Volume 16 Number 21 1988 Nucleic Acids Research

A sequence motif in many polymerases

Patrick Argos

European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, FRG

Received August 4, 1988; Accepted September 20, 1988

ABSTRACT

A 15-residue sequence motif has been found in many polymerases from various species and involving DNA and RNA dependence and product. The motif is characterized by a Tyr-Gly-Asp-(Thr)-Asp core flanked by hydrophobic spans five residues in length. An mRNA maturase segment is also suggested to display the motif pattern. The aspartates may be important in polymerase function by acting directly in catalysis and/or by binding magnesium.

INTRODUCTION

In 1984 Kamer and Argos (1) discovered a 14-residue sequence motif found in 15 viral reverse transcriptases and RNA-directed RNA polymerases. Central to the pattern was a four-residue core where tyrosine and glycine formed the consensus at positions 1 and 2 while aspartic acids invariably occupied positions 3 and 4. On either side of the central region were five residues which were often hydrophobic. It was suggested that this span represented an active processing region. Since 1984 many further viral sequences (a present total of 41) have displayed the pattern in their respective RNA-directed DNA or RNA polymerases (Figure 1). Recently, Wong et al. (2) have reported the primary sequence of the human DNA-directed DNA polymerase alpha catalytic polypeptide, generally agreed to be the principal polymerase in eukaryotic replication. They also report homologies in seven sequence spans (roughly 25 residues in mean length) with DNA-directed DNA polymerases from yeast, various human viruses, and bacteriophages (Figure 1). Their region I span, which is the most strongly conserved of the seven, contains a central segment composed of the almost universally invariant residues Tyr-Gly-Asp-Thr-Asp. It is suggested here that this sequence region bears strong resemblance to the RNA-directed polymerase cores discussed previously, even including the two five-residue hydrophobic flanking spans (Figure 1).

