Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria

Tatyana V.Ilyina[1] and Eugene V.Koonin[1,2*]
[1]Institute of Microbiology, Academy of Sciences, 117811 Moscow, Russia and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received March 11, 1992; Revised and Accepted May 26, 1992

ABSTRACT

An amino acid motif was identified that consists of the sequence HisHydrHisHydrHydrHydr (Hydr—bulky hydrophobic residue) and is conserved in two vast classes of proteins, one of which is involved in initiation and termination of rolling circle DNA replication, or RCR (Rep proteins), and the other in mobilization (conjugal transfer) of plasmid DNA (Mob proteins). Based on analogies with metalloenzymes, it is hypothesized that the two conserved His residues in this motif may be involved in metal ion coordination required for the activity of the Rep and Mob proteins. Rep proteins contained two additional conserved motifs, one of which was located upstream, and the other downstream from the 'two His' motif. The C-terminal motif encompassed the Tyr residue(s) forming the covalent link with nicked DNA. Mob proteins were characterized by the opposite orientation of the conserved motifs, with the (putative) DNA-linking Tyr being located near their N-termini. Both Rep and Mob protein classes further split into several distinct families. Although it was not possible to find a motif or pattern that would be unique for the entire Rep or Mob class, unique patterns were derived for large subsets of the proteins of each class. These observations allowed the prediction of the amino acid residues involved in DNA nicking, which is required for the initiation of RCR or conjugal transfer of single-stranded (ss) DNA, in Rep and Mob proteins encoded by a number of replicons of highly diverse size, structure and origin. It is conjectured that recombination has played a major part in the dissemination of genes encoding related Rep or Mob proteins among the replicons exploiting RCR. It is speculated that the eucaryotic small ssDNA replicons encoding proteins with the conserved RCR motifs and replicating via RCR-related mechanisms, such as geminiviruses and parvoviruses, may have evolved from eubacterial replicons.

RATIONALE AND APPROACH

Rolling circle replication (RCR) is one of the basic mechanisms by which circular replicons replicate (1). These replicons (Table 1) include small isometric and filamentous single-stranded (ss) DNA bacteriophages (prototyped by phiX174 and M13, respectively; reviewed in ref. 2), a number of ssDNA plasmids (termed so for the existence of single-stranded circular intermediates in their replication) replicating primarily but not exclusively in gram-positive bacteria (reviewed in refs. 3,4), and P2 and related temperate dsDNA bacteriophages (reviewed in ref. 5). A specific version of RCR including the cell to cell transfer of the displaced DNA strand is utilized in the conjugal mobilization of different types of bacterial plasmids (reviewed in refs. 6–8), and in the transfer of Ti plasmids from Agrobacterium to plant cells (reviewed in ref. 9). Recently strong evidence has been reported for the RCR replication of a very different class of circular ssDNA replicons, the plant geminiviruses (10,11). A modification of RCR, the so-called rolling hairpin mechanism, has been implicated in the replication of animal parvoviruses whose genome is linear ssDNA with terminal hairpins (reviewed in ref. 12). Strikingly, RCR also has been demonstrated to be the mode of replication of tiny circular RNAs pathogenic for plants and animals, viroids (virusoids) and hepatitis delta virus, respectively (reviewed in ref. 13).

Apparently, in all DNA replicons that replicate via RCR, it is initiated by a protein encoded by the replicon itself. These proteins possess a DNA nicking-closing and a topoisomerase-like activities (e.g. refs. 14,15). In phage phiX174, by far the best understood RCR system, the phage-encoded A protein nicks the *ori* site in the viral strand of the double-stranded replicative form and remains covalently linked to the 5′ end of the cleaved strand. The 3′ end is then extended by the DNA polymerase

