primase is encoded by herpesviruses; these appear to be unrelated to the DnaG and EP families, but also possess a conserved DXD dyad associated with the active site (Dracheva *et al.*, 1995).

Thus bacteria and eukaryotes use unrelated primases for genome replication, whereas archaea are peculiar in possessing both bacterial and eukaryote-type primases in an otherwise typically "eukaryotic" replication system (Leipe *et al.*, 1999). Here we report the detection of previously unrecognized homologs of the catalytic subunit of EP in several groups of bacteria.

Sequence analysis of all hitherto recognized members of the EP family revealed the presence of three conserved motifs that are likely to participate in catalysis (Figure 1). Motif I contains the DxD dyad that is found in several unrelated classes of nucleotidyl transferases, including

other primases, and probably is the divalent cation-binding site of these proteins. Motif II contains the distinctive +GhH signature where '+' is a positively charged residue and 'h' is a hydrophobic residue. This motif is probably located in a loop and could function as part of the nucleotide-binding site with positively-charged residues interacting with the phosphoester bond under attack. Motif III contains a highly conserved aspartate that could form the third cation-coordinating ligand and contribute to the nucleophilic attack by the 3' OH.

An iterative PSI-BLAST search (Altschul *et al.*, 1997) of the non-redundant protein sequence database (National Center for Biotechnology Information, NIH, Bethesda), started with the *Aeropyrum pernix* homolog of the small EP subunit and using expectation (E) value of 0.1 as the cut-off for including sequences in the position-specific scoring matrix, resulted in the retrieval of all known archaeal, eukaryotic and viral members of the EP family, and in addition, a family of bacterial proteins from actinomycetes and *Bacillus subtilis*. Searches of the incompletely sequenced microbial genomes resulted in the detection of additional members of this bacterial family in *Pseudomonas aeruginosa*, *P. putida* and *Sinorhizobium meliloti*. A reverse search with the conserved central portion of the *Mycobacterium tuberculosis* protein Rv0269c used as the query similarly resulted in the specific retrieval of the entire EP family. Examination of the database search outputs showed that these bacterial proteins contained counterparts to all the three motifs that are typical of small subunits of EPs (Figure 1). Multiple alignment analysis using the MACAW program with the Gibbs sampling method for block search (Schuler *et al.*, 1991; Neuwald *et al.*, 1995) showed that motifs I, II and III were the most conserved blocks in EPs and their potential bacterial counterparts. For each of these motifs the probability of occurring by chance in the given set of proteins was $<10^{-18}$. Furthermore, a HMMER2 (Eddy, 1998) search of the individual microbial genomes with a hidden Markov model generated from the multiple alignment of the typical EPs detects the mycobacterial proteins of the above mentioned family as the best hits without generating any higher scoring alignments with proteins from other bacterial proteomes. Although the similarity between these bacterial proteins and the EPs was subtle and not highly statistically significant in the context of entire NR database, the complete bidirectional specificity of the database searches, and even more importantly, the conservation of the three signature motifs lead us to conclude that these are indeed homologs of the EPs. Furthermore, the conservation of each of the aforementioned residues implicated in catalysis, with a single Q for H substitution in motif II of the *B. subtilis* YkoU protein (Figure 1), strongly suggests that these bacterial proteins possess primase activity.

The newly detected bacterial members of the EP superfamily notably differ from the archaeal-eukaryotic primases in the regions immediately C-terminal to Motif I and in the region between motif II and III (Figures 1 and 2).

Received June 12, 2000; accepted June 12, 2000. *For correspondence. Email koonin@ncbi.nlm.nih.gov; Tel. 301-435-5913; Fax. 301-480-9241.

