Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

doi:10.1016/S0022-2836(02)01033-1 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 324, 47–62

JMB

The Palm Subdomain-based Active Site is Internally Permuted in Viral RNA-dependent RNA Polymerases of an Ancient Lineage

Alexander E. Gorbalenya[1,2]*, Fiona M. Pringle[3], Jean-Louis Zeddam[4] Brian T. Luke[1], Craig E. Cameron[5], James Kalmakoff[3] Terry N. Hanzlik[4], Karl H. J. Gordon[4] and Vernon K. Ward[3]

[1]*Advanced Biomedical Computing Center, Science Applications International Corporation/National Cancer Institute, P.O. Box B, Frederick MD 21702-1201, USA*

[2]*Department of Medical Microbiology, Center of Infectious Diseases, Leiden University Medical Center Room L04-036, Albinusdreef 2 Postbus 9600, 2300 RC Leiden The Netherlands*

[3]*Department of Microbiology University of Otago, P.O. Box 56, Dunedin, New Zealand*

[4]*CSIRO Division of Entomology, G.P.O. Box 1700 Canberra, ACT 2601, Australia*

[5]*Department of Biochemistry & Molecular Biology Pennsylvania State University 201 Althouse Laboratory University Park, Philadelphia PA 16802, USA*

**Corresponding author*

Template-dependent polynucleotide synthesis is catalyzed by enzymes whose core component includes a ubiquitous αβ palm subdomain comprising A, B and C sequence motifs crucial for catalysis. Due to its unique, universal conservation in all RNA viruses, the palm subdomain of RNA-dependent RNA polymerases (RdRps) is widely used for evolutionary and taxonomic inferences. We report here the results of elaborated computer-assisted analysis of newly sequenced replicases from *Thosea asigna* virus (TaV) and the closely related *Euprosterna elaeasa* virus (EeV), insect-specific ssRNA + viruses, which revise a capsid-based classification of these viruses with tetraviruses, an Alphavirus-like family. The replicases of TaV and EeV do not have characteristic methyltransferase and helicase domains, and include a putative RdRp with a unique C–A–B motif arrangement in the palm subdomain that is also found in two dsRNA birnaviruses. This circular motif rearrangement is a result of migration of ~22 amino acid (aa) residues encompassing motif C between two internal positions, separated by ~110 aa, in a conserved region of ~550 aa. Protein modeling shows that the canonical palm subdomain architecture of poliovirus (ssRNA +) RdRp could accommodate the identified sequence permutation through changes in backbone connectivity of the major structural elements in three loop regions underlying the active site. This permutation transforms the ferredoxin-like β1αAβ2-β3αBβ4 fold of the palm subdomain into the β2β3β1αAαBβ4 structure and brings β-strands carrying two principal catalytic Asp residues into sequential proximity such that unique structural properties and, ultimately, unique functionality of the permuted RdRps may result. The permuted enzymes show unprecedented interclass sequence conservation between RdRps of true ssRNA + and dsRNA viruses and form a minor, deeply separated cluster in the RdRp tree, implying that other, as yet unidentified, viruses may employ this type of RdRp. The structural diversification of the palm subdomain might be a major event in the evolution of template-dependent polynucleotide polymerases in the RNA–protein world.

Keywords: RNA viruses; RNA polymerases; evolution; protein permutation; ancient palm subdomain

Present addresses: F. M. Pringle, Department of Microbiology, University of Alabama at Birmingham, 845 19th Street S. BBRB 373/17, Birmingham, AL 35294, USA; J.-L. Zeddam, IRD, Pontificia Universidad Católica del Ecuador, Laboratorio de Bioquímica, Apartado 17-01-2184, Quito, Ecuador.

Abbreviations used: aa, amino acid; CD, conserved domain; EeV, *Euprosterna elaeasa* virus; IBDV, infectious bursal disease virus; IPNVJ, infectious pancreatic necrosis virus strain Jasper; PV, poliovirus; TaV, *Thosea asigna* virus; dsRNA, double-stranded RNA; ssRNA + , positive-stranded RNA; RdRp, RNA-dependent RNA polymerase; HMM, Hidden Markov Model; ORF, open reading frames; nt, nucleotide; TDPP, template-dependent polynucleotide polymerase.