* To whom correspondence should be addressed at: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

```
                                1           2           3
                                                        *  *
                            --------    ---------    ----------
consensus                   futltxxx    xpHuHuuux    uxxYuxkxxx
                              uyp        u   a             h
1  phiX174  A    (200-390)  FDTLTLAD 53 RLHFHAVHF 70 VGFYVAKYVN  A04239
2  G4       A    (241-431)  FDTLTLAD 53 RLHFHAVHF 70 VGFYVAKYVN  A04240
3  S13      A    (209-399)  FDTLTLAD 53 RLHFHAVHL 70 VGFYVAKYVN  JS0450
4  SPV4     gp2  (76-218)   FVTLTYSD 50 RPHYHICFF 44 -ANYTARYTT  H29825
5  Chp1     P5   (140-284)  VLILTYDN 45 RMHWHMIVF 48 I-FYVARYVQ  JUO348
6  BP186    A    (319-487)  FYTLTAPS 48 TPHWHMLMF 47 TG-YVAKYIS  S10632
7  EC67 2/3 (357-400;1-116) FYTITCPS 48?TVHWHLMCF 45 TS-YIAKYIS  JQ0852/3
8  PHASYL   ARP  (87-249)   FLTLTFRD 37 RIHYHLLVA 56 IGRYVGKYIS  S02390
9  pEPLX    RAP  (59-232)   FITLTLPP 50 ALHLHIVMV 92 ASAYMGKYLS M81382(GB)
10 pHGN1    REP  (103-238)  MVTLTAST 54 YAHIHLGVF 58 LGAYLAAYMA  S06780
11 pGRB1    REP  (94-226)   MVTLTASS 48 YVHIHLGVF 57 LGAYLAAYMA  S10152
12 pEHSPN   REP  (50-182)   MVTLAASS 51 YVHIHLGVF 47 LGAYLAAYMA  S00941

13 pBAA1    REP  (87-226)   FLTLTVRN 48 HPHFHVLIP 64 ISKYPVKDTD A32059
14 pFTB14   REP  (119-258)  FLTLTVRN 48 HPHFHVLLP 63 ISKYPVKDTD S01098
15 pLP1     REP  (108-244)  FLTLTVKN 49 NQHLHVLLF 56 TAKYEVKSAD M31323(GB)
16 pUB110   REP  (124-255)  FLTLTVKN 49 NQHMHVLVC 55 TAKYPVKDTD M19465
17 pC194    REP  (97-221)   FLTLTTPN 48 NPHFHVLIA 49 MAKYSGKDSD M64604(GB)
18 pLAB1000 REP  (106-236)  FLTLTAEN 47 HQHMHVLLF 56 TAKYQVKSKD B35390
19 pBC1     REP  (31-168)   FLTLTVRN 53 HPHFHVLLC 67 VSKYPVKDTD M64604
20 pKYM     REP  (116-252)  FLTLTVRN 46 HPHFHCLLM 59 TLKYSVKPED M38574(GB)
21 pSK89    REP  (74-197)   FLTLTTPN 48 NPHFHVLMA 49 MAKYSGKDSD M37889(GB)
22 pNost    REP  (135-262)  FVTLTVKN 50 HPHFHVLMM 49 VIKYSVKESD M81381(GB)
23 pTD1     REP  (114-240)  FITLTVKN 51 HPHYHILAA 50 VAKYSVKATD M87856(GB)

24 ABMV     AC1  (15-110)   FLTYPQCS 32 EPHLHVLIQ 36 VKSYIDKDGD X15983(GB)
25 PYMV     AC1  (15-110)   FLTYPQCS 32 EPHLHVLIQ 36 VKSYVEKDGD JUO364
26 BGMV     AC1  (15-110)   FLTYPRCT 32 EPHLHALIQ 36 VKEYIDKDGV M10070(GB)
27 TGMV     AC1  (16-110)   FLTYPQCS 32 QPHLHVLIQ 36 VKTYIDKDGD A04170
28 CLV      AC1  (14-109)   FLTYPKCS 32 EPHLHALIQ 36 VKSYLDKDGD S07594
29 BCTV      C1  (15-110)   FLTYPQCS 32 QPHLHVLLQ 36 VKSYVDKDGD X04144(GB)
30 TYLCV     C1  (13-108)   FLTYPNCS 32 EPHLHVLIQ 36 VKTYVEKDGN X15656(GB)
31 MSV       C1  (18-107)   FLTYPKCP 32 SLHLHALLQ 30 VRDYILKEPL A04171
32 DSV       C1  (15-104)   FLTYSKCD 32 SLHSHALVQ 30 VRTYILKNPV M23022(GB)
33 WDV       C1  (18-113)   FLTYPECT 32 SPHLHVLVQ 37 VRDYITKEVD B24356
34 CSMV      C1  (42-131)   FLTYPRCP 32 EPHLHAFVQ 30 TLKYCMKHPE JU0043
35 MiSV      C1  (15-110)   FLTYPHCN 32 DPHLHVLIQ 30 VfGYISKTNG D01030(GB)
36 SLCV      C1  (15-110)   FLTYPRCD 33 SPHLHCLIQ 36 VKNYITKEGD M63155(GB)
37 SSV       C1  (15-110)   FLTYSRCP 32 GYHIHVLAQ 31 VRAYAMKNPV M82918(GB)

38 pADB201  REP  (12-118)   LLVYPDSA 34 KPHYHIVLA 30 MWRYMTH--K A32259
39 pMV158   REP  (11-128)   FLLYPESI 37 KAHYHVLYI 34 MYLYLTHESK S05981
40 pE194    REP  (22-132)   FVLYPESA 32 KEHYHILVM 30 LVRYMlH--M A04487
41 pWV01    REP  (13-147)   FLLYPDSI 44 KPHYHVIYI 34 SYEYLTHESK X56954
42 pFX2     REP  (13-148)   FLLYPDSI 45 NPHYHVYIL 34 SYEYLTHESK X54310
43 pLB4     REP  (32-137)   IVVYPESL 30 KSHYHLVLN 30 AVRYLTH--M JQ0181

44 pIJ101   REP  ( 62-234)  LVTFTARH 78 HPHIHAIVL 69 LAEYIAKTQD A31844
45 pSB24    REP  ( 62-234)  LVTFTARH 76 HPHIHAIVL 71 LGEYIAKTQD S04020
46 IS801    REP  (113-272)  HLVFTLPD 50 HPHVHLSVT 83 LGRYLKKPPI S15163

47 ?pCHL1   REP  (  9-108)  FIKSPIHL 33 SSHYHALAA 42 LEAYGVKRYK S02220
48 ?pCpA1   REP  (  8-112)  IIKSSLHL 36 PSHYHALAA 42 LEAYGVKRYK X62475(GB)

49 ?ColE3(E2)REP (33-131)   IAILARFI 40 NGHAHLLYA 32 DVNYSGLICK S04456

50 ?CAA     p52  (316-411)  FATLTALG 29 GQRWHTLVP 41 TATYALKEPV M81223(GB)

51 ?CFDV    p17  ( 25-148)  CFSSTESR 48 RSHFHITIG 49 ERTYCTSTSR M29963(GB)

52 MVM      NS1  (175-243)              GWHCHVLIG 51 LLTYKHKQTK A29510
53 CPV      NS1  (127-196)              GWHCHVLLH 51 ILTYRHKQTK A29962
54 FPLV     NS1  (127-196)              GWHCHVLLH 51 ILTYRHKQTK A36608
55 MEV      NS1  (127-196)              GWHCHVLLH 51 ILTYRHKQTK A38350
56 B19      NS1  ( 79-147)              GYHIHVVtG 50 IENYLMKKIP B24299
57 ADV      NS1  (154-231)              QFHIHCCLG 60 PYKYFNKQTK A35529
58 AAV      NS1  ( 88-162)              YFHMHVLVE 56 IPNYLLPKTQ A03694
59 ADNV     NS1  (399-456)              GDHIHILFS 42 IL-YCIRYGI M37899(GB)
                                         u   a             h
consensus                               xpHuHuuux    uxxYuxkxxx
                                                        *
                                        ---------    ----------
                                            2           3
```