Sequence															DIR	PPR	REF	Species
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15				
N	L	E	V	I	Y	G	D	T	D	S	I	M	I	N	DNA	DNA	2	human alpha gene
N	L	L	V	V	Y	G	D	T	D	S	V	M	I	D	DNA	DNA	2	S.cerevisiae pol I
S	M	R	I	I	Y	G	D	T	D	S	I	F	V	L	DNA	DNA	2	herpes simplex virus
E	A	R	V	I	Y	G	D	T	D	S	V	F	V	R	DNA	DNA	2	cytomegalovirus
Q	L	R	V	I	Y	G	D	T	D	S	L	F	I	E	DNA	DNA	2	Epstein-Barr virus
R	F	R	S	V	Y	G	D	T	D	S	V	F	T	E	DNA	DNA	2	vaccinia virus
A	E	R	P	L	Y	C	D	T	D	S	I	I	C	R	DNA	DNA	30	bacteriophage PRD1
E	D	F	I	A	A	G	D	T	D	S	V	Y	V	C	DNA	DNA	2	bacteriophage T4
Y	D	R	I	I	Y	C	D	T	D	S	I	H	L	T	DNA	DNA	2	bacteriophage phi-29
P	L	K	S	V	Y	G	D	T	D	S	L	F	V	T	DNA	DNA	2	adenovirus 2
E	V	K	V	I	Y	G	D	T	D	S	V	F	I	R	DNA	DNA	31	varicella-zoster virus
M	Y	I	I	H	Y	M	D	-	D	I	L	I	A	G	RNA	DNA	7	simian retrovirus(GNLJMP)
L	I	V	I	H	Y	M	D	-	D	I	L	I	C	H	RNA	DNA	7	hamster A-particle(GNHYIH)
C	T	I	L	Q	Y	M	D	-	D	I	L	L	A	S	RNA	DNA	1	virus HTLV-I
S	T	I	V	Q	Y	M	D	-	D	I	L	L	A	S	RNA	DNA	7	virus HTLV-II(GNLJH2)
V	I	I	I	Q	Y	M	D	-	D	I	L	I	A	S	RNA	DNA	7	AIDS virus HIV-II ROD(GNLJG2)
I	V	I	Y	Q	Y	M	D	-	D	L	Y	V	G	S	RNA	DNA	3	AIDS virus HTLV-III
C	L	A	F	S	Y	M	D	-	D	V	V	L	G	A	RNA	DNA	1	human hepatitis B virus
C	L	A	F	A	Y	M	D	-	D	L	V	L	G	A	RNA	DNA	7	squirrel hepatitis B virus(JDVLS)
C	V	V	F	A	Y	M	D	-	D	L	V	L	G	A	RNA	DNA	7	woodchuck hepatitis B virus(JDVLV)
V	W	T	F	T	Y	M	D	-	D	F	L	L	C	H	RNA	DNA	7	Duck hepatitis B virus(JDVLC)
L	C	M	L	H	Y	M	D	-	D	L	L	L	A	A	RNA	DNA	1	Rous sarcoma virus
S	L	L	V	S	Y	M	D	-	D	I	L	I	A	S	RNA	DNA	3	bovine leukemia virus
L	I	L	L	Q	Y	V	D	-	D	L	L	L	A	A	RNA	DNA	1	maloney murine leukemia virus
I	Q	F	G	I	Y	M	D	-	D	I	Y	I	G	S	RNA	DNA	7	visna lentivirus(GNLJVS)
V	Q	L	Y	Q	Y	M	D	-	D	L	F	V	G	S	RNA	DNA	7	equine infectious anemia virus(GNLJEV)
K	F	C	C	V	Y	V	D	-	D	I	L	V	F	S	RNA	DNA	1	cauliflower mosaic virus
I	Y	V	L	L	Y	V	D	-	D	V	V	I	A	T	RNA	DNA	7	D.melanogaster copia transposon(OFFFCP)
K	H	C	L	V	Y	L	D	-	D	I	I	V	F	S	RNA	DNA	7	D.melanogaster 17.6 transposon(GNFF17)
V	T	I	C	L	F	V	D	-	D	M	V	L	F	S	RNA	DNA	7	S.cerevisiae Ty912 transposon(B22671)
V	S	V	I	A	Y	L	D	-	D	L	L	I	V	G	RNA	DNA	7	Dictyostelium DIRS-2 transposon(C24785)
F	R	M	I	A	Y	G	D	-	D	V	I	A	S	Y	RNA	RNA	24	coxsackievirus B3
D	R	L	L	F	S	G	D	-	D	S	L	A	F	S	RNA	RNA	24	cucumber mosaic virus
C	K	F	F	A	N	G	D	-	D	L	I	I	A	I	RNA	RNA	24	tobacco vien mottling virus
S	R	L	I	N	N	G	D	-	D	C	V	L	I	C	RNA	RNA	24	carnation mottle virus
L	I	G	P	K	C	G	D	-	D	G	L	S	R	A	RNA	RNA	24	black beetle virus
V	M	V	T	Y	G	G	D	-	D	S	L	I	A	F	RNA	RNA	24	tobacco ringspot virus
I	V	Y	Y	V	N	G	D	-	D	L	L	I	A	I	RNA	RNA	24	tobacco etch virus
G	S	L	G	I	Y	G	D	-	D	I	I	V	P	V	RNA	RNA	7	bacteriophage GA beta chain(RRBPBG)
G	T	I	G	I	Y	G	D	-	D	I	I	C	P	S	RNA	RNA	1	bacteriophage MS2 beta chain
L	R	I	L	C	Y	G	D	-	D	V	L	I	V	F	RNA	RNA	7	hepatitis A virus(GNNYHR)
N	A	S	C	A	A	M	D	-	D	F	Q	L	I	P	RNA	RNA	1	influenza P2 polypeptide
L	K	M	I	A	Y	G	D	-	D	V	I	A	S	Y	RNA	RNA	1	polio virus
I	G	L	V	T	Y	G	D	-	D	N	L	I	S	V	RNA	RNA	1	cowpea mosaic virus
L	K	I	I	A	Y	G	D	-	D	V	I	F	S	Y	RNA	RNA	7	Rhinovirus 2(GNNYH2)
L	K	I	L	A	Y	G	D	-	D	L	I	V	S	Y	RNA	RNA	24	Rhinovirus 14
V	K	V	L	S	Y	G	D	-	D	D	L	L	V	A	RNA	RNA	1	encephalomyocarditis virus
Y	T	M	I	S	Y	G	D	-	D	I	V	V	A	S	RNA	RNA	1	foot-and-mouth disease virus
K	C	A	A	F	I	G	D	-	D	N	I	V	H	G	RNA	RNA	7	middleburg virus(MNWVM)
R	C	A	A	F	I	G	D	-	D	N	I	I	H	G	RNA	RNA	1	sindbis virus
D	C	A	I	F	S	G	D	-	D	S	L	I	I	S	RNA	RNA	1	brome mosaic virus
N	F	V	V	A	S	G	D	-	D	S	L	I	G	T	RNA	RNA	1	alfalfa mosaic virus
S	R	M	A	V	S	G	D	-	D	C	V	V	K	P	RNA	RNA	7	West Nile virus(GNWVWV)
K	R	M	A	V	S	G	D	-	D	C	V	V	R	P	RNA	RNA	7	yellow fever virus(GNWVY)
I	K	G	A	F	C	G	D	-	D	S	L	L	Y	F	RNA	RNA	1	tobacco mosaic virus
A	A	Q	V	Y	A	G	D	-	D	M	S	I	D	Y	RNA	RNA	26	white clover mosaic virus
P	W	C	I	A	M	G	D	-	D	S	V	E	G	F	RNA	RNA	25	southern bean mosaic virus
-	-	H y	H y	-	Y	G M	D	-	D	-	H y	H y	-	-	CONSENSUS(Hy=hydrophobic)			

