Nucleic Acids Research, 1999, Vol. 27, No. 7 **1609–1618**

DNA polymerase β-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history

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

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [2]Department of Biology, Texas A&M University, College Station, TX 77843, USA

Received January 11, 1999; Accepted January 30, 1999

ABSTRACT

A detailed analysis of the polβ superfamily of nucleotidyltransferases was performed using computer methods for iterative database search, multiple alignment, motif analysis and structural modeling. Three previously uncharacterized families of predicted nucleotidyltransferases are described. One of these new families includes small proteins found in all archaea and some bacteria that appear to consist of the minimal nucleotidyltransferase domain and may resemble the ancestral state of this superfamily. Another new family that is specifically related to eukaryotic polyA polymerases is typified by yeast Trf4p and Trf5p proteins that are involved in chromatin remodeling. The TRF family is represented by multiple members in all eukaryotes and may be involved in yet unknown nucleotide polymerization reactions required for maintenance of chromatin structure. Another new family of bacterial and archaeal nucleotidyltransferases is predicted to function in signal transduction since, in addition to the nucleotidyltransferase domain, these proteins contain ligand-binding domains. It is further shown that the catalytic domain of γ proteobacterial adenylyl cyclases is homologous to the polβ superfamily nucleotidyltransferases which emphasizes the general trend for the origin of signal-transducing enzymes from those involved in replication, repair and RNA processing. Classification of the polβ superfamily into distinct families and examination of their phyletic distribution suggests that the evolution of this type of nucleotidyltransferases may have included bursts of rapid divergence linked to the emergence of new functions as well as a number of horizontal gene transfer events.

INTRODUCTION

The transfer of a nucleotide to an acceptor hydroxyl group is a central reaction in a variety of biological processes. This reaction is catalyzed by nucleotidyltransferases that belong to more than 10 distinct superfamilies. Within each of the superfamilies the proteins are conserved at the sequence level. By contrast, different superfamilies show little, if any, sequence similarity to each other and, in several cases, have been shown to possess different structural folds, though some general common features of their interaction with the nucleotide substrate have been proposed (1–4). The nucleotidyltransferases involved in basic biological processes include: (i) replication and repair: DNA polymerases of at least five distinct superfamilies (5), primases of at least three distinct families (3) and DNA ligases of two distinct, though distantly related families (2; L.Aravind and E.V.Koonin, unpublished observations); (ii) transcription: at least two families of DNA-dependent RNA polymerases (6); (iii) RNA processing: at least two families of polyA polymerases (7), mRNA capping enzymes (2) and at least two families of CCA-adding enzymes (8); and (iv) viral replication that, in addition to the DNA polymerases, may also involve the RNA-dependent RNA polymerase and reverse transcriptases (9,10). In addition to these fundamental processes, distinct nucleotidyltransferases are involved in more specialized pathways, such as telomere maintenance, translesion DNA synthesis during repair, immunoglobulin gene rearrangement and signal transduction, as in the case of the 2′–5′ oligoA synthetase (11).

One of the most widespread superfamilies of nucleotidyltransferases that are involved in the majority of the processes listed above, though do not perform the role of the principal replicative polymerase in any known system, is typified by the eukaryotic DNA polymerase β (hereinafter polβ superfamily). Structural comparisons indicated that polymerase β is related to kanamycin nucleotidyltransferase, and examination of the conserved residues has suggested a common active site and an evolutionary relationship between these nucleotidyltransferases (1). Further searches resulted in the unification of several additional nucleotidyltransferases involved in diverse processes under this superfamily which is characterized by a distinct (although not unique) amino acid residue pattern, namely hG[GS]x(9,13)Dh[DE]h (x indicates any amino acid and h indicates a hydrophobic amino acid) (1). The following functional groups of nucleotidyltransferases have been included in the polβ superfamily: (i) polyA polymerases, (ii) protein nucleotidyltransferases, such as GlnD and GlnE, (iii) CCA-adding enzyme, (iv) interferon-induced 2′–5′ synthetase, (v) DNA polymerase β and terminal deoxynucleotidyltransferase, and (vi) antibiotic nucleotidyltransferases (1,12). The sequence similarity between some of these families is quite low. Since each of them includes enzymes with experimentally demonstrated nucleotidyltransferase activity, it appears that the sequence and

*To whom correspondence should be addressed at: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. Tel: +1 301 435 7797; Fax: +1 301 480 9241; Email: aravind@ncbi.nlm.nih.gov

structural features common to the entire superfamily should be reliable predictors of such an activity.

Experimental studies on different members of the polβ superfamily suggest that their mode of action is simpler, compared to the more processive, larger, typically multi-subunit nucleic acid polymerases. The polβ-like enzymes appear to undergo cycles of dissociation and re-association in the course of nucleotide addition, even in template-dependent polynucleotide synthesis, e.g. by DNA polymerase β (13,14), or catalyze nucleotide polymerization without enzyme translocation as in the case of the CCA-adding enzyme (15).

Given the wide range of functions in the polβ superfamily, which is paralleled by the diversity of their protein sequences, and the vastly increased amount of sequence information coming from completely sequenced genomes, we re-investigated this superfamily. We used recently developed, sensitive computer methods in order to identify potential new groups of nucleotidyltransferases and clarify their structural and evolutionary relationships. Here we describe what seems to represent the 'minimal domain' of the polβ superfamily and demonstrate independent existence of 'minimal' nucleotidyltransferases (MNTs) in a wide range of archaea and bacteria. We show that eukaryotic Trf proteins (typified by the yeast Trf4p and Trf5p) that appear to function in chromatin condensation in conjunction with topoisomerase I belong to a large family of eukaryote-specific nucleotidyltransferases within the polβ superfamily and are related to eukaryotic polyA polymerases. The identification of this novel class of eukaryotic nuclear nucleotidyltransferases suggests the possibility of hitherto unsuspected polymerase activities involved in chromosomal dynamics. We also recognize another new family of nucleotidyltransferases that are related to GlnD and GlnE and are likely to catalyze nucleotidylation of proteins in yet unidentified bacterial and archaeal signal transduction pathways. We further show that adenylyl cyclases from *Escherichia coli* and other γ proteobacteria contain a polβ-type nucleotidyltransferase domain, which provides insight into their catalytic mechanism and likely origin in evolution. Examination of the phyletic distribution of the polβ superfamily suggests that these nucleotidyltransferases probably played an important role in nucleotide utilization from very early in evolution and have been recruited to participate in different processes that involve nucleotide polymerization, on a number of independent occasions.

MATERIALS AND METHODS

Databases, sequence analysis and structural modeling

The databases used in this study were the Non-Redundant (NR) database at the NCBI, the protein sets encoded in all publicly available completely sequenced genomes and nucleotide sequences from publicly available incomplete genomes. The principal database search method used was PSI-BLAST (16) which generates a weighted profile from the sequences detected in the first pass of a gapped-BLAST search and iteratively searches the database using this profile as the query. Normally, the expectation value cut-off for inclusion of sequences into the profile at each iteration was set at 0.01. The program constructs a position-dependent weight matrix (profile) from multiple alignments generated from the BLAST hits above a certain expectation value (e-value) and carries out iterative database searches using the information derived from this profile (16). The program also allows generation of 'checkpoint' profiles with fixed e-value cut-offs and number of iterations that can be used in searches of new databases such as complete genomes or in subsequent searches with altered e-value cut-offs (17). The estimates of statistical significance of the PSI-BLAST results are based on the extreme value distribution statistics originally developed by Karlin and Altschul for local alignments without gaps (18) and subsequently shown to apply to gapped alignments as well (16,19). While there is no analytical proof of the applicability of the Karlin–Altschul statistics to searches that use profiles as queries, extensive computer simulations showed a nearly perfect fit of the score distribution obtained in such searches to the extreme value distribution (16). Therefore, e-values reported for each retrieved sequence at the point when its alignment with the query exceeds the cut-off for the first time appear to be reliable estimates of statistical significance. Once a sequence is included in the model, e-values reported for it (and its closely related homologs) at subsequent iterations become inflated and do not accurately represent the statistical significance (20). All e-values reported here are for the first appearance of the given sequences above the cut-off.

