

An attempt to unify the structure of polymerases

Marc Delarue, Olivier Poch¹, Noel Tordo², Dino Moras and Patrick Argos³

Departments of Crystallography and ¹Biochemistry, IBMC du CNRS, 15 rue Rene Descartes, 67000 Strasbourg, ²Service Recherche, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France and ³European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, FRG

With the great availability of sequences from RNA- and DNA-dependent RNA and DNA polymerases, it has become possible to delineate a few highly conserved regions for various polymerase types. In this work a DNA polymerase sequence from bacteriophage SPO2 was found to be homologous to the polymerase domain of the Klenow fragment of polymerase I from *Escherichia coli*, which is known to be closely related to those from *Staphylococcus pneumoniae*, *Thermus aquaticus* and bacteriophages T7 and T5. The alignment of the SPO2 polymerase with the other five sequences considerably narrowed the conserved motifs in these proteins. Three of the motifs matched reasonably all the conserved motifs of another DNA polymerase type, characterized by human polymerase α . It is also possible to find these three motifs in monomeric DNA-dependent RNA polymerases and two of them in DNA polymerase β and DNA terminal transferases. These latter two motifs also matched two of the four motifs recently identified in 84 RNA-dependent polymerases. From the known tertiary architecture of the Klenow fragment of *E. coli* pol I, a spatial arrangement can be implied for these motifs. In addition, numerous biochemical experiments suggesting a role for the motifs in a common function (dNTP binding) also support these inferences. This speculative hypothesis, attempting to unify polymerase structure at least locally, if not globally, under the pol I fold, should provide a useful model to direct mutagenesis experiments to probe template and substrate specificity in polymerases.

Key words: catalytic domain/DNA polymerases/RNA polymerases/sequences/structure

Introduction

The number of available protein sequences is growing rapidly due to the facility of nucleotide sequencing techniques. One protein class often studied and sequenced is the polymerase family which is central to the duplication and expression of genes. Polymerases can use RNA or DNA as a template (RNA- or DNA-dependent); the product can also be RNA or DNA. Polymerases are found both in eukaryotes and prokaryotes, though sequencing efforts have often concentrated on those from viruses. One way to use the information contained in all these sequences is to try to align them and thereby allocate them amongst related families and subfamilies. This has been achieved for DNA-dependent DNA polymerases, where three main subfamilies have been identified. One of them contains the Klenow fragment of *Escherichia coli* polymerase I, whose three-dimensional structure is known (Ollis *et al.*, 1985a), and

polymerases from phages T7 (Ollis *et al.*, 1985b; Argos *et al.*, 1986) and T5 (Leavitt and Ito, 1989), and from *Thermus aquaticus* (Lawyer *et al.*, 1989) and *Staphylococcus pneumoniae* (Lopez *et al.*, 1989). This family will be referred to as the pol I family. For another set of DNA-dependent DNA polymerases (Wang *et al.*, 1989), as those homologous to the human polymerase α (hereafter referred to as pol α s), more than 10 sequences from various species are known. A third subfamily of DNA-dependent DNA polymerase, hereafter called the pol β type, has only two members: DNA polymerase β (Matsukage *et al.*, 1987) and terminal transferase (Peterson *et al.*, 1985; Koiwai, 1986; Zmudzka *et al.*, 1986). Until now, just one DNA-dependent DNA polymerase sequence, from the SPO2 bacteriophage (Raden and Rutberg, 1984; Jung *et al.*, 1987), resisted alignment with any of the three aforementioned types.

Clearly, the aim of these alignments, apart from evolutionary implications, is the identification of the regions essential for polymerase function, since these sequence segments should appear as the most conserved. Generally, a great number of aligned sequences will ensure sufficient variability to identify the functionally required regions. For the pol I type, the five previously aligned (Ollis *et al.*, 1985b; Argos *et al.*, 1986; Leavitt and Ito, 1989; Lopez *et al.*, 1989) sequences are sufficiently close so as not to allow confident delineation of the absolutely required motifs.

In the present work, it has been found that the polymerase from bacteriophage SPO2 can be aligned, using a sensitive method, with the polymerase portion of the Klenow fragment in the C-terminal part of the protein [the N-terminal domain has a 3'-5' exonuclease function (Freemont *et al.*, 1986)]. The total alignment of the C-terminal part of the six proteins of the pol I type is presented. The relatedness of SPO2 polymerase with those from phages T7 and T5, *S. pneumoniae*, *T. aquaticus* and *E. coli* is sufficiently distant that highly conserved regions can now be reduced to five in number. Interestingly, three of the five regions match reasonably with the three most conserved motifs of DNA-dependent DNA pol α s, suggesting that the two polymerase types may share a common tertiary fold, or at least contain similar local tertiary architecture required for similar functions. These motifs are likely to represent modules required for the polymerase structure and activity.

Searches were then performed to detect such sequence patterns in other polymerase families. All three motifs could be found in DNA-dependent RNA polymerases that consist of only one subunit (see Masters *et al.*, 1987); two motifs were found in pol β s, in the same linear arrangement and maintaining the strictly conserved residues. In addition to this, an examination of several aligned RNA-dependent RNA polymerases as well as reverse transcriptases, for which four highly conserved motifs have been highlighted (Kamer and Argos, 1984; Poch *et al.*, 1989), allows the suggestion that two of their motifs could match the two motifs shared by DNA pol Is, pol α s and pol β s. These sequence similarities are further supported by a statistically significant alignment between the entire polymerase domain of two members of these different families, namely a DNA-dependent DNA