Figure 1. Aligned polymerase sequences in the Asp-Asp region. DIR indicates "directed by" while PPR refers to the "polymerized product." REF indicates "reference." In most cases the references given cite the publications where the sequences were first reported. In the case of sequences taken from the Protein Identification Resource (7), the code names of files which contain references for the sequences are given in parentheses. The motif position numbers referred to in the text are the first listed entry in the "Sequence" heading.

The major difference involves an inserted Thr between the two Asp's. The conserved aspartates, on an exposed loop in a predicted beta hairpin structure, may bind a magnesium cation as well as act catalytically in the polymerization process.

Johnson et al. (3) have suggested that the alpha-subunit of E. coli DNA-directed RNA polymerase II contains such an Asp-Asp motif. They support their contention by a possible homology over 113 residues between contiguous parts of the alpha chain and mouse Maloney leukemia virus reverse transcriptase. However, the match is weak and questionable (4) with only about 15% residue identity. Pro-Val flanks N-terminally the Asp-Asp pair; Pro is not found in the motifs of Figure 1. Furthermore, recent sequences of the mitochondrial polymerase II alpha subunits from tobacco (5) and liverwort (6) chloroplasts do not conserve the Asp-Asp pair which has been altered to Asp-Gln. As a result, the E. coli alpha subunit span is not included in Figure 1.

DATABASE SEARCH

The uniqueness of the Asp-Asp motif in polymerases was tested by searching the entire protein sequence data base (Protein Identification Resource (7) (PIR), release no. 15, consisting of nearly 6800 primary structures) with pattern rules developed from the polymerase segments.

(1) Positions 8 to 10 must be occupied by Asp-Thr-Asp or Asp-Asp.

(2) Position 7 can be Gly, Met, Cys, Val, or Leu.

(3) Position 6 can be Tyr, Ala, Phe, Ser, Asn, Cys, Gly, Ile, or Met.

(4) At least two of the residues in positions 1 to 5 and in positions 11 to 15 must be hydrophobic (Ala, Val, Leu, Ile, Cys, Met, Phe, Tyr, His, Trp, Pro).

(5) If only two residues in positions 1 to 5 are hydrophobic, then there must also be a Ser or Gly. If only two residues in positions 11 to 15 are hydrophobic, then there must also be a Ser.

(6) Position 4 cannot contain Lys, Arg, Asp, Glu, Gln, and Asn.

(7) Positions 12 and 13 must be occupied by hydrophobic residues.

It is clear from Figure 1 that positions 6 through 10 are constrained in residue selection. The composition of the five-residue flanking regions is 60% in Ala, Ile, Val, Leu, Met, and Cys (31% expected for proteins in general (8)); 10 % in Phe and Tyr (7% expected); and 23% in Asp, Glu, Gln, Lys, Arg, Thr, and Ser (45% expected). These results emphasize the hydrophobic character of the flanks.