E-mail address of the corresponding author: a.e.gorbalenya@lumc.nl

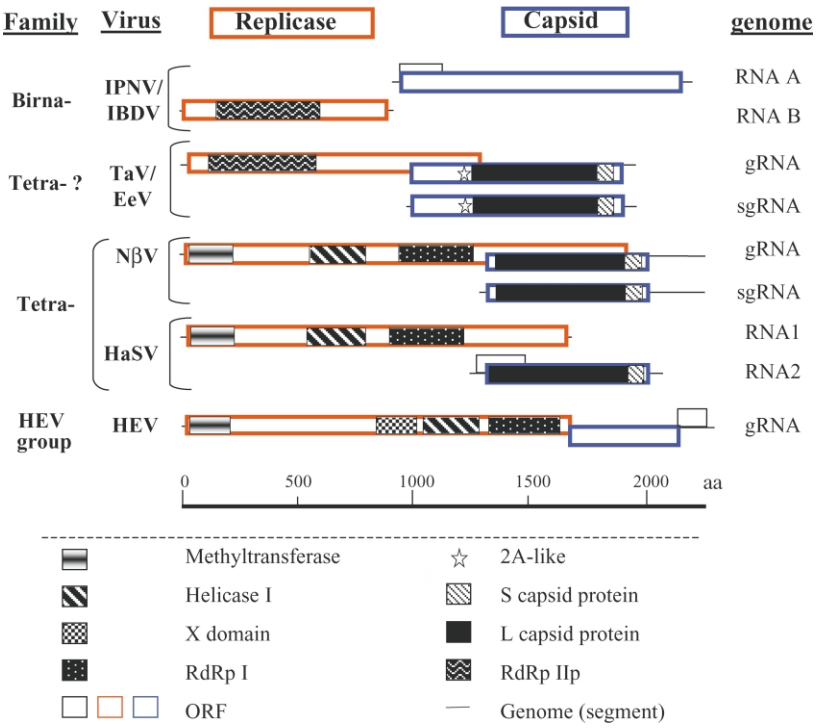


Figure 1. Mosaic domain architecture of replicase and capsid proteins of TaV and EeV. Shown are selected conserved domains (CDs) of replicative and capsid proteins of TaV/EeV, IPNVJ/IBDV, *H. armigera* stunt virus (HaSV) and *Nudaurelia* β virus (NβV) (tetraviruses), and hepatitis E virus (HEV). RNA 1 and 2 ((a) and (b)), RNA segments 1 and 2 ((a) and (b)), respectively; gRNA, genomic RNA; sgRNA, subgenomic RNA packaged into virions; RdRp-I and -IIp, Alphavirus-like RdRp and Picornavirus-like permuted RdRp, respectively. The identification of the RdRp-IIp domain in replicases of TaV, EeV, IPNV and IBDV is detailed in the text and subsequent Figures.

Introduction

The template-dependent polynucleotide polymerases (TDPPs) that replicate cellular and viral genomes are central to life. The DNA genomes of cellular organisms and the majority of DNA viruses are replicated by DNA-dependent DNA polymerases (DdDp). RNA genomes, currently found only in viruses, comprise four types: positive and negative-sense single-stranded RNA (ssRNA + and ssRNA − , respectively) viruses, double-stranded RNA (dsRNA) viruses, and RNA viruses that use reverse transcriptase for genome replication. RNA-dependent polymerase is the only enzyme universally conserved in all of the thousands of known non-satellite RNA viruses. RNA-dependent RNA polymerase (RdRp) is used to replicate the genomes of viruses with no DNA stage, and RNA-dependent DNA polymerase (RdDp; reverse transcriptase) is used by viruses with a DNA stage in the life cycle.[1]

Despite the diversity of genomes that they replicate, the TDPPs have remarkable structural conservation. All the RNA-dependent polymerases, and many DNA-dependent polymerases, employ a fold whose organization has been likened to the shape of a cupped right hand with three subdomains, termed fingers, palm and thumb.[2] Only the palm subdomain, composed of a four-stranded antiparallel β-sheet with two α-helices packed beneath, is well conserved among all of these enzymes.[3–9] The palm subdomain comprises several ordered sequence motifs, with motifs A, B and C[10] being the most prominent. Motifs A and C are conserved in the TDPPs of all cellular organisms and viruses.[5,11–13] In RdRps, motif A ($DX_{4-5}D$, where X is a non-conserved residue) contains two Asp residues separated by four or five residues, while motif C (GDD) contains an Asp-Asp dipeptide, which is often preceded by a Gly.[10] In TDPPs other than RdRps, only the N-terminal Asp residues in motifs A and C are conserved[14] at the end of a β-strand of motif A and in the turn of the β^β hairpin of motif C. These Asp residues are spatially juxtaposed, bind divalent cations, Mg^{2+} and/or Mn^{2+}, and are crucial for catalysis. Motif B forms a long α-helix and is conserved in RNA-dependent polymerases, and, at the secondary structure level, in other polymerases.[4,6] Motif B contains a residue (Asn in RdRp) that contributes to the discrimination between dNTPs and NTPs and thus determines whether RNA or DNA is synthesized.[6,15] Hence, all three motifs are indispensable for proper functioning of polymerases. This structural and functional conservation implies that palm subdomains of all TDPPs may have evolved from a common and ancient ancestor. RdRps also share the palm motif D (α^β structure), and motif E (β^β structure), which is located at the palm–thumb interface; these motifs may not be readily recognized in sequences of every RNA virus.

