DNA polymerase β belongs to an ancient nucleotidyltransferase superfamily

DNA polymerases are classified into four families on the basis of sequence similarities[1]. DNA polymerase β, from the eukaryotic 'X' family of DNA polymerases, is smaller (40 kDa) and simpler than the other polymerases. It operates in a stepwise rather than processive fashion, dissociating from the template–primer after the addition of each nucleotide. Here, we show by structure and sequence comparison that several remote relatives of DNA polymerase β form a new superfamily of nucleotidyltransferases involved in diverse biological functions that range from DNA repair to regulation of biosynthetic pathways and antibiotic resistance.

The structure of polymerase β was solved recently[2,3] and, surprisingly, it was different from the other known RNA or DNA polymerase structures, which had demonstrated a common shape with subdomains called 'palm', 'fingers' and 'thumb' by analogy to a right hand holding a rod of DNA[4]. Instead, the FSSP database[5] of structurally similar proteins reports a highly significant structural match for the catalytic domain of DNA polymerase β with that of kanamycin nucleotidyltransferase (Fig. 1). The kanamycin nucleotidyltransferase structure[6] was solved at low resolution but the putative active site, with a bound Zn^{2+}, is structurally equivalent to that seen in the ternary-complex structure of DNA polymerase β[2]. As both enzymes catalyse similar chemical reactions, the strong structural similarity suggests an evolutionary relationship between them, despite a very low sequence identity of 10% over 81 structurally equivalent residues.

To further investigate the evolutionary aspects of this unexpected three-dimensional structural similarity, the Prosite[7] signature for DNA polymerase β was relaxed just enough to recognize also kanamycin nucleotidyltransferase. Surprisingly, a scan of the SWISS-PROT database using the pattern G[SG][LIVMFY]xR[GQ]$x_{5,6}$D[LIVM][DE]-[CLIVMFY]$_3$ additionally picked up another family of nucleotidyltransferases, the protein-P_{II} uridylyltransferase. This enzyme reversibly uridylates the small bacterial protein P_{II}, which acts in the control of the activity of glutamine synthetase. The sequences of these uridylyltransferases are fully consistent with the structural alignment of the catalytic domains of DNA polymerase β and kanamycin nucleotidyltransferase (Fig. 2). Further refinement cycles of profile/pattern searches identified four more families of nucleotidyltransferases as probable structural and functional homologs in the emerging superfamily. The biological functions of the enzyme families are summarized in Table I. All catalyse the same chemical reaction, the coupling of nucleoside triphosphates to a free hydroxyl group via elimination of pyrophosphate (the target group on the modified protein is not known for the GlnD or GlnE enzymes). Moreover, secondary structure predictions[8] for the individual families were consistent with the proposed fold.

Mapping sequence conservation derived from the multiple alignment of all seven nucleotidyltransferase families onto the known structures reveals the main determinants of the active-site motif (Fig. 2). (1) A hydrophobic stripe along one side of the first helix provides structural support for the β-sheet. (2) The β-strands in the β-hairpin that forms the base for the phosphate-binding loop are hydrophobic. The residue between the conserved carboxylates, at the beginning of the second strand, points inwards to the core, and always has an aliphatic sidechain (or Met). (3) The short helix in the phosphate-binding loop starts immediately after the first β-strand. The peptide group of the invariant glycine at the junction is in *cis* conformation in the crystal structure of rat DNA polymerase β. (Only the Cα coordinates are deposited for kanamycin nucleotidyltransferase.) The short helix ends in strong capping residues (such as glycine). (4) The conserved aspartates (glutamates) bind a divalent cation, which is essential for catalysis. Mutation experiments with rat DNA polymerase β have shown that glutamate is not tolerated in place of D190 or D192. Aspartate is indeed strictly invariant in the first position, but in