Rules (1) to (6) are obeyed by all the spans in Figure 1; rule

(7) is violated by only four of the 55 sequences. Searching the entire databank with all seven constraints yields an error rate of one in 600 protein sequences; non-polymerase proteins containing compatible Asp-Asp regions are exemplified by the beta chain of Azotobacter vinelandii nitrogenase, mouse dihydrofolate reductase, and E. coli L-arabinose binding protein. A relaxation of the constraints (rules (1) to (5)) resulted in an error rate of one in 150 sequences. Use of the core conservation (rules (1) to (3)) and two hydrophobic residues in five on either side (rule (4)) gave a one in 50 error.

DISCUSSION

All the proteins passing through the filter of rules (1) to (4) were scrutinized for possible polymerase function. Four transposon segments which obeyed all seven constraints are listed in Figure 1. An examination of the references reporting the sequences showed that the authors themselves noted these segments as potentially important for transposon reverse transcriptase activity. Another suggested motif was from yeast mitochondrial cytochrome b mRNA maturase, responsible for intron splicing and maturation of the mRNA product. Since this function may involve joining RNA segments, the maturase could catalyze a reaction similar to polymerization. The Asp-Asp sequence span is shown in Figure 2. The identified region was noted by Lazowska et al. (9) to be homologous to an omega(+) open reading frame in the large ribosomal gene. Both Asp's remained invariant. Only this region was found conserved in the two sequences,

	Sequence														DIR	PPR	Name, Species, PIR code
	1	2	3	4	5	6	7	8	10	11	12	13	14	15			
336	A	L	A	I	W	I	M	D	D	G	C	K	L	G	RNA	RNA	maturase,S.cerevisiae,MRBY
1233	L	K	L	N	H	L	V	D	D	K	M	H	A	R	DNA	RNA	pol II beta,E.coli,RNECB
948	L	K	L	I	H	Q	V	D	D	K	I	H	G	R	DNA	RNA	pol II beta,tobacco,RNNTB
943	L	K	L	I	H	Q	V	D	D	K	I	H	A	R	DNA	RNA	pol II beta,liverwort,RNLVB
1093	Q	R	L	R	H	M	V	D	D	K	I	H	A	R	DNA	RNA	pol II beta,S.cerevisiae,RNBY2L
997	Q	R	L	K	H	M	V	D	D	K	I	H	S	R	DNA	RNA	pol II beta,D.melanogaster,RNFF2L

Figure 2. A possible catalytic Asp-Asp region in yeast cytochrome b mRNA maturase (see PIR entry MRBY for the sequence). Sequences are also listed for a possible Asp-Asp span in beta chains of DNA-directed RNA polymerase II's from E. coli (27) (RNECB), tobacco chloroplast (5) (RNNTB), liverwort chloroplast (6) (RNLVB), yeast (11) (RNBY2L), and fruit fly (28) (RNFF2L). The sequence position numbers of the first motif resdue are given in the leftmost columns.

pointing to its functional importance. Finally, another possible candidate was found within the beta subunits of DNA-directed RNA polymerase II from various species (Figure 2). The Asp-Asp span found in the C-terminal portion of the subunit is conserved in E. coli, tobacco and liverwort chloroplasts, yeast, and fruit fly. However, rules (3), (5), and (6) are violated by some of the sequences. Interestingly, inclusion of the beta subunit as a protein with an Asp-Asp motif would result in a polymerase example from all possible combinations in direction and product over RNA and DNA. The available experimental evidence (see ref. 10 for a review) points to the beta chain as the major contributor to the catalytic site (11,12). Allison et al. (13) have argued that the beta-prime subunit may be the catalytic one as they find some homology over relatively short spans with the E. coli DNA-directed DNA polymerase I with known tertiary structure (14). The homologies involve a conserved Asp-Pro-Asn-Leu segment and a Glu-X-X-Arg-Ala-X-Ala span (X is any amino acid).