Due to their universal occurrence and exceptional conservation,[16–18] RdRps, along with a few other replicative proteins, have been used for the identification and classification of RNA viruses. The phylogeny of RdRps mainly parallels the taxonomy of RNA viruses up to the supergroup level.[19] Among ssRNA + viruses, Alphavirus and Picornavirus-like supergroups[20,21] are the most

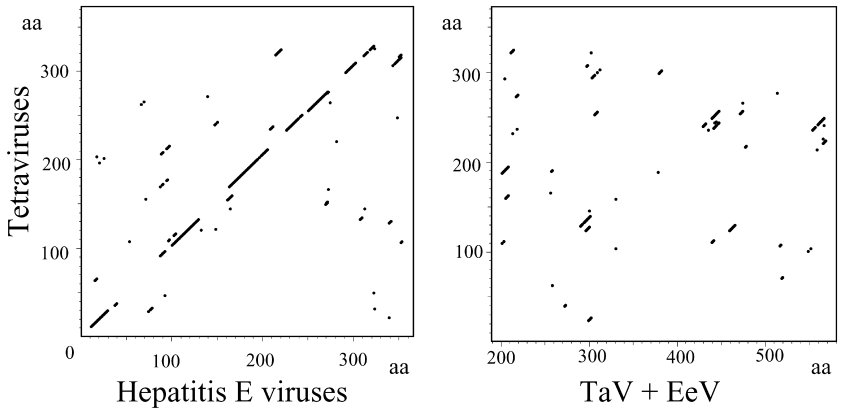


Figure 2. Profile-*versus*-profile dot-plot cross-comparisons of the tetravirus RdRps with HepEV and TaV/EeV RdRps. ClustalX-generated alignments of (putative) RdRp domains of HaSV and NβV (tetraviruses),[31] human and swine hepatitis E viruses,[75,76] and TaV and EeV (see Figure 3(b)) were converted into profiles and compared in a dot-plot fashion, as described in Materials and Methods. Shown are the dot-plots generated using a window of 23 aa residues. Matches between two profiles that were within the top 0.05% are marked by dots.

numerous, each comprising a dozen or so families.[22]

Here, we describe the analysis of the replicases of four RNA viruses from two families. Recently sequenced *Thosea asigna* virus (TaV) (Ref. 23 and this report) and *Euprosterna elaeasa* virus (EeV) (J.-L.Z., F.M.P., K.H.J.G., V.K.W., B.T.L., A.E.G. & T.N.H., unpublished results; GenBank accession number AF461742) are ssRNA + viruses provisionally classified as tetraviruses, an Alphavirus-like supergroup family whose members have only been isolated from lepidopteran insects. The second virus family is the dsRNA birnaviruses, including infectious pancreatic necrosis virus (IPNV) and infectious bursal disease virus (IBDV) that cause highly contagious diseases of young salmonid fish and chickens, respectively.[24,25]

The genomes of TaV and EeV consist of an RNA segment of ~5700 nucleotides (nt) with two open reading frames (ORFs) encoding the putative replicase (see below) and capsid proteins. The capsid precursor is expressed from a subgenomic RNA molecule which, along with genomic RNA, is packaged into virions (Figure 1)[23,26] (J.-L.Z., F.M.P., K.H.J.G., V.K.W., B.T.L., A.E.G. & T.N.H., unpublished results). The genome of birnaviruses consists of segment A (~3300 nt), encoding a precursor to the major capsid proteins, and segment B (~2900 nt), encoding the RdRp [24] (Figure 1). Counterparts of motifs A, B, and C of the palm subdomain and motif E,[10] were tentatively identified in the birnavirus RdRps through comparison with homologs encoded by ssRNA + viruses.[27,28] However, the highly conserved Asp-Asp dipeptide, which is critical for enzymatic activity, was not evident in motif C of the IPNV RdRp.[28] This is in striking contradiction to the replicative competence of birnaviruses.[29,30]

Here we resolve the above conflict, showing that the originally identified motif C in birnaviruses is fortuitous; in fact, a well-conserved motif C is present, but located upstream of motif A in RdRps of birnaviruses as well as TaV and EeV. This organization of the C–A–B motifs is unprecedented amongst viral and cellular TDPPs and yields a palm fold in which the canonical structural elements show a non-canonical connectivity. Our findings further indicate that the RdRps of TaV, EeV and birnaviruses have profoundly deviated from all known RdRps and comprise a unique ancient lineage whose very existence affects our understanding of the evolution of both polymerases and RNA viruses.