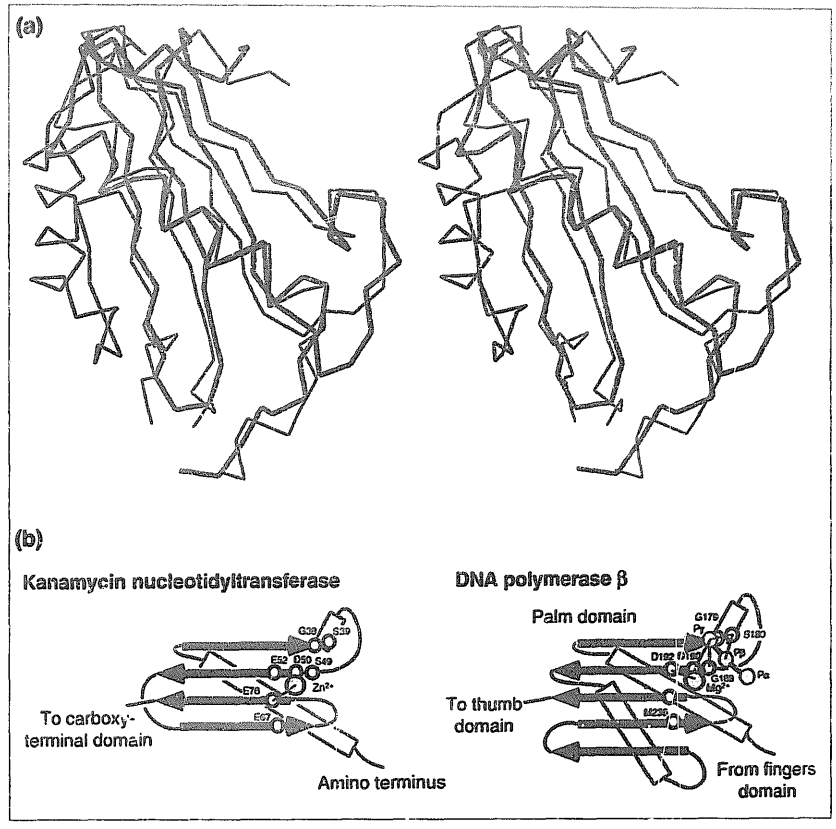

Figure 1

Structure comparison of kanamycin nucleotidyltransferase and DNA polymerase β. (a) Stereo superimposition of the Cα traces of the common catalytic domains of kanamycin nucleotidyltransferase (thicker line) and DNA polymerase β (thinner line). Eighty-one pairs of Cα atoms are superimposed with a root-mean-square positional deviation of 2.3 Å. Drawn in Whatif[10]. (b) Topology diagrams of the two enzymes. The structural core of the common catalytic domain contains a single helix packing against a four-stranded β-sheet with unusual (mixed antiparallel and parallel) topology (red). At the connection between the second and third β-strands of kanamycin nucleotidyltransferase, the DNA polymerase has a large insertion consisting of one helix and the fifth, edge-strand of the sheet (blue). The most strongly conserved sequence positions map to the region between the two first β-strands, including the structurally conserved short helix (green). Open rings show the approximate positions of important residues. Kanamycin nucleotidyltransferase has a Zn^{2+} in the middle of a crown of negatively charged residues (D50, E52, E67 and E76); the carboxy-terminal domain of the other subunit in a homodimer was also implicated in Zn^{2+} binding (E141, E142, E145)[6]. The structure of DNA polymerase was solved at higher resolution, revealing the positions of the required Mg^{2+} and triphosphate[2]. Thin lines denote protein–ligand interactions.