machinery, whereas A protein still bound to the 5′ end of the strand being displaced and complexed with Rep helicase is translocated along the template resulting in the formation of a looped rolling circle. When the replication proceeds the complete circle and the *ori* site is regenerated on the progeny strand, A protein cleaves this site and is transferred to the progeny strand to initiate a new round of replication, whereas the parental strand is concomitantly ligated yielding a single-stranded circle. Thus A protein mediates not only initiation but also termination of RCR. A very similar mechanism has been demonstrated for several ssDNA plasmids of gram-positive bacteria (3; 16−18).

Numerous complete or partial sequences of RCR replicons have been reported. Comparisons of the amino acid sequences of the RCR initiators have led to the delineation of three families of ssDNA plasmid-encoded proteins, with proteins belonging to each of them being obviously related to one another but apparently not to proteins of the other families (15; 19−21). In addition, limited similarities have been noticed between the short sequences surrounding the (putative) DNA-linking Tyr residues in the RCR initiator proteins of various groups of ssDNA plasmids and isometric phages (3, 18,20−22).

We were interested in comparing the sequences of all known RCR initiator proteins in an attempt to reveal putative motifs universally associated with this function and to gain some insight into the evolution of RCR replicons. The results reported in this paper show that an unexpectedly broad class of RCR replicons but not all of them share conserved amino acid sequence motifs, and that the genes for the respective RCR proteins may have a common origin.

Alignments of the RCR initiator proteins of three groups of bacterial plasmids prototyped by pT181 (19), pUB110 (20,21,23), and pMV158 (15) have been published. In addition, we generated an alignment of the A proteins of the phiX174 group, the related proteins of small ssDNA viruses infecting *Spiroplasma* and *Chlamydia*, and A proteins of two P2-related phages. Data base screening using the program BLASTP (24) revealed significant similarities only among members of the same family. On the other hand, inspection of the alignments showed, somewhat unexpectedly, that the pUB110 family, the pMV158 family, and the phage family (but not the pT181 family) each encompassed three best conserved motifs that seemed to be related in all three families in terms of both specific amino acid residues conserved and the relative location of the motifs in the polypeptides. The most prominent of these motifs that had the formula HHydrHHydrHydrHydr (Hydr—bulky hydrophobic amino acid residue) was exploited to screen the Non-Redundant Database (NRDB), which is created in the National Center for Biotechnology Information (NCBI) by merging together the non-redundant entries from PIR, Swissprot, and the translated versions of Genbank, using the program DBSITE (J.-M.Claverie, NCBI). Briefly, numerical weight is ascribed to each amino acid residue in each position of the motif, which is a function of the frequency of the given residue in the multiple alignment, and the data base is screened for sequences containing segments scoring above an empirically defined cut-off. The cut-off values were selected so that either one or two of the bulky hydrophobic residues (irregardless of their position in the motif) were allowed to be substituted by any other residue, to detect putative relevant sequences with deviations from the motif formula. The sequences thus retrieved were further scrutinized for the presence of appropriately located segments resembling the other two motifs, and for their possible functional relevance to RCR (using the available literature). The significance of the revealed relationships was checked using the multiple alignment program OPTAL (25). This approach has led to the delineation of two vast classes of related proteins, one of which encompassed 'true' RCR initiators, while the other consisted of proteins mediating plasmid DNA mobilization.

The RCR initiator (Rep) protein class

This class compiled proteins mediating initiation and termination of RCR not coupled to bacterial conjugation. These proteins were characterized by a coherent arrangement of the three conserved motifs, N-1-2-3-C (Fig. 1). In some of the peripheral members of the class motif 1 was too degenerate to be recognized. Motif 3 included the (putative) DNA-linking Tyr residue(s). It has been shown that in the A protein of phage phiX174 two tyrosines separated by three amino acid residues covalently bind to the 5′