Results

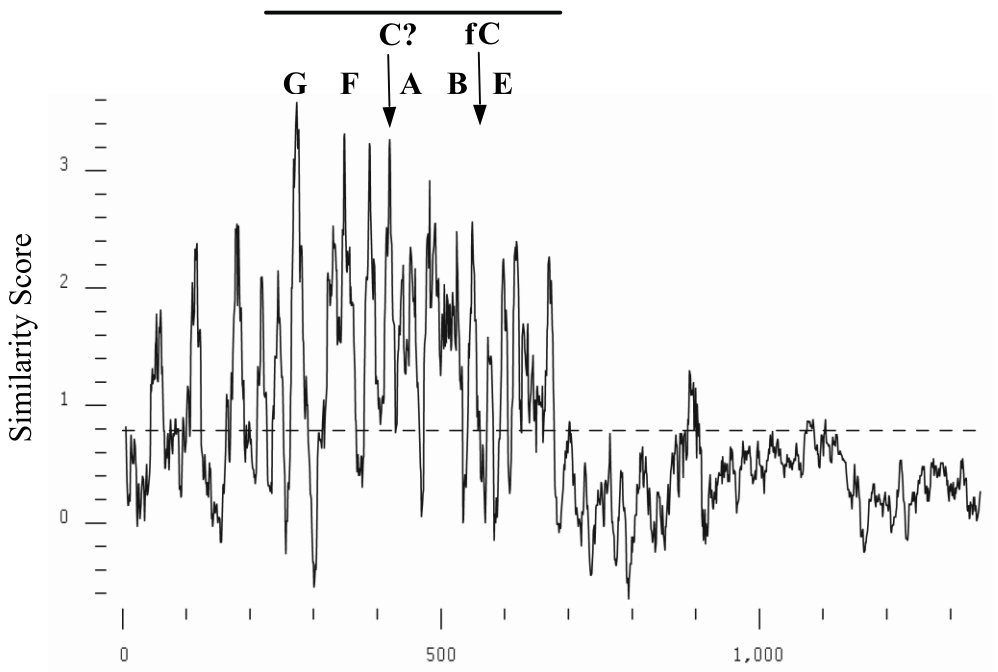
Sequencing of replicase gene of TaV

We have completed sequencing of the TaV putative replicase, of which the C terminus has been reported.[23] The ORF consists of 3771 nt encoding a protein of 1257 aa sharing ~68% identity with the homolog of the same size from EeV, whose sequence was recently determined (J.-L.Z., F.M.P., K.H.J.G., V.K.W., B.T.L., A.E.G. & T.N.H., unpublished results; GenBank accession number AF461742). TaV and EeV have very similar genome organizations and capsid precursors (Figure 1).

Replicases of ssRNA + TaV and EeV and dsRNA birnaviruses are most similar, and they are distantly related to RdRps of Picornavirus-like viruses

On the basis of the conservation of the capsid proteins among TaV, EeV and the well established tetraviruses *Helicoverpa armigera* stunt virus (HaSV) and *Nudaurelia capensis* β virus (NβV)[23] (J.-L.Z., F.M.P., K.H.J.G., V.K.W., B.T.L., A.E.G. & T.N.H., unpublished results), it was expected that putative replicases of TaV and EeV would have the methyltransferase, helicase and RdRp domains that form the backbone of the replicases of tetraviruses (Figure 1).[31,32] Surprisingly, this conservation was not evident in sensitive profile *versus* profile dot-plots (Figure 2, and data not shown). This is opposed to the conservation between replicases of distantly related insect tetraviruses (HaSV

a



Position in alignment of replicases of TaV, EeV, IBDV and IPNV

b

```
                 *         20         *         40         *         60         *         80         *        100
IBDV  : LTRVPSRMLVLTGDVDGDFEVEDYLPKNLKSSSGLPYVGRTKGETIGEMIAISNQFLRELSTLLKQGAGTKGSNKKKLLSMLSDYWYLSCGLLFPAAER : 324
IPNVJ : LRWAPSWLINYTGDLSTD---KSYLPHVTIKSSAGLPYIGKTKGDTTAEALVLADSFIRDLG-----RAATSADPEAGVKKTITDFWYLSCGLLFPKGER : 308
TaV   : FPYDPQKLVDVSGELSELV----LGIKTNAIASAGPPYW-RTKRDALPDMLDCVLPLLYDHI--------VRKDL-TTLRNKHPELFLAEC---KNKTDR : 269
EeV   : FPYDPQKLVDVSRELSDLV----LEIKTNATASAGPPYW-RSKRDALPDMLDCVLPLLYQHI--------VDGDL-ATLKNKHPELFLAEC---KNKTDR : 269
                                  -----G-----

                 *        120         *        140         *        160         *        180         *        200
IBDV  : VDKSTWLTKTRNIWSAFSPTHLMISMITWPVMSNSPNNVLNIEGCPSLVKFNPFRGGLNRIVEWI-----LAPEEPKALVYADN--TYIVHSN--TWYSI : 415
IPNVJ : YTQVDWDKKTRNIWSAFYPTHLLLSMVSTPVMNESKLNITNTQ-TPSLYGFSPFHGGMDRIMTIIRDS--LDNDEDLVMIYADN--TYILQDN--TWYSI : 401
TaV   : YEVESLGEKTRPYFSHPFHLSALVSVLS-QSFSGALKIMTEDSTSFNAYGFSWTNGGAEDLAIWARQAGEAGKKPPRIACYGDDTDIYYRKDGKLYRICP : 368
EeV   : YEVEKLGEKTRPYFSHPFHLSALVSVLS-QSFSKALKIMTEESTSYNAYGFSWTNGGAEDLVIWARGAGEVGKEPPRIACYGDDTDIYYRQKGRLYRICP : 368
                ---F---                                                                  ------C?-------       ---

                 *        220         *        240         *        260         *        280         *        300
IBDV  : LLEKGEANCTRQHMQAAMYYILTRGWS-DNGDPMFNQTWATFA---MNIAPA--LVVDSSCLIMNLQIKTYGQGMGNAATFLNNHLLSTLVLDQWNLMRQ : 509
IPNVJ : LLEKGEANCTPQHMQAMMYYLLTRGWTNEDGSPRYNPTWATFA---MNVAPS--MVVDSSCLLMNLQLKTYGQGSGNAFTFLNNHLMSTIVVAPWVKAGK : 496
TaV   : LFKQMDGSVDATTIEAVVDYVVD---AHVKQYPTARQFWEEVGKLWVEMATQSPFLIDGIKVYR--KMQKDGLMIGVVGTTLFDTVKSALAYNDWADQ-- : 461
EeV   : LFKQMDGSVDETTIKAVVDYVVH---AHTSQYPLAKGFWLEVGKLWVEMATSSPFLIDGTLVYR--KKQKDGLMTGVVGTTLFDTVKSALAYKDWADQ-- : 461
        --A------                                                                  ------------B----------

                 *        320         *        340         *        360         *        380         *        400
IBDV  : PRPDSDEFKSIEDKLGINIKIERSIDDIRGKLRQLVPLA-QPGYLSGG--VEPEQSSPTVELDLLGWSATYSKDLG----IYVPVLDKERLFCSAAYPKG : 602
IPNVJ : PNPMTKEFMDLEEKTGINIKIERELKNLRETIVEAVETAPQDGYLADGSDLPPIRPGKAVELDLLGWSAIYSRQME----MFVPVLENERLIASAAYPKG : 592
TaV   : --LMFGSLNLLEEKYAIEFFKNKHGLVIKEGTWKPALVNEDPGF-----------GELWTRQKFLGLQLKVVRREN--EKVYVPNLPFEDWLTMWVTPRS : 546
EeV   : --LLFGDLELLKEKKAIKFFKEKHGLVIKEGTWNPTAVTEDPGY-----------DELWTRQKFLGLQIKVVKHADTDEKVFVPTLPYSDWLSMWVTPRA : 548
                          -----fC----                                        ------E------

                 *        420         *
IBDV  : ---------VENKSLKSKV-GIEQAYKVVRYEALRLVGG : 631
IPNVJ : ---------LENKALARKP-GAEIAYQIVRYEAIRLVGG : 621
TaV   : KYRSKETETMRERTLFDRARGLLVTGAVFDERARGLMGA : 585
EeV   : RLGHKESYTMRQRTLFDRARGLLVTGAAFDKQARELMGS : 587
```