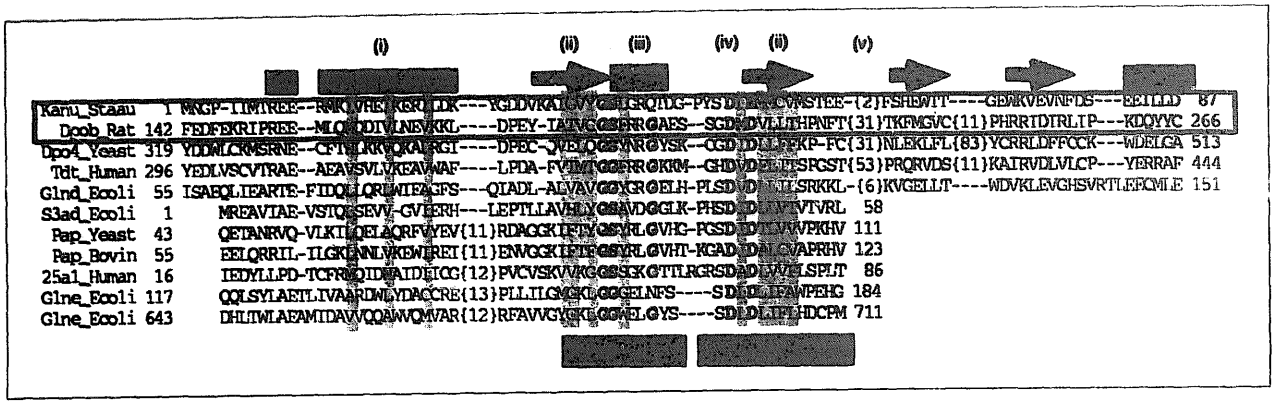

Figure 2

Multiple alignment of the catalytic domains of nucleotidyl transferases taken from representative sequences of each subfamily. The alignment of the top pair of sequences (boxed) was determined structurally using the Dali program[11]. The other sequences are aligned using only sequence information[12]. The sequences are labelled by their SWISS-PROT[13] codes. Only part of each sequence is shown; residue numbers are given for the ends. Dashes denote deletions; the number of inserted residues is shown inside curly brackets. Sites labelled (i)–(iv) are discussed in the text. Site (v) is the 31-residue insertion in DNA polymerase β (Dpob_Rat) that contains an α–β unit not present in kanamycin nucleotidyltransferase (Kanu_Staau; blue segment in Fig. 1). Especially strongly conserved columns are boxed; grey shading denotes conserved hydrophobicity. The secondary structure in the crystal is given at the top (grey bars: helix; arrows: strand). The black boxes at the bottom indicate the region used to derive a profile using the sequences of subfamilies 1–3 in Table I (top five sequences in this figure and their relatives). This profile specifically recognizes families 4–7 of Table I (bottom six sequences and their relatives).

Table I. Nucleotidyl transferase subfamilies identified by sequence-pattern search

Subfamily	SWISS-PROT/PDB identifiers	Size (aa)	Biological function[a]	E.C. number	Rank[b]
1 Kanamycin nucleotidyltransferase	kanu_staau, _bacsp 1KAN	253	Inactivates a wide range of aminoglycosides used as antibiotics, using ATP, GTP or UTP as the nucleoside monophosphate donor[10].	2.7.7.-	[5,6]
2 Polymerase family X					
• DNA polymerase β, yeast DNA polymerase IV	dpob_rat, _human, dpo4_yeast 1BPB, 1RPL	334	Repair and gap filling in DNA synthesis.	2.7.7.7	[10,16,18]
• DNA polymerase β-like protein	o174_asfb7	174	Hypothetical protein from African swine fever virus.		[7]
• DNA nucleotidyl exotransferase	tdt_human, _bovin, _mouse, _chick, _xenla	506–520	Catalyses the random addition of deoxynucleoside 5′-triphosphate to the 3′-end of a DNA initiator. For example, it adds nucleotides at the junction of rearranged immunoglobulin heavy chain and T-cell receptor gene segments during the maturation of B and T cells.	2.7.7.31	[8,9,11,14,15]
3 Protein-P_{II} uridylyltransferase	glnd_ecoli, _salty, _klepn, nfrx_azovi	887–899	Modify the P_{II} (GlnB) regulatory protein by (de)uridylylation. P_{II} indirectly controls the transcription of the glutamine synthase gene (*glnA*).	2.7.7.59	[1–4]
4 Streptomycin 3′-adenylyltransferase	s3ad_ecoli, _klepn, _entfa, _staau, _agrtu	255–263	Mediate bacterial resistance to the antibiotics streptomycin and spectomycin.	2.7.7.47	[12,13,19, 20,32]
5 Poly(A) polymerase	pap_yeast, _bovin	568, 738	Creates the 3′-poly(A) tail of mRNAs. Recognizes the AAUAAA sequence in mRNA.	2.7.7.19	[17,22]
6 (2′–5′) oligoadenylate synthetase	25a1_human, _mouse, _rat, 25a2_human, _mouse, 25a3_mouse, 25a6_human	358–414, 726	Induced by interferons; play an important role in mediating resistance to virus infection via activation of RNase L, which cleaves single-stranded RNAs.	2.7.7.-	[23–26,30, 31,43]
7 Glutamine synthase adenylyltransferase	glne_ecoli	945	Regulates glutamine synthase by adenylylation and deadenylylation.	2.7.7.42	[27]

[a]Biological functions taken from the SWISS-PROT database where not indicated otherwise.