Fig. 1. Conserved sequence motifs in proteins mediating RCR initiation (Rep class). The aligned motifs are excerpts of complete alignments generated by program OPTAL as previously described (25). The motifs are designated as indicated in the text. The amino acid residue numbering in each protein is shown in parentheses. The 'consensus' line includes amino acid residues conserved in all the aligned sequences (upper case; note, however, that one of the conserved His residues in motif 2 is apparently replaced by Arg in the putative initiator protein of the CAA replicon), or in at least one-half of them (lower case); U(u)−a bulky hydrophobic residue (I,L.V,M.F,Y,W); x - *no consensus in this position. The (putative) active Tyr residue(s) is marked by an asterisk(s). The proteins, for which the identification of* the motifs should be considered tentative because of the absence of closely related sequences, are denoted by question marks. Distinct groups of related proteins are separated by blanks. 1−12, superfamily I—bacteriophage A proteins and related cyanobacterial and archaebacterial plasmid Rep proteins with two (putative) active Tyr residues. In the genetic element containing the retron EC67 (sequence 7), the sequences related to the A protein of bacteriophage 186 were found in two distinct overlapping ORFs, 2 and 3, with motif 2 located upstream from the proposed initiator codon of ORF 3 (35). SPV4, *Spiroplasma* virus 4; Chp1, *Chlamydia psittaci* phage 1. 13−43, superfamily II—Rep proteins of eubacterial plasmids and geminiviruses with one (putative) active Tyr residue. 13−23−pUB110-related plasmid family. pNost is an unnamed plasmid from *Nostoc sp*. 24−37—geminiviruses. Bipartite geminiviruses: ABMV—abutilon mosaic virus, PYMV—potato yellow mosaic virus, BGMV—bean golden mosaic virus; TGMV—tomato golden mosaic virus, CLV—*Cassava* latent virus, SLCV—squash leaf curl virus,. Monopartite geminiviruses: BCTV—beet curly top mosaic virus, TYLCV—tomato yellow leaf curl virus, MSV—maize streak viruses, DSV—*Digitaria* streak virus, WDV—wheat dwarf virus, CSMV—*Chloris* striate mosaic virus, MiSV—*Miscanthus* streak virus, SSV—sugarcane streak virus. 38−43—pMV158-related plasmid family. 44−46—pIJ101-related plasmid family. The published sequence of plasmid pSB24 (46) appeared to contain a frameshift disrupting the similarity with the pIJ101 Rep protein in the C-terminal region and masking the putative active Tyr. The sequence related to that of pIJ101 was found in an alternative ORF, and the reconstructed version of the sequence is presented. IS801 is an insertion sequence from *Pseudomonas syringae* that also has been found in the indigenous plasmid pMMC7105 (47). The functional significance of the similarity between the protein encoded by IS801 and Rep proteins of pIJ101 and pSB24 remains to be elucidated. 52−59—NS1 (non-structural) proteins of parvoviruses. In these proteins motif 1 could not be identified. MVM—minute virus of mice, CPV—canine parvovirus, FPLV—feline panleucopenia virus, MEV—mink enteritis virus, B19—human parvovirus, isolate B19, AAV—adeno-associated virus, ADV—Aleutian disease of mink virus, ADNV—*Aedes albopictus* densonucleosis virus. The amino acid sequences were extracted from the PIR bank (Release 31) or were translated using the respective nucleotide sequences from GenBank (Release 71). For each sequence the PIR accession number, or where not available, the GenBank (GB) accession number for the respective nucleotide sequence is indicated.

Table 1. Comparison of the replicons encoding RCR initiator proteins

REPLICON	DNA	DNA	REPLICATION	ENCODED PROTEINS		REFERENCE[a]
	STRUCTURE	SIZE, kb	TYPE	Rep	Mob	
Small isometric coliphages	circular ssDNA	5.4–5.5	rolling circle	Superfamily I, two active Tyr (1–3)[b]	None	2,26
SpV4	circular ssDNA	4.4	rolling circle	Superfamily I two active Tyr (4)	None	33
Chlp1	circular ssDNA	4.9	rolling circle	Superfamily I two active Tyr (5)	None	34
Coliphage 186, retron EC67	linear dsDNA with sticky ends able to circularize	24	rolling circle	Superfamily I two active Tyr (6,7)	None ?	5,31,35
Phasyl	circular ssDNA	1.3	rolling circle	Superfamily I two active Tyr (8)	None	36,37
Cyanobacterial plasmid pEE	circular ds/ssDNA ?[c,d]	?	rolling circle ?	Superfamily I two active Tyr (9)	None	
Archaebacterial plasmids pGRB1, pHGN, pEHSPN	circular ds/ssDNA	1.7–1.8	rolling circle	Superfamily I two active Tyr (10–12)	None	38
Monopartite geminiviruses (e.g. MSV)	circular ssDNA	2.7–3.0	rolling circle	Superfamily II, one active Tyr (29–37)	None	10,39
Bipartite geminiviruses (e.g. CLV)	circular ssDNA, 2 molecules	5.1–5.5	rolling circle	Superfamily II one active Tyr (24–28)	None	11,39
Gram-positive bacterial plasmids, pMV158 family (e.g. pADB201)	circular ds/ssDNA	1.6–2.1	rolling circle	Superfamily II one active Tyr (38; 41–43)	None	3, 4, 15
Gram-positive bacterial plasmids, pMV158 family (e.g. pMV158)	circular ds/ssDNA	3.7–5.5	rolling circle	Superfamily II one active Tyr (39,40)	Family 2 (11–13)	3, 4, 15
Gram-positive bacterial plasmids, pUB110 family (e.g. pLP1)	circular ds/ssDNA	1.6–2.1	rolling circle	Superfamily II one active Tyr (13–15, 17, 19–23)	None	3, 4, 18, 20–23
Gram-positive bacterial plasmids, pUB110 family (e.g. pUB110)	circular ds/ssDNA	3.3–4.5	rolling circle	Superfamily II one active Tyr (16,18)	Family 2 (14, 17)	3, 4, 18, 20–23
Gram-positive bacterial plasmids, pIJ101 family	circular ds/ssDNA ?	3.7–8.8	rolling circle ?	Separate group within the 'Rep' class (44–46)	None	3
Chlamydial plasmids pCHL1, pCpA1	circular ds/ssDNA	7.5	rolling circle ?	Separate family within the 'Rep' class (47,48)	None	
Promiscuous plasmids, IncQ family	circular dsDNA	8.7–12.6	theta	None	Family 3 (21–23)	40
Promiscuous plasmids IncI, P families	circular dsDNA	app. 120	theta	None	Family 1 (1–4)	8, 32
Agrobacterial Ti plasmids	circular dsDNA	app. 100	theta	None	Family 1 (8–10)	9, 27
Gram-negative bacterial F factor-related plasmids	circular dsDNA	app. 100	theta	None	Separate group within the Mob class (19,20)	6, 7, 30
Gram-negative bacterial ColE2,3 plasmids	circular dsDNA	app.7	?	Separate group within the 'Rep' class (49)	None	41
Parvoviruses	linear ssDNA	4.0–5.5	rolling hairpin	Separate family within the 'Rep' class (52–59)	None	12
Coconut foliar decay virus (circovirus ?)	circular ssDNA	1.3	rolling circle ?	Separate group within the 'Rep' class ? (51)	None	42
Chicken anaemia agent (circovirus ?)	circular ssDNA	2.2	rolling circle ?	Separate group within the 'Rep' class ? (50)	None	43
Gram-positive bacterial plasmids, pT181 family (e.g. pS194)	Circular ds/ssDNA	4.4–4.6	rolling circle	pT181 family unrelated to the 'Rep' class	Family 1 (5–7)	3, 4, 14, 19
Gram-positive bacterial plasmids, pT181 family (e.g. pT181)	circular ds/ssDNA	4.4–4.5	rolling circle	pT181 family unrelated to the 'Rep' class	Family 2 (15–18)	3, 4, 14, 19