Figure 3. The sequence conservation in replicases of TaV/EeV and IBDV/IPNVJ. ClustalX-generated alignments of replicases of two pairs of viruses, TaV/EeV and IBDV/IPNVJ, were used to produce an across-virus-families alignment that included a match identified using psi-Blast. Positions of selected conserved RdRp blocks are marked. C?, putative permuted C motif; fC, fortuitous canonical C motif.[27,28] (a) A plot of the conservation along the alignment of four viruses. The plot was produced using the Plotsimilarity of the GCG package with a window of 10 aa residues and the Blosum62 scoring table. The level of average similarity is marked with a broken line. Top bar, a portion of the alignment that is presented in (b). (b) The most conserved part of the alignment of four viruses. Residue position in each replicase is shown at the right side. Red, green and yellow backgrounds highlight columns with 100% identity, not less then 75% conserved residues, 75% identity, respectively. Groups of conserved residues are: N, D, Q, E; K, R, H; F, Y, W; A, C, I, L, V, M; S, T.

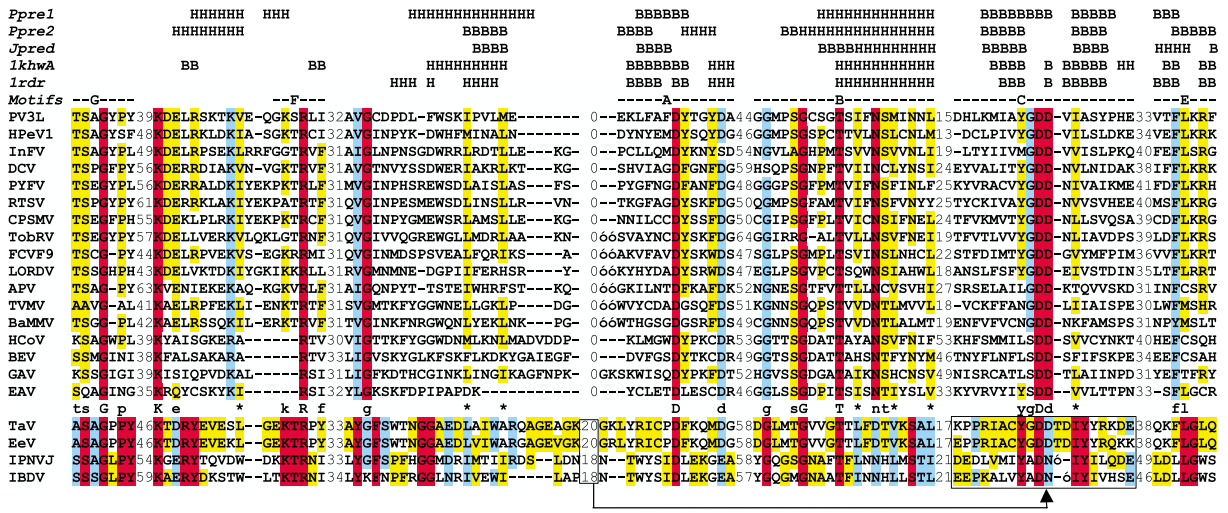


Figure 4. Sequence alignment of RdRps of selected RNA viruses employing the canonical and permuted palm subdomains. The RdRps of TaV/EeV/birnaviruses were converted into the canonical form by relocating the motif C? sequence (18–20 aa; boxed) downstream of the motif B (arrow). Excerpts from an alignment of the canonical 58RdRps comprising RdRps of 58 Picornavirus-like viruses and Nidoviruses that were proved to be among those that are most similar to the replicases of TaV, EeV and birnaviruses (17 viruses, top set) and the quasi-canonical RdRps (four viruses, bottom set) are presented. Red, blue and yellow backgrounds highlight columns with 100% identity, 75% identity or 100% conserved residues, 50% identity or 75% conserved residues, respectively, for the two sets separately. Groups of conserved residues are: N, D, Q, E; K, R, H; F, Y, W; I, L, V, M; A, S, T. Residues most conserved in two sets of viruses are featured in the line separating the two sets. Upper and lowercase residues, absolutely and partly conserved residues, respectively; ∗ , I, L, V and M. The positions of motifs are shown. The intermotif distances are given between a pair of respective motifs, except for the distances between motifs B and C, and C and E of the bottom group, which are the distances separating the insertion position of the motif C from motifs B and E, respectively. Top five lines highlight residues forming β-strands (B) and α-helices (H) in the tertiary structures of RdRps from the calicivirus RHDV (1khwA; A chain) and the picornavirus PV type 1 (1rdr), or predicted secondary structure elements by the Jpred for alignment of RdRps of TaV, EeV, IPNVJ and IBDV, or psi-Pred for the IPNVJ RdRp (Ppre1) and TaV RdRp (Ppre2). Virus families and groups, viruses, and the NCBI protein (unless other specified) IDs are listed below. Picornaviridae, human poliovirus type 3 Leon strain (PV3L, 130503) and parechovirus 1 (HPeV1, 6174922); Unclassified insect viruses, infectious flacherie virus (InFV, 3025415) and *Acyrthosiphon pisum* virus (APV, 7520835); "CrPV-like" group, *Drosophila* C virus (DCV, 2388673); Sequiviridae, rice tungro spherical virus (RTSV, 9627951) and parsnip yellow fleck virus (PYFV, 464431); Comoviridae, cowpea