[b]Ranks in profile search against the SWISS-PROT database. The profile was constructed and generalized for the block marked by thick bars in Fig. 2 using GCG programs. Ranks for sequences used to generate the profile are in brackets. No generalization of the profile was done on the four most strictly conserved positions G[SG]...D.[DE] (scores were set to +150 for allowed amino acid type, otherwise −100; the scale of scores in the unmodified profile was −100 to +150). The profile was generated using sequence weights that gave each of the subfamilies 1–3 a total weight equal to one. The gap open and elongation penalties were 4.5 and 0.15. SWISS-PROT accession numbers of the sequences follow. Subfamily **1**: P05057, P05058; **2**: P06766, P06746, P25625, P42494, P04053, P06526, P09838, P36195, P42118; **3**: P27249, P23679, P41393, P36223; **4**: P04826, P08881, Q07448, P04827, P14511; **5**: P29468, P25500; **6**: P00973, P11928, Q05961, P04820, P29080, P29081, P29728; **7**: P30870. PDB, Protein Data Bank.

kanamycin nucleotidyltransferase the second position has become a glutamate.

None of the relationships between the seven nucleotidyltransferase families identified as sharing a common catalytic domain is detected using standard sequence-search methods[9]. In this case, the structural alignment of two remote relatives led to a particularly successful generalization of the Prosite signature for the polymerase β family because the conserved structural motif characteristic of the superfamily contains two regions with a variably sized gap in between (Fig. 2). Importantly, all elements essential for catalysis and plausibly required for the domain fold are conserved in the families reported here. All seven enzyme families are therefore expected to achieve catalysis by the same conserved mechanism. The discovery that DNA polymerase β has relatives not directly involved in DNA synthesis calls for caution in drawing mechanistic parallels between DNA polymerase β and the other three DNA/RNA polymerases of known structure.

References

1 Ito, J. and Braithwaite, D. (1991) *Nucleic Acids Res.* 19, 4045–4047
2 Sawaya, M. *et al.* (1994) *Science* 264, 1930–1935
3 Davies, J. *et al.* (1994) *Cell* 76, 1123–1133
4 Moras, D. (1993) *Nature* 364, 572
5 Holm, L. and Sander, C. (1994) *Nucleic Acids Res.* 22, 3600–3609
6 Sakon, J. *et al.* (1993) *Biochemistry* 32, 11977–11984
7 Bairoch, A. and Boeckmann, B. (1992) *Nucleic Acids Res.* 20, 2019–2022
8 Rost, B. and Sander, C. (1993) *Proc. Natl Acad. Sci. USA* 90, 7558–7562
9 Scharf, M. *et al.* (1994) in *Second International Conference on Intelligent Systems for Molecular Biology* (Altman, R. *et al.*, eds), pp. 348–353, AAAI Press
10 Vriend, G. (1990) *J. Mol. Graphics* 8, 52–56
11 Holm, L. and Sander, C. (1993) *J. Mol. Biol.* 233, 123–138
12 Gribskov, M., McLachlan, M. and Eisenberg, D. (1987) *Proc. Natl Acad. Sci. USA* 84, 4355–4358
13 Bairoch, A. (1992) *Nucleic Acids Res.* 20, 2013–2018

LISA HOLM AND CHRIS SANDER

EMBL-HD, D-69012 Heidelberg, Germany.

The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors

Intracellular proteins involved in regulatory and signal transduction processes frequently contain regions of localized similarity to otherwise unrelated proteins. These homology domains usually mediate the specific interaction with other proteins or subcellular structures. Only the relatively well-conserved examples are detectable by conventional pairwise sequence comparison techniques. The discovery of the more divergent functional domains requires application of specialized methods with increased sensitivity, together with a rigorous statistical evaluation of the results.