[a] Selected references describing functional characterization and/or gene organization of the respective replicons are included; where the available data were limited to sequences, references are not indicated.

[b] The numbers of the respective sequences in Fig. 1 (Rep proteins), or in Fig. 2 (Mob proteins) are indicated in parentheses.

[c] The DNA structure of several groups of plasmids is designated ds/ssDNA to emphasize the existence of ssDNA replicative intermediates.

[d] The question marks indicate that data on the respective item are non-available or uncertain.

```
                                     3           z              2a
                                     *
                                ___________  _______________  ___________
consensus                       xxxxxnxYxx   xxxxHuUUxSfxxge  uxuaxxuHxdx
                                                      w   t     u
1      RP4     TraI   (13-130)  AGL-AN-YIT 41 DKTYHLIV-SFRAGE 22 HQRISAVHNDT
2      R751    TraI   (13-130)  AEL-VK-YIT 41 DKTYHLLV-SFRAGE 22 HQRVSAVHHDT
3      R64     NikB   (47-178)  SRL-VD-YAT 56 DPVFHYIL-SWQSHE 22 HQYVSAVHTDT
4    pTF-FC2   MobA   ( 1- 76)             12 DTINHYVL-SWREGE 22 HQAIYGLHADT
5      pS194   Rlx    (10-114)  SRA-IN-YA- 33 VQA-HtVIQSFKPGE 20 YQVAVYTHTDK
6      pC223   Rlx    (10-114)  SRA-IN-YA- 33 NEG-HVVIQSFKPNE 20 HQVAVYTHNDT
7      pC221   Rlx    (10-114)  SRA-IN-YA- 33 IQA-HTVIQSFKPGE 20 HQVAVYTHTDK
8      pTiA5   VirD2  (20-147)  INQ-LE-YLS 48 ELTTHIIV-SFPAGT 25 YNYLTAFHIDR
9      pTiA6   VirD2  (20-147)  INQ-LE-YLS 48 DLTTHIIV-SFPAGT 25 YNYLTAYHVDR
10     pRiA4   VirD2  (20-147)  INQ-LE-YLS 48 ELTTHIIV-SFPAGT 25 YNYLTAFHIDR

11     pMV158  Mob    (33-145)  RSH-LN-YEL            73         IAYA-SVHLDE
12     pGI2    Mob    (34-144)  KSE-QN-YDL            72         tLYA-MVHMDE
13     pE194   Mob    (34-146)  ETY-KN-YDL            73         MLYA-TVHLDE
14   pLAB1000  Mob    (33-144)  RSH-LN-YDL            73         IRYA-VVHMDE
15    pT181    Mob    (33-144)  KTY-LN-YDL            73         LLYA-TVHMDE
16    pT913    Mob    (33-144)  RSH-EN-YDL            73         IAYA-TVHVDE
17    pUB110   Mob    (33-144)  RTR-EN-YDL            73         IAYA-TVHNDE
18    pTX14-3  Mob    (27-152)  RLH-ENIYFV            85         AVYNMVLHDDE

19 ?R100       TraI   (69-170)  KGR-PG-YDL            56         LVMALFNHDTS
20 ?F          TraI   (69-170)  RHR-PG-YDL            56         LVMALFNHDTS

21    pSC101   Mob    (13-135)  ASPHAD-YIA            80         YQFA--IHNP-
22    RSF1010  MobA   (13-131)  ARAKAD-YIQ            78         PYLA--IHA--
23    pTF1     MobL   (17-169)  ATGAAA-Y--            91         AAVA--LHAP-
                                                                    u
consensus                       xxxxxnxYxx                       uxuaxxuHxdx
                                       *
                                _______3___                      ____2a_____
```

```
                                  2
                                 ______
consensus                   xxpHuHuuuxxuxxxxx
1     RP4      TraI     0   DNLHIHIAINKIHPTRH NA
2     R751     TraI     0   DNLHIHIAINKIHPTRN NA
3     R64      NikB     0   DNLHVHVAVNRVHPETG B38529
4   pTF-FC2    MobA     0   DNLHLHLAINRVHPETL M57717(GB)
5     pS194    Rlx      0   DHYHNHIIINSVNLETG S00935
6     pC223    Rlx      0   DHVHNHIVINSIDLETG X12831(GB)
7     pC221    Rlx      0   DHYHNHIVINSVDLETG A04494
8     pTiA5    VirD2    0   DHPHLHVVVNRRELLGH B37763
9     pTiA6    VirD2    0   DHPHLHVVVNRRELLGH B25063
10    pRiA4    VirD2    0   DHPHLHVVVNRRELLGH S06884

11    pMV158   Mob      0   STPHMHMGVVPF-ENGK A33952
12    pGI2     Mob      0   ATPHMHIGVMPITEDNR S02050
13    pE194    Mob      0   RVPHMHFGFVPLTEDGR A04486
14  pLB1000    Mob      0   KTPHMHMGIVPFDDDKK A35390
15  pT181      Mob      0   KTPHMHYGVVPITDDGR J01764
16  pT913      Mob      0   KTPHMHLGVVPM-RDGK S05987
17  pUB110     Mob      0   QTPHMHLGVVPM-RDGK M19465(GB)
18  pX14-3     Mob      0   ANPHLHINYVPNFESSR X56204(GB)

19  ?R100       TraI    4   PQLHTHVVVANVTQHNG  S10660
20  ?F          TraI    4   PQLHTHAVVANVTQHNG  M54796(GB)

21  Mob    pSC101       7   EQPHAHIMFS--ERIND  X01654
22  MobA   RSF1010      3   ENPHCHLMISE-RIN-D  JH0126
23  MobL   pTF1        28   GNWHAHILLSACHVQPD  S12190
consensus                   xxpHuHuuuxxuxxxxx

                              ____2____
```