The main source of artifacts that may arise in database searches and are inevitably amplified in PSI-BLAST iterations are low complexity regions in protein sequences that typically correspond to non-globular domains (21). In order to avoid such artifacts, but also prevent the loss of any relevant information, all searches in this study were run directly and after masking the low complexity regions in the query sequences using the SEG program (22) and the COILS program (23) (window length 21) which identifies coiled coil regions (a special case of low complexity). The SEG program was applied with two sets of parameters, namely the standard ones used by default with the BLAST family programs [window length (W) 12, trigger complexity 2.2, extension complexity 2.5] and the parameters optimized for the detection of non-globular domains in proteins [W = 45, trigger complexity 3.4, extension complexity 3.75].

Additionally, the recently developed tool that combines local alignment searches with pattern searches, PHI-BLAST (24), was used to assess subtle relationships for proteins that do not have homologs with sufficiently diverged sequences and, therefore, failed to produce effective profiles in the PSI-BLAST analysis. Under PHI-BLAST, the statistical significance of database hits is estimated using the same Karlin–Altschul statistics as employed in PSI-BLAST but for a reduced search space defined by the sequences that contain a given pattern. Therefore, e-values reported by PHI-BLAST are not directly comparable to those from gapped BLAST (or PSI-BLAST). Nevertheless, these statistical estimates help assess the relevance of conserved patterns detected in sequences (24). Alternatively, single-motif blocks were used to generate a weighted matrix to search the database using the MoST program (25) with a cut-off of r = 0.005.

The likelihood of an alignment of two sequences being indicative of a structural similarity was determined using the ZEGA program (26). Briefly, the probability that a given (or greater) alignment score is observed between two protein sequences that actually correspond to different structures is calculated using an analytical function derived from the distribution of alignment scores for sequences of proteins with known three-dimensional structures that have the same fold and those with different folds. The alignments are constructed using a modification of the Needleman–Wunsch algorithm (27) with zero end gap penalties.

Multiple alignments were constructed by using the Gibbs sampling option of the MACAW program (28,29) to detect

conserved motif blocks in a set of protein sequences, followed by global alignment with the clustalX program (30). Both alignment procedures used the Blosum series of matrices. Similarity-based single linkage clustering was carried out using the GROUPER script of the SEALS package (31) with serial gapped-BLAST bit score cut-offs in the range of 40–70.

Protein secondary structure was predicted using the PHD program, with multiple alignments used as queries (32). Manipulations with protein three-dimensional structures were conducted using the SWISS-PDB viewer version 3 and homology modeling was carried out by generating a structural alignment in SWISS-PDB viewer and then submitting it for modeling by PROMODII (33) which uses the Gromos energy minimization script (34). Large-scale sequence analysis was handled using the SEALS program package (31).

RESULTS AND DISCUSSION

Delineation of the polβ-type nucleotidyltransferase superfamily using profile searches

Using a number of different starting points for iterative database search, we were able to transitively establish relationships between most members of the polβ nucleotidyltransferase superfamily at statistically significant levels (Table 1; Fig. 1). For example, a PSI-BLAST search initiated with the sequence of a newly predicted nucleotidyltransferase from *Schizosaccharomyces pombe* (gi 3426138) resulted in the recovery of several distinct types of nucleotidyltransferases including polyA polymerases, 2′–5′ A synthetase, archaeal CCA-adding enzymes, several small uncharacterized proteins from archaea and bacteria as well as some of the antibiotic nucleotidyltransferases at e-values below the 0.01 threshold. In addition, uridylyltransferases (GlnD), DNA polymerase β and bacterial CCA-adding enzymes were detected in this search with higher (less significant) e-values. Subsequent searches performed using these proteins as queries transitively connected the entire polβ superfamily at e-values <0.01 (Fig. 1). In addition to the earlier described members, we detected three new families (Table 1). The prediction of a nucleotidyltransferase activity for each of them was supported by statistically significant similarity to proteins that possess experimentally demonstrated nucleotidyltransferase activity (Fig. 1 and see below).

Inspection of sequences of the experimentally studied poxviral polyA polymerases (35) has shown the presence of the polβ-type nucleotidyltransferase signature motif (12) which suggested that these enzymes contain a similar domain; however, in none of our searches, did these sequences emerge with statistically significant e-values. In order to evaluate this relationship further, we carried out searches using the PHI-BLAST program, with the poxvirus sequences and the aforementioned pattern as queries. This analysis provided some additional support (e-value of 0.08 with the *Methanococcus jannaschii* CCA-adding enzyme) for the distant but evolutionarily and functionally relevant relationship between the poxvirus polyA polymerases and the polβ nucleotidyltransferase superfamily.

In the course of these searches, we also unexpectedly observed that γ proteobacterial adenylyl cyclases typified by *E.coli* CyaA are distantly but specifically related to the polβ-type nucleotidyltransferases. Below we describe the newly identified families of polβ-type nucleotidyltranferases as well as the evolutionary implications of the classification and phyletic distribution of this superfamily.

Table 1. Sequence-based classification and phyletic distribution of DNA polymerase β-type nucleotidyltransferases[a]

	Bacteria	Archaea	Eukaryotes
GROUP 1			
Family 1 DNA polymerase X	Deinococcus Thermus Aquifex Bacillus	Methanobacterium	Crithidia Saccharomyces Animals (vertebrates) African swine fever virus
Family 2 Terminal deoxynucleotidyl transferase (TdT)	-	-	Animals (vertebrates) Schizosaccharomyces
GROUP 2			
Family 1 CCA adding enzyme (Archaeal type)	-	All archaea	-
Family 2 PolyA polymerase (Eukaryotic type)	-	-	Schizosaccharomyces Saccharomyces Caenorhabditis (2) Vertebrates (>2)
Family 3 TRF4/5	-	-	Saccharomyces (2) Schizosaccharomyces (5) Caenorhabditis (12) Arabidopsis(4) Vertebrates Plasmodium
Family 4 2′-5′ A polymerase	-	-	Animals (vertebrates)
GROUP 3			
Family 1 PolyA Polymerase/ CCA adding enzyme (Bacterial type)	E coli(2) Bacillus(1) Aquifex(2) Synechocystis(2) Haemophilus (2) Mycobacterium(1) Chlamydia(2) Deinococcus(2) Helicobacter(1) Treponema(1) Borrelia(1)	-	Saccharomyces Animals Plants
GROUP 4			
Family 1 Uridylyl transferase (GlnD)	E coli Klebsiella Azotobacter Haemophilus Mycobacterium Pseudomonas	-	-
Family 2 Adenylyl transferase	Same as above + Aquifex	-	-
(GlnE)	which lacks GlnD		
Family 3 CBS and cNMP-binding domain-containing nucleotidyltransferases	Aquifex Rhodobacter Vibrio Pseudomonas Chlorobium	Archaeoglobus	-
GROUP 5			
Family 1 "Minimal" nucleotidyltransferases	Synechocystis(3) Aquifex Deinococcus(2?) Thermotoga Clostridium Enterococcus Staphylococcus	Methanococcus (10) Archaeoglobus (13) Methanobacterium (3) Pyrococcus (4) Sulfolobus	-
GROUP 6			
Family 1 Kanamycin nucleotidyltransferases	Staphylococcus	-	-
GROUP 7			
Family 1 Streptomycin Nucleotidyltransferases	Staphylococcus Pseudomonas E. coli Enterococcus	-	-
GROUP 8			
Family 1 Poxviral polyA polymerases	-	-	Poxviruses
GROUP 9			
Family 1 γ proteobacterial adenylyl cyclases	Escherichia Haemophilus Yersinia Aeromonas Pseudomonas	-	-

[a]In cases where the given organism encodes more than one representative of a family, the number is indicated in parentheses.