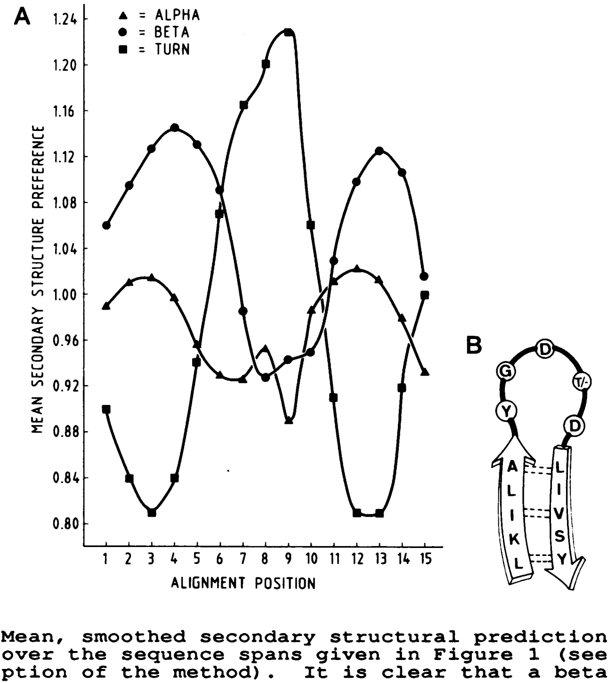

Figure 3A. Mean, smoothed secondary structural prediction preferences over the sequence spans given in Figure 1 (see ref. 29 for a description of the method). It is clear that a beta-hairpin (strand-loop-strand) is predicted.
Figure 3B. Depiction of the putative beta-hairpin using the sequence from rhinovirus with a Thr inserted as found in the motifs of DNA-directed DNA polymerases. The doublet dotted lines indicate mainchain hydrogen bonds in the anti-parallel strands.

A search of the PIR sequence data base yielded at least 12 unrelated and nonpolymerase proteins that displayed the first pattern while nearly 30 obeyed the second. Of course, site directed mutagenesis and other empirical tests are required for both the maturase and beta pol II molecules.

Two recent experiments lend considerable credance to the importance of the core segment in replicase activity. Inokuchi and Hirashima (15) examined mutants involving the Gly in position 7 in the RNA-directed RNA polymerase beta subunit of bacteriophage Q-beta. They found considerably reduced replicase activity in vivo for the altered molecules while mutation of another Gly far removed from the core sequence site had only a slight inhibitory effect. Hizi et al. (16) inserted five residues in place of the conserved Tyr (position 6) in the Asp-Asp motif of human immunodeficiency virus reverse transcriptase; the polymerizing function was destroyed.

The average secondary structure prediction over all the segments of Figure 1 was definitive (Figure 3A): a strand-loop-strand structure (the "beta-hairpin") with the Asp-Asp or Asp-Thr-Asp contained in an exposed loop (depicted in Figure 3B). Mean predictions only over sequences with a specific nucleotide dependence and product consistently pointed to the beta-hairpin structure. An exposed loop could easily accommodate the inserted Thr found in the DNA polymerases.

Several active site scenarios can be imagined: both aspartates are directly involved in polymerase activity; both aspartates bind magnesium (or some appropriate cation(s)) which in turn acts catalytically or in binding phosphate; one aspartate binds the magnesium cation while the other acts catalytically directly; the aspartic amino acids are conserved for purely structural reasons. There is ample experimental evidence supporting the importance of magnesium in RNA-directed RNA polymerases (17), reverse transcriptases (18), and RNA-directed DNA polymerases (19). An examination of the known tertiary protein architectures in the Brookhaven Data Bank (20) shows three examples where an aspartic acid binds a magnesium ion directly: E. coli elongation factor EF-Tu (21), yeast phosphoglycerate kinase (22), and E. coli DNA polymerase I. However, there are no examples of two consecutive Asp's or two separated by one residue binding zinc, copper, magnesium, manganese, or iron. The closest possibility was the Glu8-Leu9-Asp10 structure binding the manganese cation in concanavalan A (23). It could thus be possible that one Asp binds an appropriate cation while the other acts directly in catalysis. Though it is appealing to assign a function to the often conserved Tyr at position 6 as phosphate binding, the lack of invariance would diminish this proposal.