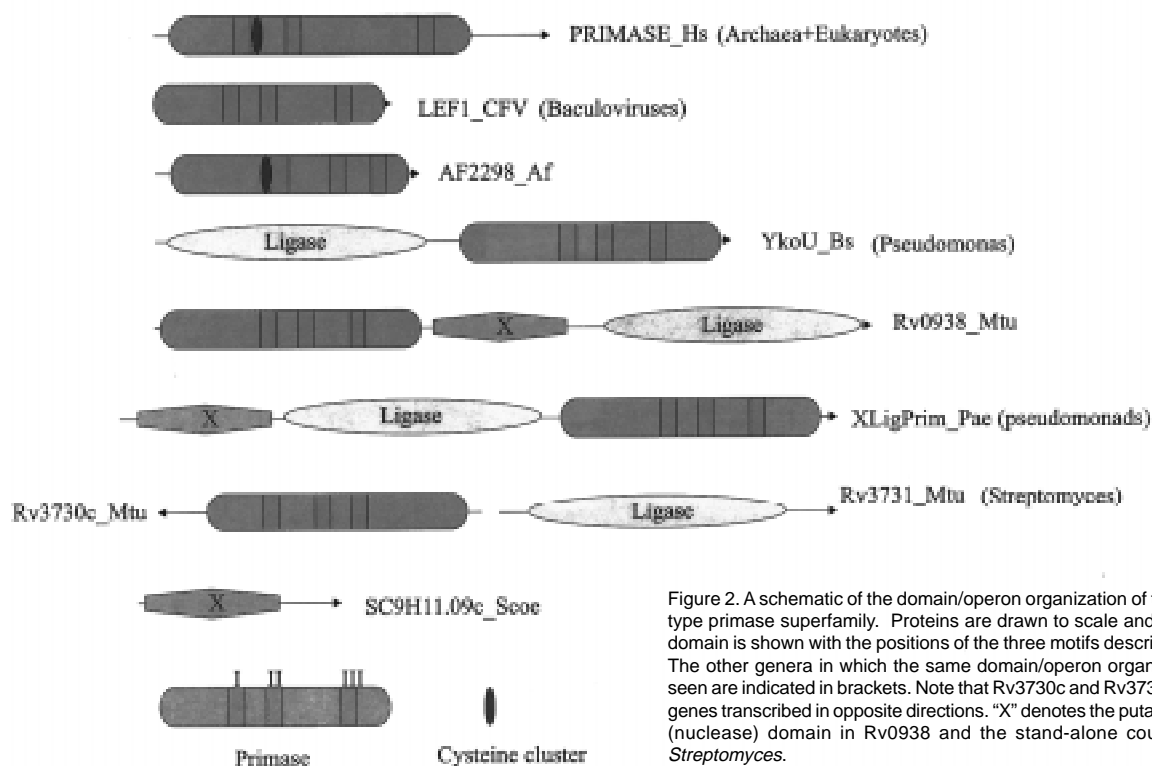


Figure 2. A schematic of the domain/operon organization of the eukaryotic-type primase superfamily. Proteins are drawn to scale and each primase domain is shown with the positions of the three motifs described in the text. The other genera in which the same domain/operon organization can be seen are indicated in brackets. Note that Rv3730c and Rv3731 are separate genes transcribed in opposite directions. "X" denotes the putative enzymatic (nuclease) domain in Rv0938 and the stand-alone counterpart from *Streptomyces*.

In the former region, all the classic EPs, with the exception of those from baculoviruses and two of the archaeal ones, contain an inserted cysteine cluster that is likely to coordinate a divalent cation (Figure 1). The second region of difference shows poor conservation even among the typical EPs and is likely to form a variable α -helical insert. These difference may, in part, account for the relatively low statistical significance of the alignments between the archaeal-eukaryotic and bacterial members of the EP superfamily (see above) despite the conservation of the predicted catalytic motifs.

The domain architectures and operon organization of the bacterial EP homologs are suggestive of a distinct function in DNA repair (Figure 2). Rv0938 from *M. tuberculosis*, YkoU from *B. subtilis* and their homologs from two pseudomonads are fused to archaeal-eukaryotic-type ATP-dependent DNA ligases. SC4C6.19 from *Streptomyces coelicolor* and Rv3730c from *M. tuberculosis* are encoded by genes that are adjacent to genes for ATP-dependent DNA ligases (Figure 2). This strongly supports functional coupling between the EP homologs and ATP-dependent DNA ligases of these bacteria. ATP-dependent ligases are indispensable components of the archaeal-eukaryotic DNA replication machinery, but in bacteria, they are present only sporadically in several lineages (Leipe *et al.*, 1999). In a phylogenetic analysis, bacterial ATP-dependent ligases formed two distantly related, well-supported clusters. One cluster includes proteins from *Haemophilus*, *Neisseria* and *Campylobacter* that group with the ligases of large eukaryotic DNA viruses, whereas the second cluster, which shows a clear affinity with archaeal ligases, includes ligases from those bacterial genomes that also encode EP homologs (LA, unpublished). This tree topology and the association between the EP homologs

and ligases suggest that the entire operon including genes for a DNA ligase and an EP homolog has been acquired by bacteria from an archaeal source, which was followed by sporadic lateral dissemination in diverse bacterial lineages (Figure 2). In the currently available archaeal genomes, the genes for the small EP subunit and the ligase are not adjacent, but this hypothesis suggests that further sampling of archaeal genomes is likely to reveal such an operon.

The poorly conserved large subunit of the EPs could not be detected in bacteria that encoded the catalytic subunit homologs (or any other bacterial genomes). While these could have diverged beyond recognition, the spatial and probable functional association with the ligases seems to suggest that bacterial homologs of the small EP subunit function independently of the large subunit. All bacteria encode a NAD-dependent ligase, which is only very distantly related to the ATP-dependent ligases (Aravind and Koonin, 1999; Singleton *et al.*, 1999) and functions in bacterial DNA replication (Kornberg and Baker, 1992). Hence, the bacterial ATP-dependent ligase-EP homolog combination is unlikely to contribute to replication and instead may be involved in long-patch DNA repair where it could couple the ligation of Okazaki fragments with priming of new fragments. However, the presence of nucleases and topoisomerases in the Toprim superfamily of primases suggests a degree of caution with respect to the prediction of the biochemical function of the bacterial EP homologs. It is, in principle, possible that their primary activity is that of a nuclease.