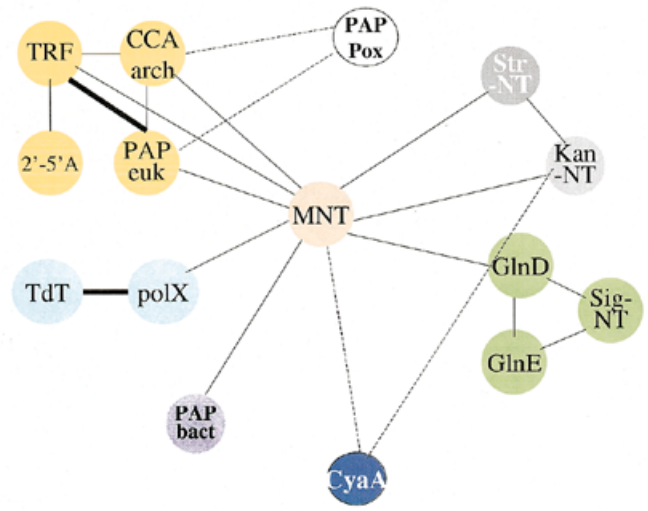


Figure 1. The polβ nucleotidyltransferase superfamily: transitive closure. Families are shown by circles, and those families that belong to the same group (Table 1) are indicated by identical color. Thick connecting lines indicate an e-value <0.01 in a single-pass BLAST for at least one pair of members of the given two families, and thin lines indicate an e-value <0.01 in at least one iterative PSI-BLAST search. Broken lines indicate that only limited, not statistically significant similarity was detectable in PSI-BLAST searches (see text). Abbreviations: CCA arch, archaeal CCA-adding enzymes; 2′–5′ A, 2′–5′ oligoA synthetases; TRF, TRF4/5 proteins; PAP euk, eukaryotic polyA polymerases; PAP bact, bacterial polyA polymerases/CCA-adding enzymes; MNT, minimal nucleotidyltransferases; TdT, terminal nucleotidyltransferases; polX, DNA polymerases of the X family; GlnD, protein uridylyl transferases; GlnE, protein adenylyl transferases; Sig-NT, new family of predicted nucleotidyltransferases involved in signal transduction; Kan-NT, kanamycin nucleotidyltransferases; Str-NT, streptomycin nucleotidyltransferases; PAP pox, poxviral polyA polymerases; CyaA, γ-proteobacterial adenylyl cyclases.

New families of nucleotidyltransferases

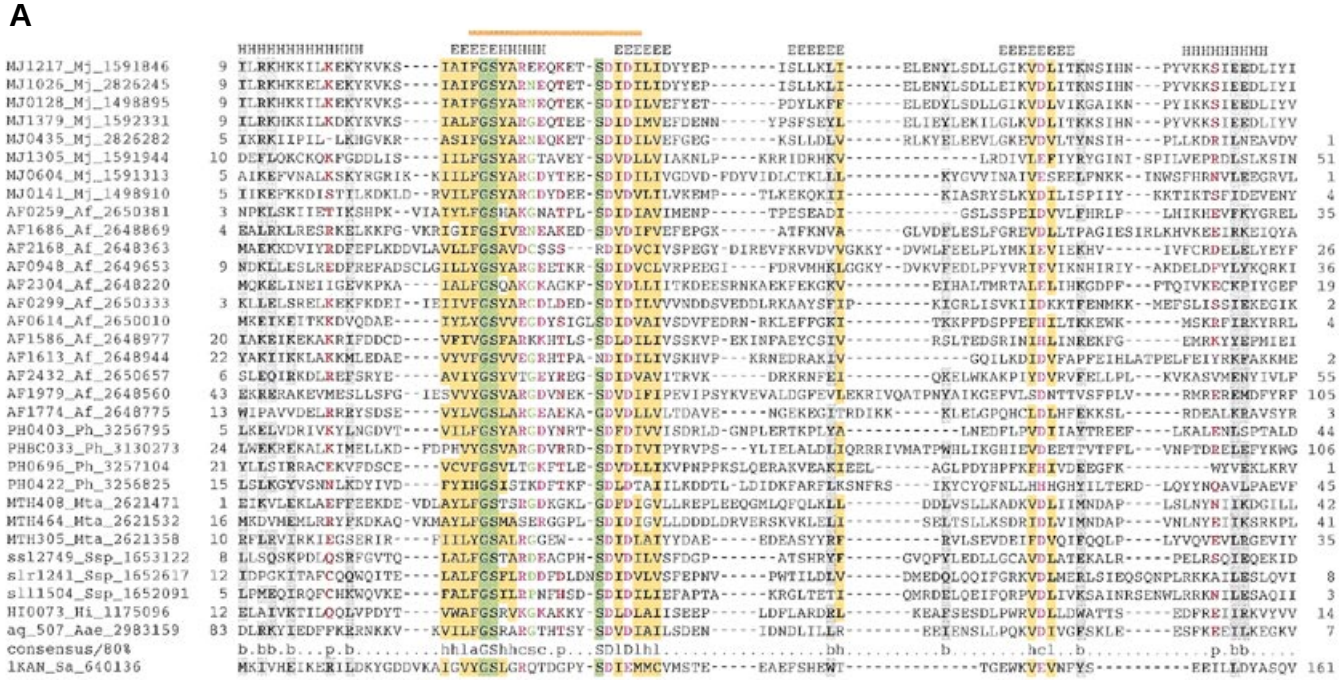
Archaeal and bacterial MNTs. In the course of previous comparative analyses of prokaryotic genomes, it was noticed that the archaeon *M.jannaschii* and some bacteria, such as *Synechocystis* sp. and *Haemophilus influenzae*, encode small proteins (85–120 amino acids) that show moderate similarity to some of the known nucleotidyltransferases and contain the nucleotidyltransferase signature motif (36). Our analysis of the three newly available archaeal genomes and completely or partially sequenced genomes of a variety of other archaea and bacteria showed that these proteins are universally present in archaea in a varying number of copies and are sporadically found in bacteria; thus far, no eukaryotic members of this family were identified (Table 1). PSI-BLAST searches with several members of this family as queries retrieved the sequences of many known polβ superfamily members with statistically significant e-values (Fig. 1). Multiple alignment of all these proteins and secondary structure predictions, followed by structural comparisons with the kanamycin nucleotidyltransferase using the Zega procedure (26), which showed significant structural similarity ($P < 10^{-6}$), suggests that they represent the minimal domain of the polβ nucleotidyltransferase superfamily (therefore we designate them 'minimal' nucleotidyltransferases, or MNT).

The conserved region of the MNTs includes approximately 90 amino acid residues which correspond to the core domain of kanamycin nucleotidyltransferase (37) (Fig. 2A) that has been structurally aligned with DNA polymerase β (1). The MNT domain consists of a poorly conserved N-terminal α-helix followed by a four-strand β-sheet, with a short α-helix inserted between strands 1 and 2, and another, variable helix, placed at different angles in different members of the superfamily, following strand 4 (Fig. 3). The glycine-rich proximal portion of the polβ signature motif is located in a 'squiggle' between strand 1 and the short inserted helix, whereas the distal DxD motif is in the beginning of strand 2; the two portions of the signature are spatially juxtaposed so that they can cooperate in holding the NTP substrate (Fig. 3). A third conserved negative charged residue is in strand 4 and is spatially very close to the DxD so that the three residues can coordinate the same metal cation (Fig. 3).

The conservation of the nucleotidyltransferase core and particularly the negatively charged metal-chelating residues (Figs 2A and 3) lead to the confident prediction that the MNTs indeed possess nucleotidyltransferase activity. Their small size, however, leaves very little beyond the core catalytic domain to help in specific substrate recognition as seen in other, larger members of the polβ superfamily. In most of the genomes that encode the MNTs, they are accompanied by another conserved, small protein (e.g. *M.jannaschii* proteins MJ1216 and MJ0127) that contains a characteristic Rx(4)HxY motif (36 and data not shown) and typically is encoded by an open reading frame adjacent to an MNT gene. This protein family shows no detectable similarity to any proteins with known functions; nevertheless, the tight correlation between this gene and the MNTs in terms of phyletic distribution and localization in the genome is suggestive of a functional interaction. Specifically, the uncharacterized small protein might function as a cofactor for the MNTs forming a complex with them and thereby providing assistance in substrate recognition. There is, so far, no clue as to the nature of this substrate; given the ubiquity of the MNTs in the archaea, elucidation of their specificity will be of particular interest.