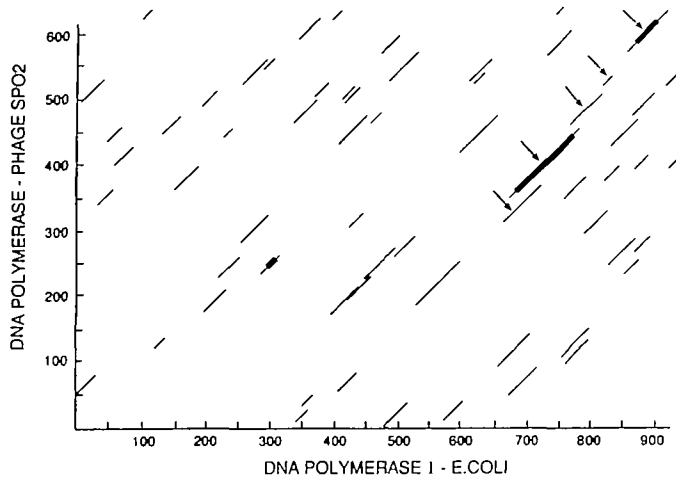


Fig. 1. Homology search matrix (Argos, 1987) between the Klenow fragment of DNA poly I from *E. coli* and DNA polymerase from bacteriophage SPO2. The search window lengths ranged from 5 to 35 in steps of 2; the search peaks are plotted over the entire window length, with the largest value dominating when overlap occurs from the multilength windows. The peak values (S) are scored as a number of standard deviations (σ) above the matrix mean for each window length. Thin lines indicate $4.5\sigma < S < 5.5\sigma$ and thick lines refer to $5.5\sigma < S < 6.0\sigma$. The path used as a basis to effect the alignment of Figure 2 is indicated by arrows. The remaining alignments shown in Figure 2 are founded on simultaneous inspection of all six sequences.

polymerase and an RNA-dependent DNA polymerase. The alignment of the C-terminal most motif, containing two acidic groups, is in agreement with those suggested by Argos (1988) amongst RNA-dependent RNA polymerases, reverse transcriptases, and DNA-dependent DNA pol α s.

An examination of these conserved regions relative to the known Klenow fragment tertiary architecture positions them in areas likely to interact with nucleotides. In fact, many biochemical data (Joyce and Steitz, 1987; Larder et al., 1987b) point to the importance of these motifs in the catalytic process, in the different families of polymerases. The design of site-directed mutagenesis experiments should benefit from the model suggested here, implying possible common tertiary segments for various polymerases, if not a common topology.

Results

Pairwise sequence comparisons were done by a search procedure based on residue characteristics (Argos, 1987). The resulting alignments of the C-terminal polymerase portion of the Klenow fragment of DNA-dependent DNA polymerase I from *E. coli* and polymerases from bacteriophages T7, T5, *S. pneumoniae* and *T. aquaticus* are similar to the ones given by Argos et al. (1986), Leavitt and Ito (1989), Lopez et al. (1989) and Lawyer et al. (1989) respectively. The DNA polymerase sequence from bacteriophage SPO2 was also compared to the *E. coli*, *S. pneumoniae*, T5 and T7 polymerases. The search matrix and

E. COLI	NHSEELTLRL	AELEKKAHEI	AGREFNLSST	KQLQTLIFEK	QGIRPL...K	KTFGGAPSTS	EEVLEELALD	YPLPKVILEY	R. GLAKLK...	...STYTDKL
S. PNEU	EMQAEDELVI	ENLTOEYIEL	AGREFVNSP	KQLGVLLFEK	LGL. PLEYTK	KTKTG. YSTA	VDVLERLAPI	AFIVKILIDY	R. QIAKIQ...	...STYVIGL
T. AQUA	ALSLEVAEEI	ARLEAEVFRL	AGHPFHLNER	DQLERVLFDE	LGL. PAL. GK	TEKTKGRSTS	AAVLEALREA	HPIVEKILQY	R. ELTKLK...	...STYIDPL
SPO2	KHALACDEQY	QAGLIAEAKK	LTGLPHFNSI	AQLKQVLEEK	.GLTISSLAK	DK.....I	EELIENTNDE	TV. HRVLR	RQMART...	...SVKRYLA
T5	NLANKAREKLY	TYPEVKOLEQ	DNEAFVNSV	KQLRVLLFDY	VGLTPT. .GK	LIDTGDADST	AEALNELATQ	HP IAKTLLI	R. KLTKLI...	...STYVEKI
T7	REPCELDTRE	YVAGAPYTPV	EHVVTFPSR	DHIOKQLOEA	.GVVPT. .IK	YTDKGAPVVD	DEVLEGVVVD	DPEKQAAIDL	IKKEYIMIGKR	IQQBAEGDKA

E. COLI	PLMINPKTGR	VHTSYBQ. AV	TATGRLSSTD	P. NLQNI. V	RNEEGRRIQ	AF.....	I	APEDYVTVSA	DYSQIELRIM
S. PNEU	QDWILA. DGK	IHTRYVQ. DL	TQTGLSSVD	P. NLQNI. A	RLBQGRIRK	AFV.....	P	ENEDSVLLSS	DYSQIELRVL
T. AQUA	PDLIHPRTGR	LHTRFNQ. TA	TATGRLSSTD	P. NLQNI. V	RTPLGQRIR	AF.....	I	AEEGNLLVAL	DYSQIELRVL
SPO2	MEKALCPDNR	VRGLLQFYGA	SRTGRVAGRL	V. QVONL. PON	KIEDLDTARN	LLKGGHYEAI	ELLYGQVFPV	LSQLVRTAFI	P	SEGNQFYVS	DSATIEARVI
T5	LLSIDA. DGC	IRTFGHE. HM	TTSGRLSSSG	KLANQQLP..	RDESIINGCV	V	APPGYRVIAW	DLTTAEVYYA
T7	WLRYVAEDGK	IHGSVNP. NG	AVTGRATHAF	P. NLAQIPGV	RSPYGEQCRA	AFGAEHHLDG	I	TGKFWVQAGI	DSGIELRCL