Fig. 2. Conserved sequence motifs in proteins mediating initiation of conjugal transfer of plasmid DNA (Mob class). The designations are as in Fig. 1. The motifs are designated as in the Rep proteins, with motif containing the (putative) active Tyr designated 3 in spite of its location upstream from motif 2; z—a specific motif found in the Mob proteins of family 1; 2a—the upstream portion of motif 2 that is separated by a spacer from the downstream portion in sequences 19–23. 1–10, family 1—Mob proteins of IncI and IncP plasmids (1–4), gram-positive bacterial ssDNA plasmids of the pT181 family (5–7), and Ti plasmids of *Agrobacterium* (8–10). For the plasmid pTF-FC2 only a partial sequence has been reported. 11–18, family 2—Mob proteins of gram-positive bacterial ssDNA plasmids. 21–23, family 3—Mob proteins of IncQ plasmids. NA—accession number not available, the sequences were from ref. 32.

Table 2. Unique sequence motifs and patterns in RCR proteins

MOTIF/PATTERN[a]	SET OF PROTEINS SELECTED
2[b]–[PAU]HUH[AU][CU][AU][c]	Bacteriophage A proteins and related proteins with two (putative) active Tyr residues
3–Y[TU]A[KR]Y	
1–[FILV][ILV][ILVT]YP	Rep proteins of pMV158-related plasmids and geminiviruses (one active Tyr)
2–H[ILVFYWST]H[ILVMAC][ILVMFYW][ILVMFYWAC]	
3–[ILVMAST]x_2Y[ILVMAC]x[KH][d]	
2–HxDx2[PU]HxHUxU	Mob proteins of Ti plasmids, IncP and IncI plasmids, and Gram-positive bacterial ssDNA plasmids (families 1 and 2 of the Mob class)[e,f]

[a] We define a motif as a constellation of conserved amino acid residues that may include short spacers of strictly defined length, and a pattern as a group of motifs that may be separated by spacers of arbitrary length (44). Search for a pattern included consecutive screening of NRDB with the respective motifs.

[b] The motifs are numbered as in the text and in Figs. 1 and 2.

[c] The residues shown in brackets are alternatives; U–bulky hydrophobic residue.

[d] x–any residue.

[e] One irrelevant sequence was retrieved upon screening NRDB with this motif.

[f] A more specific version of this motif has been described by Pansegrau and Lanka as the identifier of a set of Mob proteins coinciding with our family 1 (45).

end of the nicked viral strand, and a model of their alternate participation in the cleavage-ligation reaction has been suggested (26). The conservation of these two Tyr residues was a hallmark of a distinct superfamily within the Rep class including mainly bacteriophage A proteins but also Rep proteins of certain halobacterial and cyanobacterial plasmids. Another large superfamily brought together the (putative) RCR initiator proteins of two families of eubacterial plasmids, and unexpectedly of plant geminiviruses (Table 1, Fig. 1). These proteins appear to have only one active Tyr residue. Its tentative identification by site-directed mutagenesis has been described for the plasmid pKYM (22).

The mobilization (Mob) protein class

The proteins of the Mob class contained only two universally conserved motifs, which were oriented differently from the Rep proteins, with the (putative) active Tyr being located N-terminally of motif 2 (Fig. 2). Experimental identification of this residue in VirD2 protein of a Ti plasmid has been reported (27). This class included at least three distinct families, with one of them

(Family 1) uniting Mob proteins of such diverse replicons as Ti plasmids, on the one hand, and small ssDNA plasmids from Gram-positive bacteria, on the other hand (Table 1). The proteins of this family contained additional well defined motifs (Fig. 2). Although the sequence conservation around the putative active Tyr was poor in Mob proteins (Fig. 2), its assignment for the proteins of the families 1, 2 and 3 was confirmed by statistically significant alignments of relatively closely related sequences. Caution is due in the interpretation of the putative motifs in TraI proteins of F-related plasmids that had no close relatives to corroborate the assignments.