The TRF family of eukaryotic, chromatin-associated nucleotidyltransferases. Our analysis showed that yeast TRF4 and TRF5 proteins that are involved in chromatin condensation (38) and their highly conserved homologs found in all eukaryotes belong to the polβ nucleotidyltransferase superfamily (Fig. 1). For example, a PSI-BLAST search initiated with the human polyA polymerase sequence detects the TRF4 sequence and the sequence of its homolog from *S.pombe* at the second iteration with e-values <0.001. Conversely, the TRF4 sequence hits eukaryotic PolyA polymerases with e-values ~0.001 in the first pass and detects the MNTs, 2′–5′ A synthetases and some of the aminoglycoside nucleotidyltransferases in subsequent PSI-BLAST iterations. The multiple alignment of the TRF family proteins and eukaryotic polyA polymerases contains eight conserved motifs (Fig. 2B), with the probability of occurrence by chance in the given set of proteins in the range of 10^{-4}–10^{-20} as computed using the MACAW program. The most highly conserved motif includes the polβ superfamily signature with the two metal-chelating aspartates. Motif 4 contains the conserved aspartate that corresponds to the third metal-chelating residue seen in the MNT domain (Fig. 3). The distal motifs 5–8 (Fig. 2B) are outside the minimal domain and, accordingly, are expected to belong to a distinct domain. This region of extended sequence conservation shared by the TRF protein and polyA polymerases could also be identified in the 2′–5′ A synthetases and the archaeal CCA-adding

A



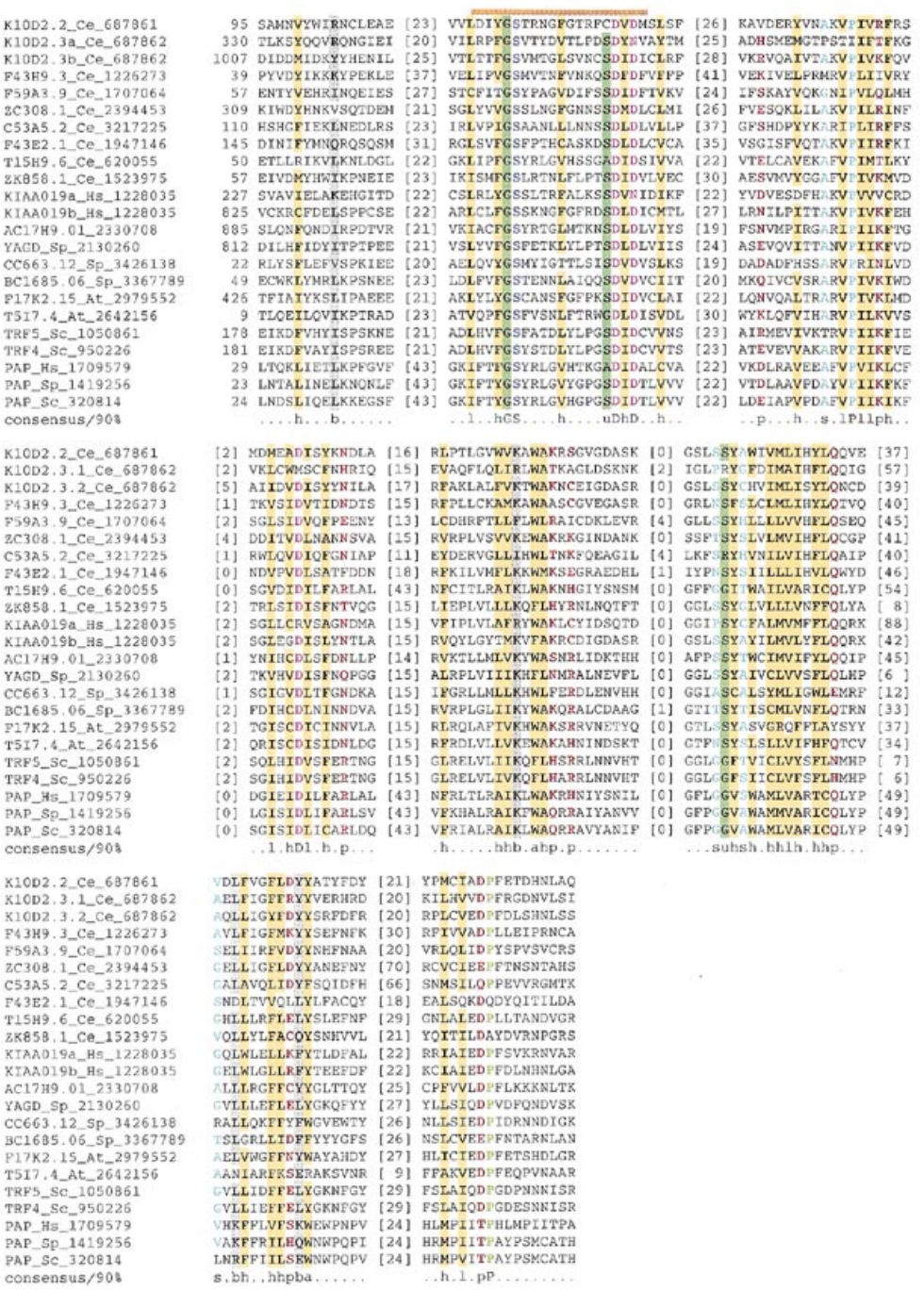
enzymes, though it is not as strongly conserved as in the former set (data not shown).

The extended sequence conservation with the polyA polymerases, which includes the intact active site, suggests that TRF proteins not only possess nucleotidyltransferase activity but, more specifically, catalyze polynucleotide synthesis. TRF4 has been originally identified as a gene whose mutation is synthetic lethal when combined with topoisomerase I mutations (39). Further studies have shown that TRF4 mutations were lethal also when combined with mutations in the SMC1 gene which encodes an ABC superfamily ATPase involved in chromosome condensation, and complex formation between Trf4p and SMC1 has been demonstrated (38). The TRF5 gene that encodes a protein closely related to Trf4p complements TRF4 mutations when overexpressed and appears to have overlapping functions since TRF4/5 double mutants are unviable (40). Phenotypic studies on these double mutants indicate that Trf4/5 proteins function in chromatin condensation and chromosome assembly during mitosis (40). A plausible role for active nucleotidyltransferases in these processes is suggested by the interaction between Trf4/5p and topoisomerase I. It seems likely that the TRF family enzymes catalyze DNA synthesis required to repair gaps that may be introduced as a result of topological manipulations during DNA condensation.

The ubiquity and high level of conservation of the TRF family nucleotidyltransferases in eukaryotes suggest that whatever the exact details of their function(s), they are essential for chromosome condensation and segregation in all eukaryotes. While *Saccharomyces cerevisiae* encodes only two TRF family nucleotidyltransferases, other eukaryotes have considerably larger numbers of these proteins (Table 1) which may be due to partial functional differentiation. The existence of such differentiation is supported by the different domain architectures found in TRF family proteins such as, for example, WD40 repeats and Zn fingers in proteins from *S.pombe* and humans, respectively (Fig. 4); these particular accessory domains may mediate the association of the nucleotidyltransferase with chromatinic complexes or directly with DNA. Furthermore, proteins from *Caenorhabditis elegans* (K10D2.3) and humans (KIAA019) show duplication of the entire nucleotidyltransferase domain, with replacements of the metal-chelating residues in motifs 2 and 4 in the N-terminal domain (Figs 2B and 4). This is reminiscent of a similar inactivation of the N-terminal copy of the nucleotidyltransferase domain in the large isoform of the 2′–5′ oligoA synthetase. These apparently inactive domains may be involved in allosteric regulation of the nucleotidyltransferase activity of these proteins. Examination of the phyletic distribution of the TRF family proteins (Table 1) shows that the common ancestor of animals, plants and fungi encoded at least three distinct forms of this nucleotidyltransferase (one apparently had been lost in the yeast lineage); furthermore, at least one copy is detectable in the genome of the earlier branching *Plasmodium falciparum*.