The mycobacterial protein Rv0938 contains an additional conserved globular domain between the EP and ligase domains (Figure 2). In *Pseudomonas*, this domain occurs in the same polypeptide, but N-terminally of the

ligase-primase fusion, and in *S. coelicolor* it is a stand-alone protein (Figure 2). This domain is additionally encoded in the genomes of *Sinorhizobium* and *Dehalococcoides* where the domain architecture is hard to determine due to incomplete sequencing. This domain contains three characteristic motifs with conserved aspartates and histidines, which is compatible with an enzymatic function (data not shown). The fusion in *M. tuberculosis* and *Pseudomonas* along with the phyletic co-occurrence of this domain (protein) with EP homologs and archaeal-type ligases suggests that it could be a nuclease that functions in the predicted DNA repair pathway that also involves the EP-Ligase combination.

Identification of EP homologs in bacteria provides an interesting new direction for experimental investigation of DNA repair system and once again emphasizes the probable major role of horizontal gene transfer in the evolution of prokaryotes. With these findings, the phylogenetic and functional complementarity between bacterial-type and eukaryotic-type primases is becoming even more striking, with DnaG-type primases probably recruited for repair in archaea and eukaryotes, and EP-type primases apparently assuming a similar role in some of the bacteria. Additionally, this analysis may clarify certain functional features of the eukaryotic and archaeal primases themselves. While the importance of the DXD motif of the EPs has been previously recognized (Barrett *et al.*, 1996; Evans *et al.*, 1997), the remaining details of the catalytic domain remain unknown. From the conservation pattern and predicted structural elements, it is clear that the catalytic domain of these primases is unrelated to the Toprim domain. Furthermore, beyond cation-binding mediated by the DxD motif, there are likely to be mechanistic differences between these two types of primases, as suggested by the presence of the unique motifs II and III, for which there are no counterparts in the Toprim domain. The identification of these motifs could help in clarifying the two different chemistries that result in a similar polymerization reaction in the different branches of life.

Acknowledgements

Preliminary sequence data for unfinished genomes was obtained from The Institute for Genomic Research website at <http://www.tigr.org>.

References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25: 3389-3402.
- Aravind, L., and Koonin, E.V. 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* 287: 1023-1040.
- Aravind, L., Leipe, D.D., and Koonin, E.V. 1998. Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucl. Acids Res.* 26: 4205-4213.
- Arezi, B., Kirk, B.W., Copeland, W.C., and Kuchta, R.D. 1999. Interactions of DNA with human DNA primase monitored with photoactivatable cross-linking agents: implications for the role of the p58 subunit. *Biochem.* 38: 12899-12907.
- Barrett, J.W., Lauzon, H.A., Mercuri, P.S., Krell, P.J., Sohi, S.S., and Arif, B.M. 1996. The putative LEF-1 proteins from two distinct Choristoneura fumiferana multiple nucleopolyhedroviruses share domain homology to eukaryotic primases. *Virus Genes.* 13: 229-237.
- Dracheva, S., Koonin, E.V., and Crute, J.J. 1995. Identification of the primase active site of the herpes simplex virus type 1 helicase-primase. *J. Biol. Chem.* 270: 14148-14153.

- Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics.* 14: 755-763.
- Evans, J.T., Leisy, D.J., and Rohrmann, G.F. 1997. Characterization of the interaction between the baculovirus replication factors LEF-1 and LEF-2. *J. Virol.* 71: 3114-3119.
- Keck, J.L., Roche, D.D., Lynch, A.S., and Berger, J.M. 2000. Structure of the RNA polymerase domain of *E. coli* primase. *Science.* 287: 2482-2486.
- Kornberg, A. and Baker, T. 1992. *DNA Replication*, 2nd Edn. New York, NY: W. H. Freeman and Co.
- Leipe, D.D., Aravind, L., Grishin, N.V., and Koonin, E.V. 2000. The bacterial replicative helicase DnaB evolved from a RecA duplication. *Genome Res.* 10: 5-16.
- Leipe, D.D., Aravind, L., and Koonin, E.V. 1999. Did DNA replication evolve twice independently? *Nucl. Acids Res.* 27: 3389-3401.
- Neuwald, A.F., Liu, J.S., and Lawrence, C.E. 1995. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* 4: 1618-1632.
- Santocanale, C., Foiani, M., Lucchini, G., and Plevani, P. 1993. The isolated 48,000-dalton subunit of yeast DNA primase is sufficient for RNA primer synthesis. *J. Biol. Chem.* 268: 1343-1348.
- Schneider, A., Smith, R.W., Kautz, A.R., Weisshart, K., Grosse, F., and Nasheuer, H.P. 1998. Primase activity of human DNA polymerase alpha-primase. Divalent cations stabilize the enzyme activity of the p48 subunit. *J. Biol. Chem.* 273: 21608-21615.
- Schuler, G.D., Altschul, S.F., and Lipman, D.J. 1991. A workbench for multiple alignment construction and analysis. *Proteins.* 9: 180-190.
- Singleton, M.R., Hakansson, K., Timson, D.J., and Wigley, D.B. 1999. Structure of the adenylation domain of an NAD⁺-dependent DNA ligase. *Structure Fold. Des.* 7: 35-42.