E. COLI	AHLSDKGLL	TAFAE.....	.GKDIHRAT	.AAEVFGLPLET	VTSEQRSAK	AINFGLIYGM	SAFGLARQLN	IPRK
S. PNEU	AHISKDEHLI	KAFQE.....	.GADIHTST	.AMRVFGIERPDN	VTANDRRNAK	AVNFGVVYGI	SDFGLSNNLG	ISRK
T. AQUA	AHLSDENLI	RVFOE.....	.GRDIHTET	.ASVWFGVPREA	VDFLBRRAK	TINFGVLYGM	SAERLSQELA	IPYE
SPO2	AWLAGEWRL	EVFN.....	.THGKIYEA	.SAAQMFKVPVESIT	KGSPLEKGR	VAELALGYQG	GRGALIQMGA	INMGLAEGEL	PE. LVKAMRT	
T5	AVLSGDRNMQ	QVF INMRNEP	DRYDPFHSNI	.AHMVFKLQ	.CEPRDVKK	LFPALRQAK	AITFGILYGS	GFARVAHSVN	EALLEQAART	GEFPVECTVA	
T7	AHFARFDNG	EYAHEIL...	.NGDIHTKN	QLAAEL....PTENAK	TFIYGFLYGA	GDEKIQGVG	AGKE

E. COLI	EAQKYMDLYF	ERYPGVLEYM	ERTRAQAKEQ	GY.....	.VETLDGR	LYLPDIKSSN	GAR.....RAAAERA	AIRAPMGGTA
S. PNEU	EAKAYIDTYF	ERFPGIKNYM	DEVVREARDK	GY.....	.VETLTKR	RELPDINSRN	FNI.....RGFAEAT	AIRNSPIGSA
T. AQUA	EAQAFIERYF	QSFPKVRAMI	EKILEEGRRR	GY.....	.VETLFGRR	RYVPDLARV	KSV.....REAAERM	AIRMPVGGTA
SPO2	ANKKIVKFWY	DVEAAAKAV	KERKPKVLQH	GLTFLYESGI	LFVQLPSGR	LAYAKPKLEL	DERFGREALT	YEGKLESQW	GRLNITYGGKL	VEN. IVQATA
T5	DAKEYIETFY	GQFPQLKRWI	DRCHDQIKNH	GF.....	IYSHFRGR	L. HNIHSED	RGV.....QGEIRS	GFRAIQSAS
T7	RGKELKCKFL	ENTPAIAAALR	ESIQQTLVES	SQ.....	WVA	GEQQVKKR	W. IKGLDGRK	VHV.....RSPHA

E. COLI	ADI IKRAMIA	VDANLQAEQP	RV.....	RM	IMQVEDELVF	EVHK...DOV	DAVAKQIHQL	MEN....CTR	LDVPLLVEVG	SGENWQDAH
S. PNEU	ADILKIAMIQ	LDKALVAGGY	QT.....	KM	LLQVEDEIVL	EVPK...SEL	VENMKLVKQT	MEE....AIQ	LSVPLIADEN	EGATWYEAQ
T. AQUA	ADIMKIAMVK	LFPRLPEMGA	RM	LLQVEDELVL	EAPK...ERA	EAVARLAKEV	MEG....VYP	LAVPLEVEVG	IGEDWLSAK
SPO2	RDCLAITLMR	LDNAGY....	KT	VMLVDEAVL	DVPRGR. NEL	DKVEAIMGEF	IS....WA	KGLPLTADGF	VTDYKRD.
T5	SDSLILGAVD	ADNEIISLGL	EQEM....	KI	VMLVDESVVA	IVREDLIDQY	NEILLIRNIQK	DRG....ISI	PGCP IGIDSD	SEAGGSRDY
T7	ALICKLWI IK	TEEMLVEKGL	KHGWGDGFAY	MAWVEDEIQV	GCRT...EEI	AQVVIETAQE	AMRWVGDHWN	FRCLLDTEGR	MGNWAIACH	

Fig. 2. Multiple sequence alignment of the C-terminal regions of *E. coli* DNA polymerase I, *S. pneumoniae*, *T. aquaticus*, bacteriophages SPO2, T5 and T7 polymerases. The major conserved motifs are numbered 1-5. The number of the first amino acid of the displayed sequence is: 554, 511, 458, 213, 352 and 310 for *E. coli*, *S. pneumoniae*, *T. aquaticus*, SPO2, T5 and T7 respectively.

alignment pattern taken for SPO2 and *E.coli* polymerases are shown in Figure 1; it is clear that strong regions of homology exist at the 4.5 SD or higher level. No such strong relationships could be found between SPO2 and *S.pneumoniae*, *T.aquaticus*, T7 or T5 polymerases. Figure 2 shows the alignment of the C-terminal part of these six polymerase sequences; this multiple

alignment was obtained by manual adjustment of the different but closest pairwise alignments. A conservation profile resulting from this alignment was also calculated (data not shown); this profile is based on a five-residue window and the score is simply the normalized sum of the matrix elements corresponding to the mutations observed in all the different pairwise alignments. The