Putative signal transducing nucleotidyltransferases. GlnD (41) and GlnE (42) proteins are nucleotidyltransferases that participate in the regulation of glutamine synthetase in bacteria by transfer of uridylate and adenylate to tyrosine residues of GlnB and glutamine synthetase, respectively (42). We identified a third family of nucleotidyltransferases that is distantly related to GlnD and GlnE and might be involved in signaling. These proteins are encoded by several bacteria, namely *Vibrio*, *Pseudomonas*, *Rhodobacter*, *Chlorobium* and *Aquifex*, and the archaeon *Archaeoglobus fulgidus*. They are readily detected in iterative PSI-BLAST searches seeded with the sequences of GlnE and GlnD proteins. For example, a search initiated with the sequence of the nucleotidyltransferase domain of the *H.influenzae* GlnE protein recognized the sequence of the *A.fulgidus* protein from the new family with an e-value of 10^{-4} in the second iteration and retrieved all other members with e-values <0.01 in subsequent iterations. In reciprocal searches, the proteins of the new family specifically retrieved GlnE and GlnD sequences before other

B



members of the polβ superfamily. A multiple alignment of this new family with GlnD and GlnE shows the hallmark features of active nucleotidyltransferases as well as several additional conserved motifs (Fig. 2C). Given their specific relationship with GlnE and GlnD, it is likely that the predicted nucleotidyltransferases of the new family catalyze nucleotidylation of specific proteins.

In addition to the nucleotidyltransferase domain, these proteins contain N-terminal cNMP-binding and CBS domains (Fig. 4) which are typical components of signal- transducing systems (43). This association with ligand-binding domains is reminiscent of GlnD that also contains a predicted amino acid-binding domain (L.Aravind and E.V.Koonin, unpublished observations) and is regulated by glutamine (44). It is likely that the newly identified nucleotidyltransferases sense cAMP and possibly other ligands and in response to their concentrations, regulate activities of other proteins through nucleotidylation. Given the presence of these enzymes in at least two major bacterial pathogens, namely *Vibrio cholerae* and *Pseudomonas aeruginosa*, identification of their targets is of major interest.

C

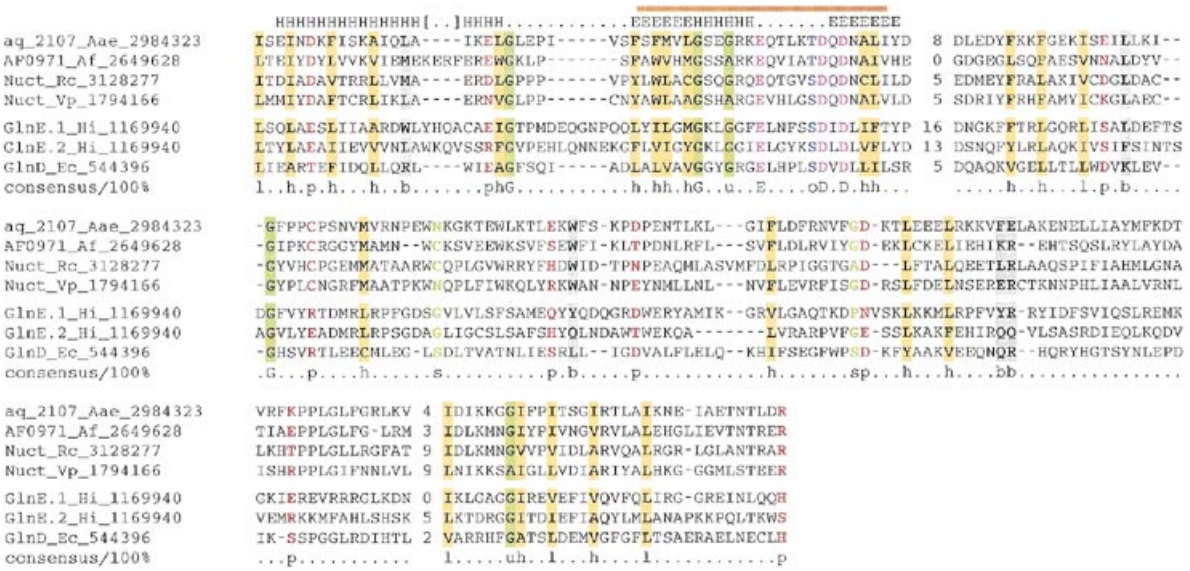


Figure 2. (Above and previous pages) Multiple alignments of new families of predicted nucleotidyltransferases. The alignments were constructed on the basis of the PSI-BLAST results using the ClustalW program. The left column includes the protein names from the SWISS-PROT database or gene names, and the Gene Identification (GI) numbers (after the underscore). The species abbreviations are: Aae, *A.aeolicus*; Af, *A.fulgidus*; Amac, *Allomyces macrogynus*; At, *Arabidopsis thaliana*; Ce, *C.elegans*; Ec, *E.coli*; Hi, *H.influenzae*; Hs, *Homo sapiens*; Mj, *Methanococcus jannaschii*; Mta, *Methanobacterium thermoautotrophicum*; Ph, *Pyrococcus horikoshii*; Rc, *Rhodobacter capsulatus*; Sc, *S.cerevisiae*; Sp, *S.pombe*; Ssp, *Synechocystis* sp.; Vp, *Vibrio parahaeomlyticus*. In each panel, a consensus derived using the indicated percentage cut-off is shown, and the respective alignment columns are highlighted using differential coloring; b, 'big' residues (E,K,R,I,L,M,F,Y,W); h, hydrophobic residues (A,C,F,I,L,M,V,W,Y); l, aliphatic residues (I,L,V,A); o, alcoholic residues (S,T); s, small residues (A,C,S,T,D,N,V,G,P); u, 'tiny' residues (G,A,S); p, polar residues (D,E,H,K,N,Q,R,S,T); c, charged residues (K,R,D,E,H). The distances from the aligned regions to the protein termini and the distances between the conserved blocks, where more variable regions were omitted [(B) only], are indicated by numbers. The principal conserved motif of the polβ nucleotidyltransferase superfamily is overlined. (**A**) Archaeal and bacterial MNTs. (**B**) Eukaryotic TRF family implicated in chromatin remodeling aligned with eukaryotic polyA polymerases. The sequence of kanamycin nucleotidyltransferase for which the crystal structure is available (PDB code 1KAN) is shown below the consensus line, and secondary structure elements derived from this structure are shown above the alignment [E indicates extended conformation (β-strand); H indicates α-helix]. PAP, polyA polymerase. (**C**) Bacterial and archaeal nucleotidyltransferases implicated in signal transduction. The upper block includes members of the new family, and the lower block includes previously identified uridylyl and adenylyl transferases.

Proteobacterial adenylyl cyclase is a divergent member of the polβ superfamily

There are three types of adenylyl cyclases in bacteria and archaea: (i) the eukaryotic type that is found fused to a variety of domains, including protein kinase domains, and is particularly abundant in *Mycobacterium tuberculosis*, Myxobacteria and Cyanobacteria (45); (ii) a small adenylyl cyclase identified in all archaea, some bacteria and animals (46; L.Aravind and E.V.Koonin, unpublished observations); and (iii) CyaA proteins from the γ division of proteobacteria (47). It has been shown that the N-terminal half of the large (approximately 840 amino acids) CyaA protein contains the catalytic domain whereas the C-terminus contains the regulatory domain that senses the environmental conditions to which the enzyme responds (48).

In the course of our analysis of the polβ nucleotidyltransferase superfamily, we noticed that certain queries, for example an *Enterococcus faecalis* aminoglycoside nucleotidyltransferase, showed limited similarity to the N-terminal region of CyaA-type adenylyl cyclases (e-values in the range of 0.14–0.5) at convergence of the iterative PSI-BLAST searches. In spite of this limited statistical significance, the alignments between nucleotidyltransferases and adenylyl cyclases span the minimal domain of the polβ superfamily and show conservation of the principal catalytic residues. This prompted us to investigate the potential relationship in more detail. As all the γ proteobacterial adenylyl cyclases are very closely related to each other, they do not form an informative profile to facilitate the detection of subtle sequence similarities. Therefore, two alternative approaches were used. A PHI-BLAST search with the *E.faecalis* nucleotidyltransferase and the polβ signature pattern as the queries detects the CyaA-type adenylyl cyclases with e-values in the range of 10^{-3}–10^{-5}. Using the ZEGA procedure (26), the probability that the *E.faecalis* nucleotidyltransferase sequence and CyaA do not adopt the same fold was estimated at 10^{-6}–10^{-7}. The multiple alignment of the proteobacterial adenylyl cyclases and nucleotidyltransferases shows not only conservation of the catalytic motifs but also of the key hydrophobic and turn positions that comprise the scaffold of β–α–β structure (Fig. 5). Taken together, this evidence suggests that γ proteobacterial adenylate cyclases indeed are distant homologs of polβ superfamily nucleotidyltransferases.