Hodgman (32) has recently found that the Gly-Asp-Thr sequence in the DNA-directed DNA polymerases is also contained in three other RNA plant virus proteins. The latter spans are not in agreement with those presented here. It must be emphasized that the present work lists and relates the Asp-Asp motif in polymerases from all known (to the author) sequences of RNA plant viruses in contrast to Hodgman's three examples despite his efforts to find others. It must be emphasized that there are still many catalytic polymerases without a credible Asp-Asp motif in their sequences (e.g., E. coli DNA polymerases I and III). Since the sequence pattern is relatively short, locating it cannot guarantee discovery of the active site component. There are 12 unrelated proteins in a data base of nearly 6800 sequences that contain a Tyr-Gly-Asp-Asp span but bear little relationship to polymerases in function. Utilization of rules (1) to (6) yields all the polymerase examples of Figure 1 but still maintains an identification error rate of one in 150, odds that require caution.

REFERENCES

1. Kamer, G. & Argos, P. (1984) Nucleic Acids Res. 12, 7269-7282.
2. Wong, S. W. et al. (1988) EMBO J. 7, 37-47.
3. Johnson, M. S., McClure, M. A., Feng, D. F., Gray, J. & Doolittle, R. F. (1986)Proc. Natl. Acad. Sci. USA 83, 7648-7652.
4. Argos, P. (1987) J. Mol. Biol. 197, 331-348.
5. Shinozaki, K. et al. (1986) EMBO J. 5, 2043-2049.
6. Ohyama, K. et al. (1986) Nature 322, 572-574.
7. Sidman, K. E., George, D. G., Barker, W. C. & Hunt, L. T. (1988) Nucleic Acids Res. 16, 1869-1871.
8. Dayhoff, M. O., Schwartz, R. M. & Oscutt, B. C. (1978) in Atlas of Protein Sequence and Strucutre, Vol. 5, Suppl. 3, 345-358 (National Biomedical Research Foundation, Washington, D. C.).
9. Lazowska, J., Jacq, C. & Słonimski, P. P. (1980) Cell 22, 333-348.
10. Armaleo, D. (1987) J. theor. Biol. 127, 301-314.
11. Sweetser, D., Monet, M. & Young, R. A. (1987) Proc. Natl. Acad. Sci. USA 84, 1192-1196.
12. Panka, D. & Dennis, D. (1985) J. Biol. Chem. 260, 1427-1431.
13. Allison, L. A., Moyle, M., Shales, M. & Ingles, C. J. (1985) Cell 42, 599-610.
14. Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G. & Steitz, T. A. (1985) Nature 313, 762-766.
15. Inokuchi, Y. & Hirashima, A. (1987) J. Virol. 61, 3946-3949.
16. Hizi, A., McGill, C. & Hughes, S. H. (1988) Proc. Natl. Acad. Sci. USA 85, 1218-1222.
17. Boccardo, G. & Accotto, G. P. (1988) Virology 163, 413-419.
18. Tanese, N., Sodroski, J., Haseltine, W. A. & Goff, S. P. (1986) J. Virol 59, 743-745.
19. Katinka, M. D. (1987) Eur. J. Biochem. 163, 569-575.
20. Bernstein, F. C. et al. (1977) J. Mol. Biol. 112, 535-542.

21. Jurnak, F. (1985) Science 230, 32-36.
22. Watson, H. C. et al. (1982) EMBO J. 1, 1635-1640.
23. Hardman, K. D. & Ainsworth, C. F. (1972) Biochemistry 11, 4910-4919.
24. Domier, L. L., Shaw, J. G. & Rhoads, R. E. (1987) Virology 158, 20-27.
25. Wu, S., Rinehart, C. A. & Kaesburg, P. (1987) Virology 161, 73-80.
26. Forster,. R. L. S., Bevan, M. W., Harbison, S. A. & Gardner, R. C. (1988) Nucleic Acids Res. 16, 291-302.
27. Ovchinnikov, Y. A. et al. (1981) Eur. J. Biochem. 116, 621-629.
28. Falkenburg, D. Dworniczak, B., Faust, D. M. & Bautz, E. K. F. (1987) J. Mol. Biol. 195, 929-937.
29. Argos, P. (1985) EMBO J. 4, 1351-1355.
30. Jung, G., Leavitt, M. C., Hsieh, J.-C. & Ito, j. (1987) Proc. Natl. Acad. Sci. USA 84, 8287-8291.
31. Davison, A. J. and Scott, J. E. (1986) J. gen. Virol. 67, 1759-1816.
32. Hodgman, J. C. (1986) Nucleic Acids Res. 14, 6769.