----- Motif A -----			----- Motif B -----			---- Motif C ----									
	++ +	+ + +		+ + +		++++	+++	DNA-DEP DNA POL							
HUM-ALFA	858	FILLL	D	FNSLYPSIIQEPNICFTTV	66	IQKAL	K	LTA.NSM	YG	CLGFSYSRIFYAK	25	NLEVIYG	D	TDSIMIN	Pol Alpha' s
CAL-NPV	525	NAFSL	D	FNSLYLTIMIAICACLSNL	50	QKNSV	K	RTA.NSI	YG	YGFIFYKVLANY	34	TFKVVYD	D	TDSITFVL	
T4	403	YIMSF	D	LTSLYPSIIROVNISETTI	121	INQLNR	K	ILI.NSL	YG	ALGNIHFYRYDL	33	EDFIAAG	D	TDSVYVC	
EPST-B	579	FVLVV	D	FASLYPSIIQAENLCYSTM	73	KQLAI	K	CTC.NAV	YG	FTGVANGLFPC	47	QLRVIYG	D	TDSLFE	
HERPES	717	PVVVF	D	FASLYPSIIQAENLCFSTL	70	KQAAI	K	VVC.NSV	YG	FTGVQHGLLPCL	48	SMRIIYG	D	TDSIFVL	
CYTOMEG	712	PVAVF	D	FASLYPSIIQAENLCYSTL	71	KEQML	K	VTC.NAF	YG	FTGVVNGMPC	71	EARVIYG	D	TDSVFR	
VACCINIA	519	NVLIF	D	YNSLYPNVCIIFGNLSPETL	86	SMQTY	K	IVA.NSV	YG	LMGFRNSALYSY	62	RFRSVYG	D	TDSVFTE	
VARIC-Z	677	PVVVL	D	FASLYPSIIQAENLCFTTL	69	KQAAI	K	VVC.NSV	YG	FTGVVQGFPC	48	EVKVIYG	D	TDSVFR	
ADENO	538	PLYVY	D	ICGMYASAL.THPMPWGP	120	TLRSIA	K	LLS.NAL	YG	SFATKLDNKKIV	148	PLKSVYG	D	TDSLFTV	
PHI29	244	EGMVF	D	VNSLYPAQMYSRLLPYGEP	109	AIKOLA	K	IML.NSL	YG	KFASNPDTGKV	46	YDRIIYC	D	TDSIHLT	
PGKLI	644	DVLCI	D	VKELYPSASAFYDQYGSF	109	VKRNVI	K	IIM.NSL	WG	KFAKRWVNFYF	63	GAECIYS	D	TDSIFVH	
PRD1	215	DIRKY	D	VNSMYPHAMRNFRHPFSDE	94	FHNIFY	K	LLL.NSS	YG	KFAQMPENYKEW	61	AERPLYC	D	TDSIISR	
E.COLI	700	VIVSA	D	Y.SQIELRIMAHLSRDKGL	30	EQRRSA	K	AINFGLI	YG	MSAFGLARQLNI	101	RMINQVH	D	.ELVFEV	Pol I' s
S.PNEUM	648	VLLSS	D	Y.SQIELRVLVAHLSKDEHL	31	NDRRNA	K	AVNFGVY	YG	ISDFGLSNLGI	101	RMLQVH	D	.EIVLEV	
T.AQUAT	605	LLVAL	D	Y.SQIELRVLVAHLSGDENL	30	LMRRAA	K	TINFGVL	YG	MSAHRLSQELAI	99	RMLQVH	D	.ELVLEA	
SPO2	381	EYFVS	D	F.SATEARVIAFLAGEENR	32	PLRQK	K	VAELALG	YQ	GKRGALIQMGAL	131	KIVMHVH	D	.EAVLLV	
T5	496	RVLAW	D	L.TTAEVYYAAVLSGDRNM	42	ALRQAA	K	AITFGIL	YG	SGPAKVAHSVNE	103	KIVMLVH	D	.SVVALV	
T7	470	VOAGI	D	A.SGLELRCLAHEFMAREFDN	24	PTRONA	K	TFIYGFL	YG	AGDEKIQGIVGA	104	AYMARVH	D	.EIQVGC	
RAT-BETA	12	NGGIT	D	MLVELANFEKNVSOAIHKY	147					RGAESSG	D	MDVLLTH	D		Pol Beta' s
MOUSE-TDT	165	NQLFT	D	ALLDILAENDELRENGSCL	146					RGRMIGH	D	VDFLITI	D		
T3	533	LPLAF	D	G.SCSGIIQHFSSAMIRDEVG	69	VTRSVT	K	RSVMILA	YG	SKEFGFRQOVID	152	ESFALH	D	.SFGTIP	DNA-DEP RNA POL
MITO	940	QPVHQ	D	G.TCNGLQHYAALGGDVEG	49	ITRFVV	K	QTVMTNV	YG	VTVVGGATFQIAK	147	LDFAVSH	D	.SYWTHA	
DIRS-1	154	YMKVL	D	IKKAYLHVLVDPQYRDLFR	40					VSVIAYL	D	.DLLIVG	D		Gypsy-like
GYPSY	252	FFITL	D	LKSGYHQIYLAEHDREKTS	38					KICIVYV	D	.DVIIFS	D		
I-FAC	438	SLVTL	D	FSRAFDRVGVHSIIQQLQE	66					IKFNAYA	D	.DFFLLI	D		Line-like
RTCHLA	180	VLVTF	D	LQAAYNSVDIKHLMQTLQL	52					MDFTIYA	D	.NFAQVV	D		
TY912	914	YITQL	D	ISSAALYADIKKELYIRPP	55					VTICLFV	D	.DMVLF	D		TY-like
HIV1	260	SVTVL	D	VGDAFYFVPLDEDFLDKVF	49					IVIYQYM	D	.DLYVGS	D		
W-HEPB	460	QMLSL	D	VSAAFYHIPISPAAVPHLL	94					CVVFAYM	D	.DLVLGA	D		DNA virus
POLIO	1976	KLFAF	D	Y.TGYDASLSPAMFEALRM	70					LRMIAYG	D	.DVIASY	D		Plus-Strand
COXV	1951	HLLAF	D	Y.SGYDASLSPVNFACLKM	71					FRMIAYG	D	.DVIASY	D		
BTV	633	LEQII	D	F.GYGEGRVANTLWNGKRR	99					LSEQYVG	D	.DTLFTY	D		Double-Strand
INFL	300	FTITG	D	N.TKRNENQNPFRFLAMIT	113					WDGLQSS	D	.DFALIV	D		
SENDAI	658	CFLTT	D	L.KKYCLNWRPFESTALFGQ	83					VSAMVQG	D	.NQALAV	D		Minus-Strand
NK	BETA 9	HELIX L			HELIX O					BETA 12	BETA 13				
SK	bbbb b	tttttaaaaaaatttttt		aaaaa a	aaaaaa aa	ttttttttttt		ttttttttttt		tbbbbt t	ttbbbb				
S1	bbbb b	bttttttaaaaaa		aaaaa a	at.tttt	tt	ttttttttttt		ttbbbbt t	ttbbbb					
S2	bbbb b	t.ttaaaaaa		tttaa a	bbbbbt	tt	ttttttttttt		bbbbbt t	ttbbbb					
S3	bbbb a	aaaaaaaaa		bbbb b	bbbbbt	tt	ttttttttttt		ttttt t	ttbbbb					
S4	tttt t	t.ttttaaaaaattttt		bbbb b	bbbbbt	tt	ttttttttttt		aaaabb b	ttttt					
S5	bbbb b	t.ttttttttaaaaaa							bbbbbt t	ttbbbb					
K#		705			758	767				882					