In retrospect, the relationship between nucleotidyltransferases and adenylyl cyclases is not entirely unexpected as the cyclization reaction catalyzed by the latter also involves transfer of a nucleotide moiety accompanied by the release of pyrophosphate, except in this case, the acceptor is the 3′OH of the same nucleotide. In evolutionary terms, the restricted phyletic distribution of these adenylyl cyclases suggests that they might have evolved by rapid divergence from an ancestral nucleotidyltransferase, early in the γ proteobacterial lineage.

This is the second instance when an apparent relationship between adenylyl cyclases and nucleotidyltransferases has been detected. Previously it has been shown that DNA polymerases I

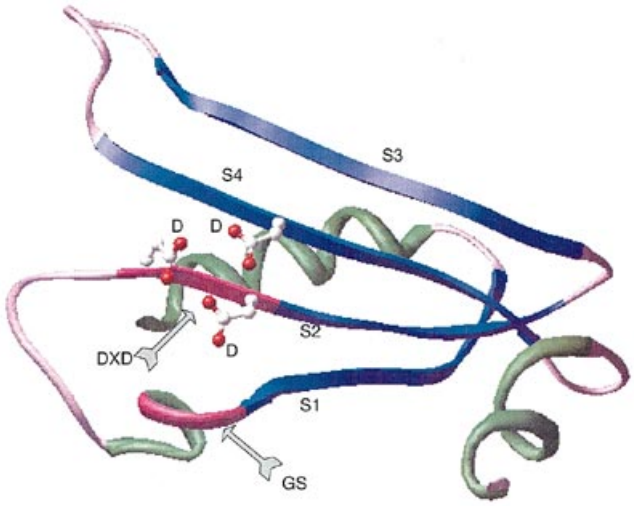


Figure 3. A structural model of the MNT domain. The sequence used for modeling was a consensus derived from the multiple alignment of the MNTs (Fig. 2A); the structure of kanamycin nucleotidyltranferase (PDB code 1kan) was as the template. The β-strands are numbered S1–S4 starting from the N-terminus. The positions of the two principal elements of the nucleotidyltransferase motif, namely the conserved glycine–serine (GS) doublet and the two conserved aspartates (DXD) are indicated. The two conserved aspartates and a third, distal aspartate that is conserved in the majority of polβ superfamily nucleotidyltransferases (Fig. 2) are shown as ball-and-stick models.

and classic 'eukaryote-type' adenylyl cyclases share a common fold and may utilize similar catalytic mechanisms (49,50). Recent structural and site-directed mutagenesis studies on the 'eukaryote-type' adenylyl cyclases have led to a suggestion of a DNA polymerase-like reaction mechanism (51). Thus adenylyl cyclases may have been derived from nucleotidyltransferases on more than one occasion in evolution, which may illustrate a general trend of the origin of signal transduction components from enzymes involved in basic processes, such as nucleic acid biosynthesis and processing.

Classification and phyletic distribution of the polβ-type nucleotidyltransferases: the evolutionary implications

All detected members of the polβ nucleotidyltransferase superfamily were classified hierarchically into groups and families using single linkage clustering with serial gapped BLAST score cut-offs in the range of 40–70 bits. A multiple alignment was constructed for each family and unique signatures (synapomorphies) were identified. The four distal motifs in the alignment of the TRF family with eukaryotic-type polyA polymerases are a clear example of such a synapomorphy (Fig. 2B). Synapomorphies also can be seen in the domain organization of some of the nucleotidyltransferase families, e.g. the newly identified family of nucleotidyltransferases implicated in signal transduction and containing a cNMP-binding domain and a CBS domain that are not found in any other nucleotidyltransferases (Fig. 4). In the absence of a sufficient number of aligned informative positions, this approach provides an alternative to conventional phylogenetic tree analysis for constructing a tentative evolutionary classification. In order to examine the phyletic distribution of the polβ superfamily, we used PSI-BLAST generated profiles for each family to extract all the family members from complete genome sequences. The families derived using these procedures and the phyletic distribution for each family are summarized in Table 1.

The striking aspect of the phyletic distribution of the nucleotidyltransferase families within the polβ superfamily is that most of them (10 of the 14 families) are confined to only one of the three divisions of life (bacteria, archaea or eukaryotes). Only the DNA polymerase X family is seen in all three divisions; even in this case, however, the family is represented (so far) in only one archaeon (*Methanobacterium thermoautotrophicum*) and the distribution in bacteria is patchy (Table 1). This suggests a major role for horizontal transfer and lineage-specific gene loss in the evolution of this family of nucleotidyltransferases. The presence of DNA polymerase X in bacterial thermophiles (*Aquifex* and *Thermus*) is compatible with the possibility of gene exchange between bacteria and archaea.

By similar logic, horizontal gene transfer seems a likely explanation for the observed phyletic distribution of the new family of putative signal-transducing nucleotidyltransferases that are found sporadically in bacteria and, so far, on a single occasion in the archaea, and the MNTs that are universal in the archaea but sporadic in bacteria (Table 1). Furthermore, given their absence in archaea, it seems likely that bacterial-type CCA-adding enzymes/polyA polymerases may have entered the eukaryotic world by horizontal transfer from organelles.

Generally, most of the families in the polβ superfamily are highly conserved but the inter-family relationships typically are distant, with only a few distinct, higher-order groups (Table 1; Fig. 1). This pattern seems to suggest a model of evolution whereby most of the families have independently and rapidly evolved from a common ancestor to occupy a particular functional niche. Such off-shoots of pre-existing families with specialized functions might have emerged also at later stages in the evolution of the polβ nucleotidyltransferase superfamily. Thus terminal deoxynucleotidyl transferases that are closely related to the DNA polymerase β family have acquired the vertebrate-specific role of generating antigen receptor diversity by template-independent nucleotide addition at the V(D)J recombination junctions (52) (we found, however, that fission yeast *S.pombe* encodes a TdT; in this case, the role of this enzyme remains unclear). Another terminal branching of this type is the 2′–5′ oligoA synthetase family that apparently had been derived from the polyA polymerases concomitantly with the origin of interferon signaling in vertebrates. In the newly described TRF family of predicted nucleotidyltransferases, a remarkable expansion is seen in multicellular eukaryotes (Table 1), which is likely to correspond to distinct and as yet unidentified functions in chromatin remodeling.

A functional, as well as an evolutionary, connection seems to exist also between the two types of nucleotidyltransferases involved in signal transduction, namely the protein uridylyl and adenylyl transferases (GlnD and GlnE), and the newly described family containing the cNMP-binding and CBS domains. Finally, it appears that the most radical and previously unsuspected transformation of the polβ-type nucleotidyltransferases, namely the evolution of proteobacterial adenylate cyclases, is yet another example of a rapid divergence linked to the emergence of a new function.

The discovery of the family of archaeal and bacterial MNTs may provide a clue as to the ancestral form of a polβ superfamily