severe mosaic virus (CPSMV, 549316) and tobacco ringspot virus (TobRV, 1255221); Caliciviridae, feline calicivirus F9 (FCVF9, 130538) and Lordsdale virus (LORDV, 1709710); Potyviridae, tobacco vein mottling virus (TVMV, 8247947) and Barley mild mosaic virus (BaMMV, 1905770); Coronaviridae, human coronavirus 229E (HCoV, 12175747) and Berne torovirus (BEV, 94017); Arteriviridae, equine arteritis virus (EAV, 14583262); Roniviridae, gill-associated virus (GAV, 9082018); putative Tetraviridae, TaV (AF82930; nt sequence) and EeV (AF461742; nt sequence); Birnaviridae, IPNVJ (133634) and IBDV (4894793). Corona-, Arteri- and Roniviridae belong to the order *Nidovirales*.[77,78]

and NβV) and mammalian viruses of another family (hepatitis E viruses) (Figures 1 and 2). Furthermore, when our analysis was extended to database searches, the only statistically significant hit (psi-Blast, Blosum62, no filter, $E = 0.004$) was recorded between the N-terminal ~330 aa regions of the putative replicase of TaV and the previously identified RdRp domain of the 845 aa replicase of a dsRNA birnavirus, IPNV.[28] This hit was expanded through iterative searches and converted into an alignment between the replicases of TaV and EeV and two birnaviruses, IPNV and IBDV, that contained conserved regions of 530–580 aa residues adjacent to the N terminus of the proteins (Figure 3(a) and (b) and data not shown). Using profile HMMER2.1-mediated searches,[33] this region in the four viruses was shown to be similar to RdRps of ssRNA + viruses of the Picornavirus-like supergroup[18] and *Nidovirales*[34] whose sequence affinity was already documented[35] (all top hits in the Genpeptides database with scores better than $E = 100$ were (putative) RdRps) (Figure 4). Accordingly, we concluded that the identified region of the TaV and EeV replicases might include a RdRp.

The conserved active site motifs associated with the palm subdomain are permuted in the (putative) RdRps of TaV, EeV and birnaviruses

Inspection of the TaV/EeV/birnavirus replicase alignment revealed the conserved variants of several sequence elements including the characteristic RdRp palm subdomain motifs A ($DX_{4-5}D$) and B ($GX_{2-3}TX_3N$), and two other, less prominent motifs, F ($RX_{1-2}I/L$)[7] and E (no consensus). The assignment of these motifs is also supported by comparative analysis of secondary structure elements predicted for RdRps of TaV, EeV, and

birnaviruses and resolved for RdRps of a calicivirus, rabbit hemorrhagic disease virus (RHDV),[9] and a picornavirus, poliovirus (PV)[6] (Figure 4). The analyzed RdRps also contain a newly recognized motif, termed G ($T/SX_{1-2}G$) (Figures 3 and 4), that is the most conserved sequence in RdRps of TaV, EeV, and birnaviruses (Figure 3(a)). In the RHDV RdRp, the G motif occupies a part of the finger subdomain and is flanked by two Lys residues (Lys114 and Lys134) that were predicted to interact with the phosphodiester backbone of the primer in the primer-template duplex.[9] One or two conserved basic residues can also be found in the vicinity of the G motif of other viruses listed in Figure 4 (data not shown). Thus, the invariant Gly and highly conserved Pro residues prominent in the G motif may have been selected to enforce the correct orientation of the adjacent basic residue(s) relative to the primer.