Fig. 3. Multiple alignments of various polymerase sequences for motifs A, B and C. The putative corresponding regions in the Klenow fragment of *E.coli* pol I structure are shown in Figure 4 as blackened for motif A, vertically striped for motif B and horizontally striped for motif C. Strictly conserved positions are boxed, highly conserved residues are bold and underlined, generally hydrophobic residues are indicated by a (+) (top line). References for DNA-dependent DNA polymerases of pol α type can be found in Wang *et al.* (1989). The sequences are taken from human polymerase alpha, Autographica californica nuclear polyhedrosis virus, phage T4, Epstein-Barr virus, cytomegalovirus, vaccinia virus, varicella-Zoster virus, cytomegalovirus, adenovirus, phage 29, yeast plasmid PGKLI and phage PRD1. DNA polymerases of the pol I type come from *E.coli*, *S.pneumoniae*, *T.aquaticus* and bacteriophages SPO2, T5 and T7. Pol β type is represented by the rat DNA polymerase β and mouse terminal transferase. Also included are DNA-dependent RNA polymerases from phage T3 and yeast mitochondria (Masters *et al.*, 1987). The ones from phages T5 and Sp6 were omitted since they are too close to the one from phage T3. The final set of RNA-dependent polymerases derive from (references in Poch *et al.*, 1989) *D.discoideum* transposon DIRS-1; *D.melanogaster* gypsy polyprotein; *D.melanogaster* I factor 2 transposon (Ifac) *C.rheinardii* intron (mitochondrial DNA) Rchla; *S.cerevisiae* transposon Ty 912; human immunodeficiency virus pol polyprotein; Woodchuck hepatitis B virus polymerase; poliovirus (strain Mahoney) genome polyprotein; Coxsackievirus B3 polyprotein; blue tongue polymerase, P1 protein from influenza virus A; Sendai virus (strain Z) and L protein. The three motifs are identified as motif A (N-terminal), motif B and motif C (C-terminal). The number of amino acids between motifs for each sequence is also indicated. The line NK shows the secondary structure as observed in the tertiary structure of the *E.coli* Klenow fragment [see Figure 4 for fold and nomenclature and Ollis *et al.* (1985a) for the structure description]. Under SK, the observed Klenow secondary structure is given for each motif position where a = helix, b = strand and t = turn; residues predicted as turn or not predicted as helix or strand are designated turn. S1, S2, S3, S4 and S5 show the mean predicted secondary structure (Argos, 1985) for pol α, pol I, pol β DNA-dependent DNA polymerases, DNA-dependent RNA polymerases and RNA-dependent polymerases respectively. Under K# the amino acid sequence number according to the Klenow fragment tertiary structure is given for the highly conserved sites; this allows better recognition of the site in the Klenow fold given in Figure 4.

matrix is a rescaled Dayhoff matrix as used by Gribskov and Burgess (1986). A threshold of 60% conservation leaves five regions of high homology, which are indicated in Figure 2. These regions are characterized by the conserved sequences (F)N••S••(Q)(L)•••L for the first region [\bullet refers to a given alignment position occupied by any amino acid, (X) indicates an amino acid almost universally conserved by the six sequences, and a single letter refers to exact conservation in all the sequences], (T)GR for region 2, D•(S)••E for region 3, K•••••Y(G) for region 4 and VHD(E) for region 5.

A multiple alignment was also built for 12 sequences of DNA-dependent DNA polymerases α s from man, yeast and several viruses. These sequences are essentially the ones presented by Wang *et al.* (1989) together with two recent additions to the NBRF sequence data bank (one polymerase from another adenovirus, Wmad12, and one from a nuclear polyhedrosis virus, Npadnapma). The resulting conservation profile using a window of five residues left three regions above a 60% conservation threshold. They are characterized by D••(S)(L)Y(P)(S), K•••N(S)•(Y)G and (Y)(G)DTDS, as previously noted by other authors (Bernad *et al.*, 1987). A comparison of the most conserved sequences for the pol I and pol α types suggests an overlap of these three regions of pol α s with regions 3, 4 and 5 of the pol I type; these regions will hereafter be called motifs A, B and C (Figure 3). These three motifs are centered on invariant residues: motif A contains a strictly conserved aspartate at the junction of a beta strand and an alpha helix, motif B contains an alpha helix with positive charges and motif C has a doublet of negative charges located in a beta-turn-beta secondary structure, as implied from the Klenow structure (Figure 4). Great variability in the lengths separating the different conserved regions was observed for the pol α s: 50–120 residues between motifs A and B (average 70) and 25–150 residues between motifs B and C (average 60). In the pol Is, these values compare respectively with 30 and 100 residues, on average.

A rat DNA-dependent DNA polymerase β (Matsukage *et al.*, 1987) can be easily aligned with a mouse terminal deoxynucleotidyltransferase (Koiwai *et al.*, 1986). Though many residues are identically conserved amongst the two sequences, an N-terminal and a C-terminal fragment may be related to motifs A and C of DNA polymerases (Figure 3).