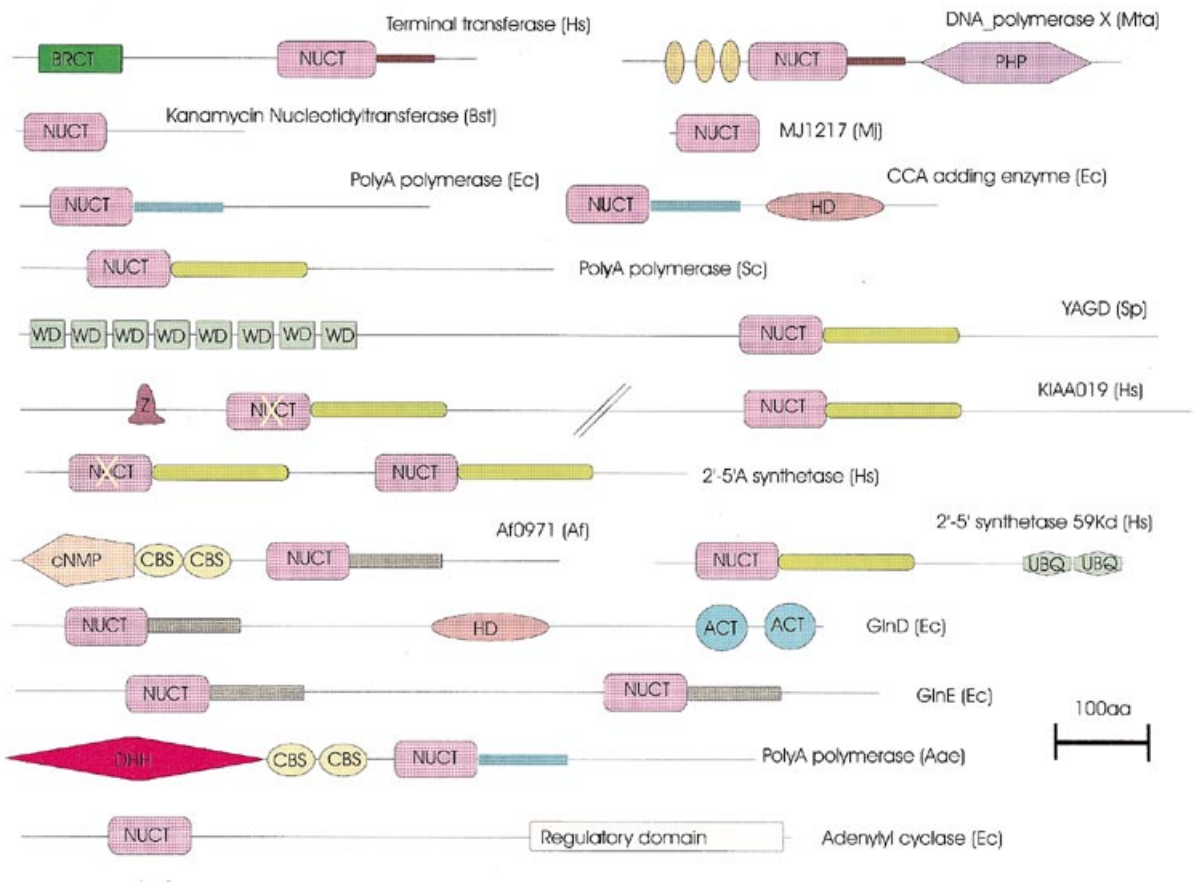
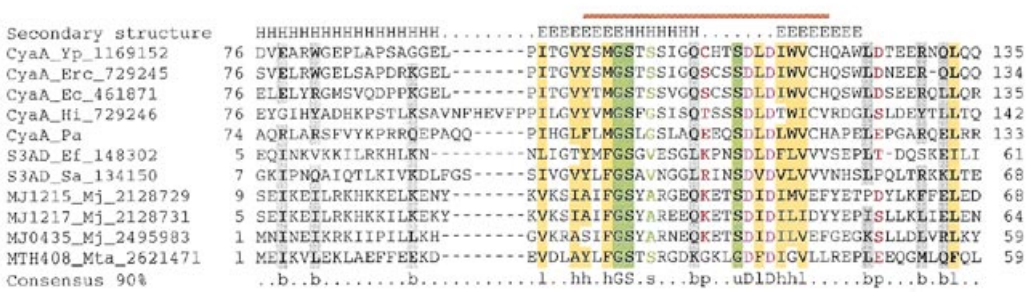


Figure 4. Distinct domain architectures of the polβ superfamily nucleotidyltransferases. The figure is roughly to scale; the double slash (//) shows that a portion of a long sequence is omitted. Domain designations: NUCT, polβ superfamily nucleotidyltransferase domain; PHP, PHP (polymerase histidinol phosphatase) superfamily phosphoesterase domain; HD, HD superfamily phosphoesterase domain; DHH, DHH family phosphoesterase domain; BRCT, BRCA1 C-terminal domain; WD, WD40 repeat; Z, Zn finger; cNMP, cNMP-binding domain; ACT, predicted ligand (probably amino acid) binding domain; UBQ, ubiquitin. The species designations are as in Figure 1.



```
Secondary structure    HHHHHHHHHHHHHHHHH.........EEEEEEEEHHHHHHHH.......EEEEEEEE
CyaA_Yp_1169152     76 DVEARWGEPLAPSAGGEL-------PITGVYSMGSTSSIGQCHTSDLDIWVCHQAWLDTEERNQLQQ 135
CyaA_Erc_729245     76 SVELRWGELSAPDRKGEL-------PITGVYSMGSTSSIGQSCSSDLDIWVCHQSWLDNEER-QLQQ 134
CyaA_Ec_461871      76 ELELYRGMSVQDPPKGEL-------PITGVYTMGSTSSVGQSCSSDLDIWVCHQSWLDSEERQLLQR 135
CyaA_Hi_729246      76 EYGIHYADHKPSTLKSAVNPHEVFPPILGVYVMGSFGSISQTSSSDLDTWICVRDGLSLDEYTLLTQ 142
CyaA_Pa             74 AQRLARSFVYKPRRQEPAQQ-----PIHGLFLMGSLGSLAQEEQSDLDLWVCHAPELEPGARQELRR 133
S3AD_Ef_148302       5 EQINKVKKILRKHILKN---------NLIGTYMFGSGVESGLKPNSDLDFLVVVSEPLT-DQSKEILI  61
S3AD_Sa_134150       7 GKIPNQAIQTLKIVKDLFGS-----SIVGVYLFGSAVNGGLRINSDVDVLVVVNHSLPQLTRKKLTE  68
MJ1215_Mj_2128729    9 SEIKEILRKHKKELKENY-------KVKSIAIFGSYARGEQKETSDIDIMVEFYETPDYLKFFELED  68
MJ1217_Mj_2128731    5 SEIKEILRKHKKILKEKY-------KVKSIAIFGSYAREEQKETSDIDILIDYYEPISLLKLIELEN  64
MJ0435_Mj_2495983    1 MNINEIKRKIIPILLKH--------GVKRASIFGSYARNEQKETSDIDILVEFGEGKSLLDLVRLKY  59
MTH408_Mta_2621471   1 MEIKVLEKLAEFFEEKD--------EVDLAYLFGSTSRGDKGKLGDFDIGVLLREPLEEQGMLQFQL  59
Consensus 90%          ..b..b.......b..........l..hh.hGS.s...bp..uDlDhhl.....bp...b.bl..
```

Figure 5. Multiple alignment of γ proteobacterial adenylyl cyclases, aminoglycoside nucleotidyltransferases and MNTs. The designations are as in Figure 2. The upper five sequences are those of γ proteobacterial adenylyl cyclases (CyaA); S3AD, spectinomycin adenylyl transferases; the four bottom sequences are those of archael MNTs. Additional species abbreviations: Erc, *Erwinia caratovora*; Ef, *E.faecalis*; Pa, *P.aeruginosa*; Sa, *S.aureus*; Yp, *Yersinia pestis*.

nucleotidyltransferase. It appears likely that these small proteins resemble the ancestral form of a polβ-like nucleotidyltransferase and, in this regard, it is of interest that in iterative database searches, the MNT sequences yielded connections with most of the other distinct protein groups within the superfamily (Fig. 1). The subsequent evolution of the polβ superfamily seems to have proceeded by accretion of additional domains. This accretion process resulted not only in the increase in the size of the nucleotidyltransferases but also in diverse domain architectures, with a variety of additional domains that possess distinct enzymatic and regulatory activities (Fig. 4). Perhaps the most notable of these architectures are the independent fusions of the nucleotidyltransfer-

ase domains with phosphoesterases of three distinct families, namely DHH (53), PHP (54) and HD (55) (Fig. 4). These fusions may be interpreted as a trend towards the evolution of bi-functional enzymes that possess both a hydrolase (nuclease) and a polymerase activities. Alternatively, as discussed previously, one of the possible functions of the phosphoesterase domains is the hydrolysis of the inorganic pyrophosphate formed during nucleotide transfer, which would drive the reaction in the direction of polymerization (54). As already mentioned, another type of domains that tend to combine with the polβ-type nucleotidyltransferases are regulatory, ligand-binding domains that link the nucleotidyltransferases to signal transduction circuits. In addition to the cNMP-binding and CBS domains found in the new family of nucleotidyltranferases, the proteins of the GlnD and GlnE families contain a distinct regulatory domain implicated in amino acid binding (the Act domain; L.Aravind and E.V.Koonin, unpublished observations). Finally, combinations with DNA-binding domains such as the helix–hairpin–helix and the C2H2 Zn finger (Fig. 4) probably help localize some of the nucleotidyltransferases on their nucleic acid substrates. Interestingly, a recently characterized member of the 2′–5′ A synthetase family contains two C-terminal ubiquitin domains which suggests interaction with the ubiquitin signaling pathway (56).