Unique sequence patterns

It appeared not to be possible to define a sequence motif or pattern that would selectively extract from the sequence bank all the RCR proteins, or at least the proteins of either of the two classes (Rep or Mob) without retrieving any false positives. However, unique patterns typical of large subsets of these proteins could be derived (Table 2) and hopefully will be useful for easy identification of RCR proteins in newly sequenced replicons.

The possible function of the 'two His' motif

Motif 2 containing two His residues embedded in a highly hydrophobic sequence is the only common denominator of the Rep and Mob classes of the RCR proteins (compare Figs. 1 and 2). The data base searches revealed the (partial) conservation of this motif, in addition to the RCR proteins, in various groups of metalloenzymes, particularly in cytochrome c oxidase polypeptide I, hemocyanins, and carbonic anhydrases. Histidine residues have been shown to act as ligands to metal centers in many enzymes. In cytochrome oxidases the 'two His' motif formula was conserved from bacteria to mammals, and at least one of the two conserved His residues has been implicated as a Cu ion ligand (28). In carbonic anhydrases, superoxide dismutases and procatechuate 3,4-oxygenase His groups located two residues apart and surrounded by hydrophobic residues have been shown to interact with the same metal ion (29). An additional typical feature of many proteins with His as a metal ligand is the presence of a Pro residue within two residues of the binding His (29). A Pro residue was found in the position preceding the first conserved His in about one-half of the RCR proteins (Figs. 1, 2). The reactions catalyzed by these proteins require Mg^{2+} or Mn^{2+} (e.g. refs. 2, 30). Thus it is tempting to speculate that the conserved His residues in the 'two His' motif function as ligands to these metal ions.

Summary of functional predictions

This analysis highlighted the previously unsuspected relationship between the proteins involved in the DNA replication of plant geminiviruses (detailed elsewhere, Koonin & Ilyina, submitted) and animal parvoviruses, and procaryotic RCR proteins. These findings are compatible with what is known of the replication mechanisms of these viruses (see above). Also, the conserved motifs delineated here are predicted to be of crucial importance for the functions of RCR proteins and may serve as plausible targets for site-directed mutagenesis experiments. In particular, the active Tyr residue(s) was predicted for numerous RCR proteins, including A protein of bacteriophage 186 (Fig. 1), and TraI proteins of IncF and IncP plasmids that are objects of intensive studies (30–32). The strength of prediction is the highest when the functional relevance of the motifs could be confirmed by their conservation in an alignment of a family of definitely related sequences. If such evidence was not available, the predictions should be treated with some reserve (Figs. 1,2).

Some implications for the evolution of RCR replicons

It seems unlikely that a similar arrangement of the three conserved RCR motifs, as observed in the Rep proteins (Fig. 1), could have evolved independently in several evolutionary lineages. The hypothesis of divergent rather than convergent origin of the proteins of this class is supported by the fact that these motifs are not universally required in RCR-mediating proteins as shown by the comparative sequence analysis of the Rep proteins of the pT181-related plasmids and filamentous phage gene II proteins (ref. 19, and E. V. K. and T. V. I., unpublished observations). The same notions apply to the Mob protein class. The relationship between these two classes that have only one motif in common remains uncertain.

Related RCR proteins are encoded by extremely diverse replicons, from the smallest and most primitive such as phasyl (apparently the smallest known DNA replicon) and some of the ssDNA plasmids, of which these proteins are the only products, and up to such relatively large and complex as phage 186, *E.coli* F factor, or Ti plasmids (Table 1). Some of the small plasmid replicons encode both a Rep protein and a Mob protein, and there are several cases when of two plasmids with closely related Rep proteins one encodes a Mob protein, whereas the other lacks the respective gene (Table 1). This is compatible with the so-called cassette concept of the evolution of plasmids of gram-positive bacteria, which conjectures that these plasmids consist of two relatively independent gene cassettes, the replication one and the mobilization one that are readily exchangeable (19, 23). Recombination, both at the level of fusion and/or separation of gene portions encoding different domains of the RCR proteins, and at the level of the exchange of the genes encoding RCR proteins between different replicons, appeared to have made a major contribution to the evolution of this type of DNA replication.

Finally, these findings raise the question of the origin of small eucaryotic ssDNA replicons, such as geminiviruses, parvoviruses, and circoviruses, from procaryotic plasmids or phages.

ACKNOWLEDGEMENTS

E.V.K. is grateful to Dr. J.-M.Claverie for providing the program DBSITE, and to Drs. D.J.Lipman and J.C.Wootton for critical reading of the initial draft of this manuscript.

REFERENCES

1. Gilbert, W., and Dressler, D. (1968) *Cold Spring Harbor Symp. Quant. Biol.*33, 473–484.
2. Baas, P. D., and Jansz, H. S. (1988) *Curr. Top. Microbiol. Immunol.* 136, 31–70.
3. Gruss, A., and Ehrlich, S. D. (1989) *Microbiol. Rev.* **53**, 231–241.
4. Novick, R. P. (1989) *Annu. Rev. Microbiol.* **43**, 537–565.
5. Bertani, L. E., and Six, E. W. (1988) In The Bacteriophages (R. Calendar, ed.), Plenum Press, New York, pp. 73–143.
6. Willetts, N., and Wilkins, B. (1984) *Microbiol. Rev.* **48**, 24–41.
7. Willetts, N., and Skurray, R. (1987) In F. C . Neihardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter & H. E. Umbarger (ed.), Escherichia coli and Salmonella typhimurium: cellular and molecular biology, vol. 2. American Society for Microbiology, Washington, DC, pp. 1110–1133.
8. Guiney, D. G., and Lanka, E. (1989) In C. M. Thomas (ed.), Promiscuous plasmids of Gram-negative bacteria. Academic Press Inc., London, pp. 27–56.