For DNA-dependent RNA polymerases, the three motifs could only be found in the proteins made of only one subunit. Four different sequences of this type are known, the two most distantly related being from phage T3 and yeast mitochondria (Masters *et al.*, 1987). The distances between motifs compares well with the ones of the pol I type. The sequence fragments corresponding to the motifs are presented in Figure 3, together with those found in DNA pol α s, pol Is and pol β s.

Secondary structure predictions averaged over each set of sequences (DNA pol. α s, pol Is and pol β s and DNA-dependent RNA pols) were made according to the procedures of Argos (1985) and are given in Figure 3. Except for pol β s, for which only two sequences are known, they often show agreement amongst themselves as well as with the pol I structure, especially for motifs A and C.

Recently, Poch *et al.* (1989) identified four motifs shared by all RNA-dependent RNA, and DNA polymerases (84 different sequences). Two of their motifs were found to be similar to two of those contained in the DNA-dependent polymerases; namely, motifs A and C. These two motifs maintain the strictly conserved residues in the same environment. The mean secondary structure predictions agree with the observed structure in the corresponding

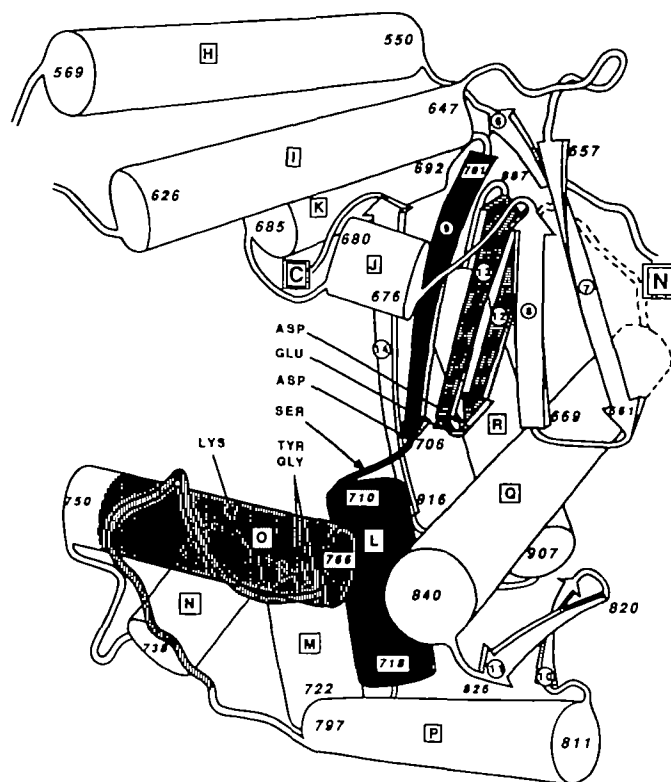


Fig. 4. An illustration of the tertiary folding pattern of the polymerase domain from the *E. coli* Klenow fragment. The sketch is similar to that given by Ollis *et al.* (1985a). The three differently shaded regions, in moving from the N-terminus to the C-terminus, correspond respectively to motifs A (black region), B (vertical stripes) and C (horizontal stripes), whose sequences are given in Figure 3. Sequence positions relative to the secondary structures are numbered. β -strands are numbered sequentially and α -helices are designated by letters according to the alphabet. The structural position of the conserved amino acids is indicated in each of the motifs.

regions of the Klenow fragment (Poch *et al.*, 1989). Sequence fragments from 12 representative RNA-dependent polymerases (i.e. from each of the subfamilies: reverse transcriptases, plus-, minus- and double-strand RNA polymerases) are shown aligned to those matched in the DNA pols (Figure 3).

Several viral RNA-dependent RNA polymerase and reverse transcriptase sequences were then compared to those from pol Is and pol α s using the same sequence comparison technique as above (Argos, 1987). Among the different pairwise comparisons, extensive sequence similarities were found between the viral hepatitis B RNA-dependent DNA polymerase from Woodchuck (Galibert *et al.*, 1982) and Herpes simplex virus DNA polymerase, a member of the DNA pol α family (Gibbs *et al.*, 1985). The matrix is shown in Figure 5 and the resultant alignment in Figure 6. It is noteworthy that 31% of the aligned residues are identical. In this alignment, two conserved regions can be delineated that correspond to motifs A and C. The homology between these two sequences provides a possible link between the DNA- and RNA-dependent polymerases.

Discussion

DNA-dependent DNA polymerases

Sequence similarities. In this work, a new member was added to the pol I type of DNA-dependent DNA polymerases. An analysis of the multiple alignments for the pol I type as well as for the pol α type pointed to three conserved motifs held in

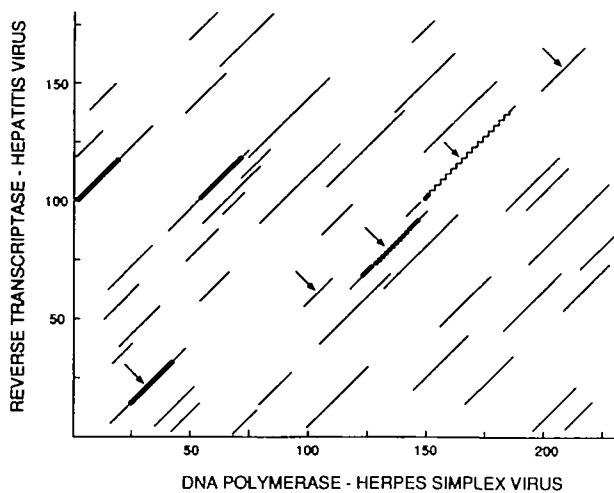


Fig. 5. Homology search (Argos, 1987) between Woodchuck hepatitis B RNA-dependent DNA polymerase and Herpes simplex DNA-dependent DNA polymerase. The search window lengths ranged from 9 to 35 in steps of 2; the search peaks are plotted over the entire window length, with the largest value dominating when overlap occurs from the multiprobes. The peak values (S) are scored as a number of standard deviations (σ) above the matrix mean for each window length. Thin lines indicate $3.2\sigma < S < 3.9\sigma$; thick lines, $4.0\sigma < S < 4.3\sigma$; circles $4.4\sigma < S < 4.7\sigma$; and jagged lines, $4.8\sigma < S < 5.0\sigma$. The path used as a basis to effect the alignment shown in Figure 6 is indicated by arrows.