However, and consistent with a previous observation on the birnavirus RdRps,[28] the key catalytic motif comprising two aspartate amino acid residues flanked by two stretches of hydrophobic residues (motif C), proved to be lacking in the canonical positions in the putative RdRps of TaV and EeV. Accordingly, this region was termed fortuitous C motif (fC; Figure 3(a) and (b)). Motif D (no consensus) was similarly not found. Surprisingly, a block with the expected properties for motif C is present immediately upstream of motif A in the replicases of TaV/EeV/birnaviruses (C? in Figure 3 and C in Figure 4). It includes a GDD (TaV and EeV) or structurally similar ADN tripeptide (infectious pancreatic necrosis virus strain Jasper (IPNVJ) and IBDV), and might therefore be the authentic motif C occupying a non-canonical position in the sequence of these RdRps. This motif forms an extra block compared to the RdRps of Picornavirus-like viruses and Nidoviruses (column that includes boxed numbers in Figure 4).

If motif C? in these unusual RdRps is the functional motif C required for replicase activity, it could have been relocated without compromising the associated RdRp activity, as has been previously observed for characterized circularly permuted proteins.[36] We reasoned that such an internal sequence rearrangement or permutation, which is unprecedented for the TDPPs, might be corroborated in a truly objective manner. To verify this permutation, we have applied a specially designed computer-assisted protocol that utilizes capabilities of sensitive HMMER and rps-BLAST programs for analyzing artificially permuted sequences. Using this protocol, it was shown that the relocation of the motif C? into the canonical position specifically and selectively converts non-canonical RdRps of the TaV/EeV and IPNV/IBDV lineages into the quasi-canonical RdRps (Figure 4; for details see Materials and Methods and Figures 8 and 9). The latter are indistinguishable from the real biological sequences that are distantly related to RdRps of Picornavirus-like viruses (Pfam database accession number PF00608[37]).

Collectively, the above observations independently inferred a non-canonical C–A–B motif arrangement for replicases of each of the TaV/EeV and, IPNVJ/IBDV lineages, thus confirming their special clustering (Figure 3; see also below).

The permuted sequences of the RdRps of TaV/EeV and birnaviruses are compatible with the palm subdomain architecture

Circularly permuted proteins are known to maintain folds of their unpermuted homologs.[36] Is the internally permuted sequence organization of RdRps from TaV/EeV/birnaviruses compatible with the canonical palm fold? To address this question, the connectivity of the tertiary structure of the PV RdRp,[6] a typical palm-based polymerase belonging to the PF00608 family, was modified to model the permuted C–A–B motif sequence arrangement. To relocate motif C upstream of motif A, the PV structure had to be cut in three loops between the following pairs of elements: αD and β1, αH and β2, β3 and αI. For the permuted structure, three new connections between the following pairs: αD and β2, β3 and β1, and αH and αI, respectively, had to be formed (Figure 5(a) and (b)). All connections affected by this permutation are confined to a restricted loop area opposite the active site where the conserved catalytic aspartate residues (D233 and D328) are positioned (Figure 5(b)). Three new connections could be modeled without major steric clashes, and the Cα–Cα distances between the termini of the major secondary structural elements in the actual and artificially permuted structures, 9.58–14.27 Å and 4.26–10.40 Å, respectively, are in similar ranges (Figure 5(b) and (c)). Thus, the permuted backbone connectivity is compatible with the spatial organization of the major secondary structure elements of the palm fold and could maintain structural integrity of the subdomain, as was observed for circularly permuted proteins.[36,38]

This rearrangement transforms the ferredoxin-like β1αHβ2β3αIβ4 fold of the palm subdomain, containing an insertion between β1 and αH elements, into a new β2β3β1αHαIβ4 structure (Figure 5(a)). In this structure, the antiparallel β-sheet of the original fold is partly freed from the covalent linkage to the αH and αI elements, which, in turn, become directly covalently linked, and the β2^β3 hairpin and β1 strand, carrying the principal catalytic Asp residues, are brought into intimate sequential proximity. Similar structural alterations are predicted for the naturally permuted RdRps of TaV, EeV and birnaviruses. They may result in unique structural properties (e.g. intra- and inter-domain mobility) of the palm subdomain that may affect, for instance, the interconversion of open (inactive) and closed (active) conformations of the RdRp active site[9] and, ultimately, functioning of these RdRps.