Conclusions

We showed that the polβ superfamily of nucleotidyltransferases is an ancient group of enzymes that has evolved in different directions in each of the three divisions of life which may suggest a very general function(s) in the common ancestor. These functions may have included participation in multiple processes, such as chain priming and template-dependent and template-independent chain elongation (57). The family of MNTs that is found in all archaea and some bacteria may resemble the hypothetical ancestral state. The subsequent evolution of the polβ superfamily seems to have involved rapid divergence accompanying the adaptation of distinct families to specific roles. Some of the reactions catalyzed by these nucleotidyltransferases, such as CCA addition and polyA synthesis, appear to have independently evolved more than once. We identified new families within the polβ superfamily which include, in addition to the MNTs, the family of eukaryotic proteins typified by yeast TRF4/5. The TRF family of proteins is predicted to catalyze yet unknown nucleotide polymerization reactions required for chromatin remodeling. The identification of this family that is represented by multiple members in all eukaryotes opens a new direction for experimental research into chromatin structure and dynamics. Another new family of bacterial and archaeal nucleotidyltransferases is predicted to be involved in signal transduction since in these proteins, the nucleotidyltransferase domain is combined with ligand-binding domains. The evolution of signal-transducing enzymes from those involved in replication, repair and RNA processing may be a general phenomenon as demonstrated by the detection of an apparent evolutionary connection between the polβ superfamily of nucleotidyltransferases and the γ proteobacterial adenylyl cyclases.

REFERENCES

1 Holm,L. and Sander,C. (1995) *Trends Biochem. Sci.*, **20**, 345–347.
2 Shuman,S. and Schwer,B. (1995) *Mol. Microbiol.*, **17**, 405–410.
3 Aravind,L., Leipe,D.D. and Koonin,E.V. (1998) *Nucleic Acids Res.*, **26**, 4205–4213.
4 Singh,K. and Modak,M.J. (1998) *Trends Biochem. Sci.*, **23**, 277–281.
5 Burgers,P.M.J. (1998) *Chromosoma*, **107**, 218–227.
6 Archambault,J. and Friesen,J.D. (1993) *Microbiol. Rev.*, **57**, 703–724.
7 Colgan,D.F. and Manley,J.L. (1997) *Genes Dev.*, **11**, 2755–2766.
8 Deutscher,M.P. (1990) *Methods Enzymol.*, **181**, 434–439.
9 Koonin,E.V., Gorbalenya,A.E. and Chumakov,K.M. (1989) *FEBS Lett.*, **252**, 42–46.
10 Xiong,Y. and Eickbush,T.H. (1990) *EMBO J.*, **9**, 3353–3362.
11 Player,M.R. and Torrence,P.F. (1998) *Pharmacol Ther.*, **78**, 55–113.
12 Yue,D., Maizels,N. and Weiner,A.M. (1996) *RNA*, **2**, 895–908.
13 Davies,J.F.,II, Almassy,R.J., Hostomska,Z., Ferre,R.A. and Hostomsky,Z. (1994) *Cell*, **76**, 1123–1133.
14 Sawaya,M.R., Pelletier,H., Kumar,A., Wilson,S.H. and Kraut,J. (1994) *Science*, **264**, 1930–1935.
15 Shi,P.Y., Maizels,N. and Weiner,A.M. (1998) *EMBO J.*, **17**, 3197–3206.
16 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
17 Wolf,Y.I., Brenner,S.E., Bash,P.A. and Koonin,E.V. (1999) *Genome Res.*, **9**, 17–26.
18 Karlin,S. and Altschul,S.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
19 Altschul,S.F. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.
20 Altschul,S.F. and Koonin,E.V. (1998) *Trends Biochem. Sci.*, **23**, 444–447.
21 Wootton,J.C. (1994) *Comput. Chem.*, **18**, 269–285.
22 Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol.*, **266**, 554–571.
23 Lupas,A. (1996) *Methods Enzymol.*, **266**, 513–525.
24 Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) *Nucleic Acids Res.*, **26**, 3986–3991.
25 Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
26 Abagyan,R.A. and Batalov,S. (1997) *J. Mol. Biol.*, **273**, 355–368.
27 Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
28 Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) *Proteins*, **9**, 180–190.
29 Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) *Protein Sci.*, **4**, 1618–1632.
30 Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.
31 Walker,D.R. and Koonin,E.V. (1997) *ISMB*, **5**, 333–339.
32 Rost,B. and Sander,C. (1994) *Proteins*, **19**, 55–72.
33 Peitsch,M.C. (1996) *Biochem. Soc. Trans.*, **24**, 274–279.
34 Guex,N. and Peitsch,M.C. (1997) *Electrophoresis*, **18**, 2714–2723.
35 Gershon,P.D., Ahn,B.Y., Garfield,M. and Moss,B. (1991) *Cell*, **66**, 1269–1278.
36 Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) *Mol. Microbiol.*, **25**, 619–637.
37 Sakon,J., Liao,H.H., Kanikula,A.M., Benning,M.M., Rayment,I. and Holden,H.M. (1993) *Biochemistry*, **32**, 11977–11984.
38 Castano,I.B., Brzoska,P.M., Sadoff,B.U., Chen,H. and Christman,M.F. (1996) *Genes Dev.*, **10**, 2564–2576.
39 Sadoff,B.U., Heath-Pagliuso,S., Castano,I.B., Zhu,Y., Kieff,F.S. and Christman,M.F. (1995) *Genetics*, **141**, 465–479.
40 Castano,I.B., Heath-Pagliuso,S., Sadoff,B.U., Fitzhugh,D.J. and Christman,M.F. (1996) *Nucleic Acids Res.*, **24**, 2404–2410.
41 Garcia,E. and Rhee,S.G. (1983) *J. Biol. Chem.*, **258**, 2246–2253.
42 Kustu,S., Hirschman,J., Burton,D., Jelesko,J. and Meeks,J.C. (1984) *Mol. Gen. Genet.*, **197**, 309–317.
43 Bateman,A. (1997) *Trends Biochem. Sci.*, **22**, 12–13.
44 Jiang,P., Peliska,J.A. and Ninfa,A.J. (1998) *Biochemistry*, **37**, 12782–12794.
45 Katayama,M. and Ohmori,M. (1997) *J. Bacteriol.*, **179**, 3588–3593.
46 Sismeiro,O., Trotot,P., Biville,F., Vivares,C. and Danchin,A. (1998) *J. Bacteriol.*, **180**, 3339–3344.
47 Mock,M., Crasnier,M., Duflot,E., Dumay,V. and Danchin,A. (1991) *J. Bacteriol.*, **173**, 6265–6269.
48 Crasnier,M., Dumay,V. and Danchin,A. (1994) *Mol. Gen. Genet.*, **243**, 409–416.
49 Artymiuk,P.J., Poirrette,A.R., Rice,D.W. and Willett,P. (1997) *Nature*, **388**, 33–34.
50 Murzin,A.G. (1998) *Curr. Opin. Struct. Biol.*, **8**, 380–387.
51 Zimmermann,G., Zhou,D. and Taussig,R. (1998) *J. Biol. Chem.*, **273**, 19650–19655.
52 Yang,B., Gathy,K.N. and Coleman,M.S. (1994) *J. Biol. Chem.*, **269**, 11859–11868.
53 Aravind,L. and Koonin,E.V. (1998) *Trends Biochem. Sci.*, **23**, 17–19.
54 Aravind,L. and Koonin,E.V. (1998) *Nucleic Acids Res.*, **26**, 3746–3752.
55 Aravind,L. and Koonin,E.V. (1998) *Trends Biochem. Sci.*, **23**, 469–472.
56 Hartmann,R., Olsen,H.S., Widder,S., Jorgensen,R. and Justesen,J. (1998) *Nucleic Acids Res.*, **26**, 4121–4128.
57 Maizels,N. and Weiner,A.M. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 6729–6734.