9. Zambryski, P. (1988) *Annu. Rev. Genet.* **22**, 1–30.
10. Stenger, D. C., Revington, G. N., Stevenson, M. C., and Bisaro, D. M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8029–8033.
11. Saunders, K., Lucy, A., and Stanley, J. (1991) *Nucleic Acids Res.* **19**, 2325–2330.
12. Berns, K. I. (1990) *Microbiol. Rev.* **54**, 316–329.
13. Symons, R. H. (1989) *Trends Biochem. Sci.* **14**, 445–450.
14. Koepsel, R. R., Murray, R. W., Rosenblum, W. D., and Khan, S. A. (1985) *Proc. Natl. Acad. Sci. USA* 82, 6045–6049.
15. De la Campa, A. G., Del Solar, G. H., and Espinosa, M. (1990) *J. Mol. Biol.* **213**, 247–262.
16. Te Riele, H., Michel, B., and Ehrlich, S. D. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2541–2545.
17. Te Riele, H., Michel, B., and Ehrlich, S. D. (1986) *EMBO J.* **5**, 631–637.
18. Gros, M. F., te Riele, H., and Ehrlich, S. D. (1987) *EMBO J.* **6**, 3863–3869.
19. Projan, S., and Novick, R. (1988) *Plasmid* 19, 203–221.
20. Bouia, A., Bringel, F., Frey, L., Kammerer, B., Belarbi, A., Guyonvarch, A., and Hubert, J.-C. (1989) *Plasmid* 22, 185–192.
21. De Rossi, E , Milano, A., Brigidi, P., Bini, F., and Riccardi, G. (1992) *J. Bacteriol.* **174**, 638–642.
22. Yasukawa, H., Hase, T., Sakai, A., and Masamune, Y. (1991) *Proc. Natl. Acad. Sci. USA* 88, 10282–10286.
23. Josson, K., Soetaert, P., Michiels, F., Joos, H., and Mahillon, J. (1990) *J. Bacteriol.* **172**, 3089–3099.
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
25. Gorbalenya, A. E., Blinov, V. M., Donchenko, A. P., and Koonin, E. V. (1989) *J. Mol. Evol.* **28**, 256–268.
26. Van Mansfeld, A. D., Van Teeffelen, H. A., Baas, P. D., and Jansz, H. S. (1986) *Nucleic Acids Res.* **14**, 4229–4238.
27. Vogel, A. M., and Das, A. (1992) *J. Bacteriol.* **174**, 303–308.
28 Saraste, M. (1990) *Q. Rev. Biophys.* **23**, 331–366.
29. Chakrabarti, P. (1990) *Protein Eng.* **4**, 57–63.
30. Inamoto, S., Yoshioka, Y., and Ohtsubo, E. (1991) *J. Biol. Chem.* **266**, 10086–10092.
31. Sivaprasad, A. V., Jarvinen, R., Puspurs, A., and Egan, J. B., (1990) *J. Mol. Biol.* **213**, 449–463.
32. Ziegelin, G., Pansegrau, W., Strack, B., Balzer, D., Kroger, M. Kruft, V., and Lanka, E. (1991) *DNA Sequence* **1**, 303–327.
33. Renaudin, J., Pascarel, M. C., and Bove, J.M. (1987) *J. Bacteriol.* **169**, 4950–4961.
34. Storey, C. C., Lusher, M., and Richmond, S. J. (1989) *J. Gen. Virol.* **70**, 3381–3390.
35. Hsu, M.-Y., Inouye, M., and Inouye, S. (1990) *Proc. Natl. Acad. Sci. USA* 87, 9454–9458.
36. Seufert, W., Lurz, R., and Messer, W. (1988) *EMBO J.* **7**, 4005–4010.
37. Gielow, A., Diederich, L., and Messer, W. (1991) *J. Bacteriol.* **173**, 73–79.
38. Sioud, M., Baldacci, G., Forterre, P., and De Recondo, A.-M. (1988) *Nucleic Acids Res.* **16**, 7833–7842.
39. Davies, J. W., and Stanley, J. (1989) Trends Genet. **5**, 77–81.
40. Drolet, M., Zanga, P., and Lau, P. C. (1990) *Molec. Microbiol.* **4**, 1381–1391.
41. Kido, M., Yasueda, H., and Itoh, T. (1991) *Nucleic Acids Res.* **19**, 2875–2880.
42. Rohde, W., Randles, J., Langridge, P., and Hanold, D. (1990) *Virology* **176**, 648–651.
43. Claessens, J. A. J., Schrier, C. C., Mockett, A. P. A., Jagt, E. H. J. M. & Sondermeijer, P. J. A. (1991) *J. Gen. Virol.* **72**, 2003–2006.
44. Hodgman, T. C. (1989). CABIOS **5**, 1–13.
45. Pansegrau, W., and Lanka, E. (1991) *Nucleic Acids. Res.* **19**, 3455.
46. Bolotin, A. P., Sorokin, A. V., Aleksandrov, N. N., Danilenko, V. N., Kozlov, Yu. I. (1985) *Dokl. Akad. Nauk SSSR* 283, 1014–1017 (in Russian).
47. Romantschuk, M., Richter, G. Y., Mukhopadhyay, P., and Mills, D. (1991) *Molec. Microbiol.* 5, 617–622.