Within the family of the forkhead-type transcription factors, sequence conservation is usually limited to the actual DNA-binding domain. Starting from the observation of an additional conserved region found in a subset of forkhead-family proteins, we applied the generalized sequence profile method[1] to find further instances of this domain, which we named the forkhead-associated (FHA) domain. We were able to detect FHA domains in at least 20 otherwise unrelated proteins; most are from yeast but there are also some from mammals and bacteria. The typical FHA domain comprises approximately 55–75 amino acids and contains three highly conserved blocks separated by more divergent spacer regions (Fig. 1). Only three positions seem to be invariant, including a histidine residue in the central part of the domain, which might be functionally important. The boundaries of the domain are not clearly defined; several subfamilies have short extensions

			β		β	β		β	β	
MNF1	MM	107	VTIGRNSS	8	GLSSFISRRHLQLSFQ	3	FYLRC.LGKNGVFVDGAFQ	sp:P42128
FHL1	SC	300	AIIGRRSE	11	GPSKSISRRHAQIFYN	5	FELSI.IGKNGAFVDDIFV	sp:P39521
FKH1	SC	76	VTIGRNTD	20	GPAKIVSRKHAAIRFN	5	WELQI.FGRNGAKVNFRRI	sp:P40466
FKH2	SC	83	VSIGRNTD	23	GPAKVVSRKHAIIKYN	5	WELHI.LGRNGAKVNFQRT	sp:P41813
est	AT	54	IILGRNSK	11	GGGMNISRNHARVFYD	5	FSLEV.LGKNGCLVEGVLH	gb:T20592
YHR5	SC	189	IIIGRYTE	17	FKSKVISRTHGCFKVD	5	WFLKDVKSSSGTFLNHQRL	sp:P38823
fraH	AS	204	VHIGKPND	11	ANSEIVSRVHADIRLE	4	HYIEDVGSSNGTYINNLPL	gb:U14553
DUN1	SC	56	TTIGRSRS	4	LSEPDISTFHAEFHLL	11	INVID.KSRNGTFINGNRL	sp:P39009
SPK1a	SC	66	WTFGRNPA	5	GNISRLSNKHFQILLG	4	LLLND.ISTNGTWLNGQKV	sp:P22216
SPK1b	SC	601	FFIGRSED	4	IEDNRLSRVHCFIFKK	18	DIWYCHTGTNVSYLNNNRM	sp:P22216
cds1	SP	60	WGFGRHKS	4	LNGPRVSNFHFEIYQG	11	VFLHD.HSSNGTFLNFERL	gb:X85040
MEK1	SC	47	VKVGRNDK	5	LTNPSISSVHCVFWCV	9	FYVKD.CSLNGTYLNGLLL	sp:P24719
Ki67	HS	27	CLFGRGIE	4	IQLPVVSKQHCKIEIH	4	AILHNFSSTNPTQVNGSVI	gb:X65550
KAPP	AT	209	VKLGRVSP	4	LKDSEVSGKHAQITWN	6	WELVDMGSLNGTLVNSHSI	gb:U09505
pks1R	STC	130	IRLGRSAD	4	LDDPDVSRMHCAVTVG	5	VSVADLGSTNGTTLDGTRV	gb:D26539
L8083.1	SC	99	LKLGRPVA	30	FDSRVLSRNHALLSCD	6	VYIRDLKSSNGTFINGQRI	gb:U19027
9346.10	SC	185	LKLGRPVT	25	FDSRVLSRNHACLSCD	5	IYIRDLKSSNGTFVNGVKI	gb:Z48784
L94705.22	SC	118	ITVGRNSS	6	CKNKFISRVHASITYL	5	VKIHC.FSMNGLIVTYRKQ	gb:U17246
c01g6.5	CE	39	KSFGRATT	13	LEPQFISRCHARVHHT	7	YLVED.ISENGTYINDRRL	gb:Z35595
zk632.2	CE	108	VVIGRIKP	5	MEHPSISRYHCILQYG	11	WHIFELGSTHGSRMNKKRL	sp:P34648

Figure 1

Alignment of all detected FHA domains. Only sequences matching established profiles with an error probability $P < 0.03$ were accepted during iterative profile refinement and are shown here. Error probabilities were estimated as described in Box 1. Amino acids are coloured according to their physicochemical properties. Numbers on the left indicate the position of the domain within the protein sequence; numbers in the alignment indicate the length of the omitted non-conserved regions. SWISS-PROT (sp) or GenBank (gb) Accession Nos of the sequence entries are shown on the right. The black bars above the alignment indicate the positions of the predicted β-strands.