common. The distances between these motifs are quite variable. However, within each polymerase type, there is already great variation in the distances between conserved regions. Moreover, it would seem unlikely that three of five most conserved sequence regions in the polymerase domain of pol Is (Figure 2) could be aligned with the three most conserved regions of DNA pol α s purely by chance. A different piece of evidence pointing to a possible link between the pol I and the pol α s comes from sequence similarities within the exonuclease domain of the T5 polymerase and pol α s (Leavitt and Ito, 1989). It is then reasonable to expect their polymerase domains to be related also. In fact, this observation has been recently extended to all exonuclease domains of pol α s and pol Is, including SPO2 (Bernad *et al.*, 1989). In addition, the secondary structure predictions, especially for motifs A and C, point to supersecondary structures for pol α s that coincide with those observed in the pol I Klenow structure.

From the sequences to structure. The three motifs A, B and C are clustered and consecutive in the C-terminal fraction of the Klenow fragment and correspond to structural features likely to interact with the DNA (Ollis *et al.*, 1985b). Figure 4 highlights the portions of the *E. coli* pol I tertiary structure corresponding to the three shared motifs. Motif A, characterized by a conserved Asp and Ser, corresponds to strand 9 and helix L. Motif B, with conserved Lys, Tyr and Gly encompasses most of helix O and the following long loop region. Motif C delineates strands 12 and 13 with the mostly conserved negative charges contained in the connecting loop of the β -hairpin. Since the N-terminal Asp of the doublet is the only one universally conserved, it is presumed to be catalytically more important. It is clear that the structural segments A and B are likely to come into contact with DNA and that the strand 12–13 hairpin loop could place at least one catalytic residue in the polymerase active site. Model building of DNA into the Klenow fragment structure also supports the importance of these regions (Ollis *et al.*, 1985a,b; Warwicker *et al.*, 1985). Furthermore, the C_{α} atom of motif C Asp882 is

within 5.5 Å of that from Asp705, conserved in motif A [C_{α} atom coordinates given in Brookhaven database (Bernstein *et al.*, 1977, file 1DPI)]. Their spatial proximity would allow both to participate in catalysis. Region 2 of pol Is (sequence TGR between strands 7 and 8—see Figures 2 and 4) is also located in the vicinity of this area but no sequence homology with the pol α type could be found for this region. Region 1 falls in an undefined part of the electron density map (see Ollis *et al.*, 1985a; Figure 4).

From structure to function. Considerable biochemical evidence points to the importance of these three motifs in the DNA polymerase activity. A synthesized *E. coli* pol I oligopeptide corresponding to the N-terminal-most two-thirds of the loop region connecting helices O and P (motif B—see Figure 4) has been shown to bind deoxynucleotide triphosphate substrates of pol I as well as duplex DNA (Mildvan, 1989). Furthermore, photo-affinity labeling with 8-azido-dATP identifies Tyr766 as a residue in the active site (Joyce and Steitz, 1987) while Lys758, also part of this motif, succumbs to chemical labeling, this time using pyridoxal phosphate (Basu and Modak, 1987). A further synthesized peptide (Shenbagamurthi *et al.*, 1988; Mildvan, 1989) corresponding to helix Q and strands 12 and 13 (motif C) was not found to bind the pol I substrate, although it apparently retains its proper folding; however, this result does not exclude that this peptide, although unable to bind dNTP on its own, can co-operate with other regions of the native protein and be just part of the binding site. In fact, His881 of pol I, which is in the loop joining strands 12 and 13 and sequentially adjacent to the conserved Asp882 of motif C, has been shown to be involved in the binding of [32 P]dTTP (Pandey *et al.*, 1987). In addition, the beginning of helix Q, which is close in space to the end of helix O (motif B) and motifs A and C, can also be labeled by phenylglyoxal (Mohan *et al.*, 1988).

Similarly, for the pol α type, genetic studies of the Herpes simplex virus polymerase (Larder *et al.*, 1987a) revealed that four out of the six identified drug-resistant polymerase mutants cluster in motifs A and B. The drug used was a dNTP analog. Other mutants involved in drug and substrate recognition were also mapped by Gibbs *et al.* (1988) in motifs A and B. This is in agreement with a central role of these motifs in dNTP binding for the pol α type, as is the case of the pol I type.

Finally, for the third DNA polymerase type (pol β), a chemical affinity-labeled [32 P]8-azido-dATP peptide in the terminal transferase allowed the mapping of the dNTP binding site; this peptide contained the sequence GHDVD that is part of motif C (Evans *et al.*, 1989) and that resembles both YGDTD and VHDE. This supports the alignment given in Figure 3, even though motif B could not be found with certainty in this pol β type of DNA polymerases. Furthermore, the number of residues between motifs A and C in the transferase (~ 150) matches well the corresponding number in the Klenow pol Is averaging about 170 residues.

Structural implications of the distance variability between motifs. Figure 3 shows the number of residues contained between each of the motifs for all the DNA pol α sequences and for the Klenow polymerase I domains. Between motifs A and B a comparable number of residues is found for both pol I and α types, while between B and C the pol α sequences generally contain considerably fewer amino acids than their putative Klenow counterparts. This region of the structure encompasses helix P, strands 10 and 11, and helix Q. Apart from the beginning of helix Q, these regions are unlikely to be in contact with the DNA, according to model building studies (Warwicker *et al.*, 1985; Figure 4), suggesting that this region could be considerably

future experiments. If reported crystals of a heterodimer of the reverse transcriptase for human immunodeficiency virus prove to be diffraction worthy (Lowe *et al.*, 1988) and if the high resolution structure of T7 RNA polymerase soon becomes available, then the final test for the hypothesis made here should be forthcoming.

References

- Argos, P. (1985) *EMBO J.*, **4**, 1351–1355.
 Argos, P. (1987) *J. Mol. Biol.*, **197**, 331–348.
 Argos, P. (1988) *Nucl. Acids Res.*, **16**, 9909–9916.
 Argos, P., Tucker, A.D. and Philipson, L. (1986) *Virology*, **149**, 208–216.
 Basu, A. and Modak, M.J. (1987) *Biochemistry*, **26**, 1704–1709.
 Bernad, A., Zaballo, A., Salas, M. and Blanco, L. (1987) *EMBO J.*, **6**, 4219–4225.
 Bernad, A., Blanco, L., Lazaro, J.M., Martin, G. and Salas, M. (1989) *Cell*, **59**, 219–228.
 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Bruce, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.
 Evans, R.K., Beach, C.M. and Coleman, M.S. (1989) *Biochemistry*, **28**, 713–720.
 Freemont, P.S., Ollis, D.L., Steitz, T.A. and Joyce, C.M. (1986) *Proteins*, **1**, 66–73.
 Galibert, F., Chen, T.N. and Mandart, E. (1982) *J. Virol.*, **41**, 51–65.
 Gibbs, J.S., Chiou, H.C., Hall, J.D., Mount, D.W., Retondo, M.J., Weller, S.K. and Coen, D.M. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 7969–7973.
 Gibbs, J., Chiou, H., Bastow, K., Cheng, Y.C. and Coen, D.M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 6672–6676.
 Gribskov, M. and Burgess, R.R. (1986) *Nucl. Acids Res.*, **14**, 6745–6763.
 Inokuchi, Y. and Hirashima, A. (1987) *J. Virol.*, **61**, 3946–3949.
 Joyce, C.M. and Steitz, T.A. (1987) *Trends Biochem. Sci.*, **12**, 277–292.
 Jung, G., Leavitt, M.C., Hsieh, J.C. and Ito, J. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 8287–8291.
 Kamer, G. and Argos, P. (1984) *Nucl. Acids Res.*, **12**, 7269–7282.
 Koiwai, O., Yokota, T., Kageyama, T., Hirose, T., Yoshida, S. and Arai, K. (1986) *Nucl. Acids Res.*, **14**, 5777–5792.
 Larder, B.A., Kemp, S.D. and Darby, G. (1987a) *EMBO J.*, **6**, 169–175.
 Larder, B.A., Purifoy, D.J.M., Powell, K.L. and Darby, G. (1987b) *Nature*, **327**, 716–717.
 Lawyer, F.C., Stoffel, S., Saiki, R.K., Myambo, K., Drummond, R. and Gelfand, D.H. (1989) *J. Biol. Chem.*, **264**, 6427–6437.
 Lazcano, A., Fastag, J., Gariglio, P., Ramirez, C. and Oro, J. (1988) *J. Mol. Evol.*, **27**, 365–376.
 Leavitt, M.C. and Ito, J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 4465–4469.
 Lopez, P., Martinez, S., Diaz, A., Espinosa, M. and Lacks, S.A. (1989) *J. Biol. Chem.*, **264**, 4255–4263.
 Lowe, D.M., Aitken, A., Bradley, C., Darby, G.K., Larder, B.A., Powell, K.L., Purifoy, J.M., Tisdale, M. and Stammers, D.K. (1988) *Biochemistry*, **27**, 8884–8889.
 Matsukage, A., Nishikawa, K., Ooi, T., Seto, Y. and Yamaguchi, M. (1987) *J. Biol. Chem.*, **262**, 8960–8962.
 Masters, B.S., Stohk, L.L. and Clayton, D.A. (1987) *Cell*, **51**, 89–99.
 Mildvan, S. (1989) *FASEB J.*, **3**, 1705–1714.
 Mohan, P., Basu, A., Abraham, K.I. and Modak, M.J. (1988) *Biochemistry*, **27**, 226–233.
 Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G. and Steitz, T.A. (1985a) *Nature*, **313**, 762–766.
 Ollis, D.L., Kline, C. and Steitz, T.A. (1985b) *Nature*, **313**, 762–766.
 Pandey, V.N., Williams, K.R., Stone, K.L. and Modak, M.J. (1987) *Biochemistry*, **26**, 7744–7748.
 Peterson, R.C., Chung, L.C., Mattaliano, R.J., White, S.T., Chang, L.M.S. and Bollum, F.J. (1985) *J. Biol. Chem.*, **260**, 10495–10502.
 Poch, O., Sauvaget, I., Delarue, M. and Tordo, N. (1989) *EMBO J.*, **8**, 3867–3874.
 Raden, B. and Rutberg, L. (1984) *J. Virol.*, **52**, 9–15.
 Shenbagamurthi, P., Muller, G.P. and Mildvan, A.S. (1988) *FASEB J.*, **2**, A588.
 Soos, R., Rose, J.P., Chung, Y.J., Lafer, E.M. and Wang, B.C. (1989) *Proteins*, **5**, 266–270 and *Proceedings of the Annual Meeting of the American Crystallographic Association in Seattle, July 23–29, 1989*, p.110 (PB27).
 Wang, T.S.F., Wong, S.W. and Korn, D. (1989) *FASEB J.*, **3**, 14–21.
 Warwicker, J., Ollis, D., Richards, F.M. and Steitz, T.A. (1985) *J. Mol. Biol.*, **186**, 645–649.
 Wilson, S.H. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 5106–5110.
 Zmudzka, B.Z., SenGupta, D., Matsukage, A., Cabiantti, F., Kumar, P. and Wilson, S.H. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 5106–5110.

Received on November 24, 1989; accepted on February 24, 1990