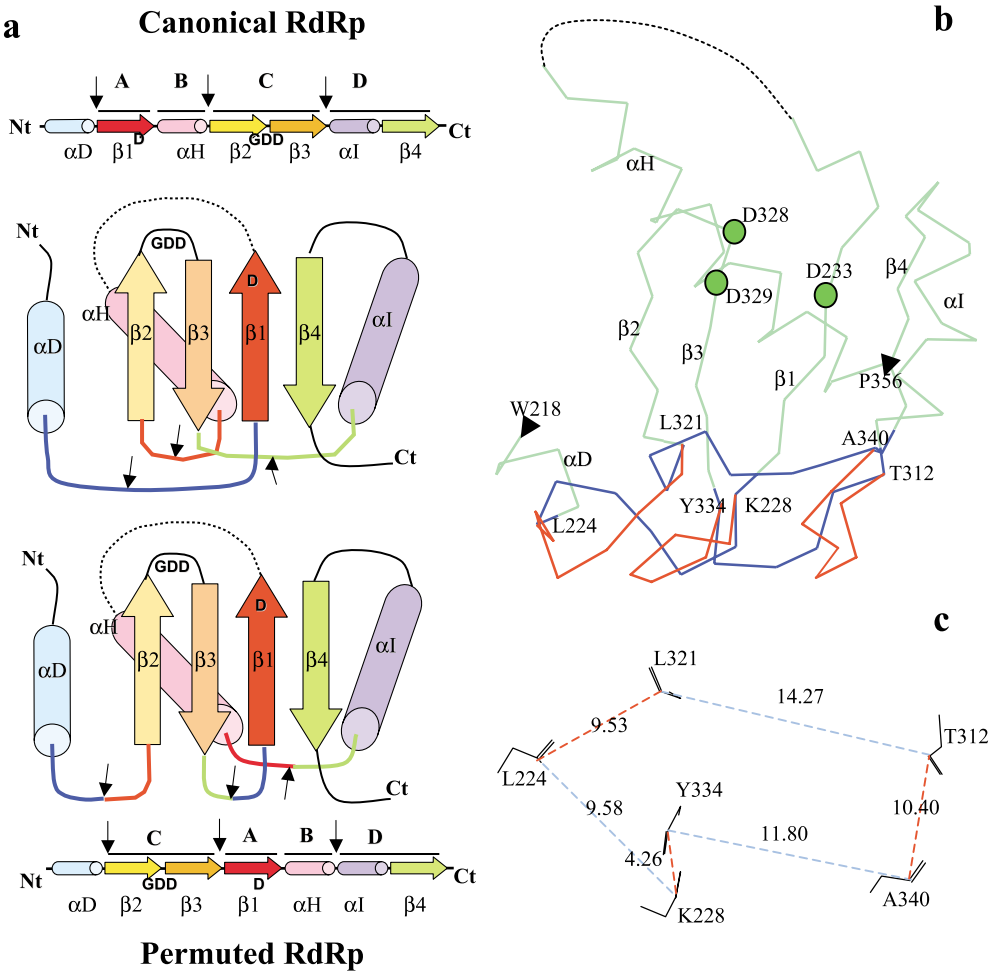


Figure 5. Structural organization of the canonical and permuted palm subdomains. The color schemes in (a), and in (b) and (c) are different. (a) Linear and tertiary organizations of sequence motifs and the major secondary structure elements in the canonical (top) and permuted (bottom) palm folds. Secondary structure elements are labeled according to the PV RdRp structure 1rdr[6] and, along with the A–D sequence motifs, shown schematically. Black arrows, positions where the backbone connectivity is broken in the canonical RdRp and reconnected in the permuted RdRp. For sake of clarity, all breaks are introduced in the middle of the loops and elements between A and B motifs omitted. Nt and Ct, N and C terminus, respectively. (b) Permutation of the palm fold of the PV RdRp. Using the Modeller suit of the Insight II package, the permutation found in the TaV/EeV/birnavirus RdRps was modeled onto the PV RdRp by relocation of the H320-E337 18 aa peptide into between E226 and E227 residues. To accommodate this re-organization, three additional mutations were introduced using loops from other proteins as templates: the foreign KVD tripeptide was inserted upstream of the 18 aa peptide and two point mutations P335 → G and E337 → P were engineered. The modeled regions were then improved using the Whatif4.99 package. The structure connecting β1 and αH contains three α-helices and the unresolved 268–290 aa region that are depicted with a broken line. The Cα track of the palm subdomain of the PV RdRp (from W218 to P356, capped by arrows) is shown in green with loops to be affected by the permutation colored in blue. The modeled permuted loops are shown in red. Green dots, active site Asp residues of motif A (D233) and motif C (D328 and D329). Also indicated are the residues at the termini of secondary structure elements that are connected by loops to be permuted. (c) Distances (in Å) between the terminal residues of the affected secondary structure elements in the actual structure of PV (blue) and in the permuted derivative (red).

The permuted RdRps form a separate lineage in a polymerase tree

To ask whether the presence of the non-canonical C–A–B motifs order in the the TaV/EeV and birnavirus lineages might be due to a single ancestral permutation, a phylogenetic analysis was conducted using an alignment of the quasi-canonical RdRps of TaV/EeV/birnaviruses and a representative set of the canonical RdRps from the 58RdRp list (Figure 4). It was found that the quasi-canonical RdRps comprise a separate, deeply rooted lineage supported by 953 out of 1000 and 78 out of 100 bootstrap trials in the neighbor-joining and parsimonious analyses, respectively (Figure 6 and data not shown). The RdRps of TaV/EeV and birnaviruses form a distinct cluster because of the sequence conservation over a long region rather than the presence of the unique motif C permutation that was reversed before the phylogenetic

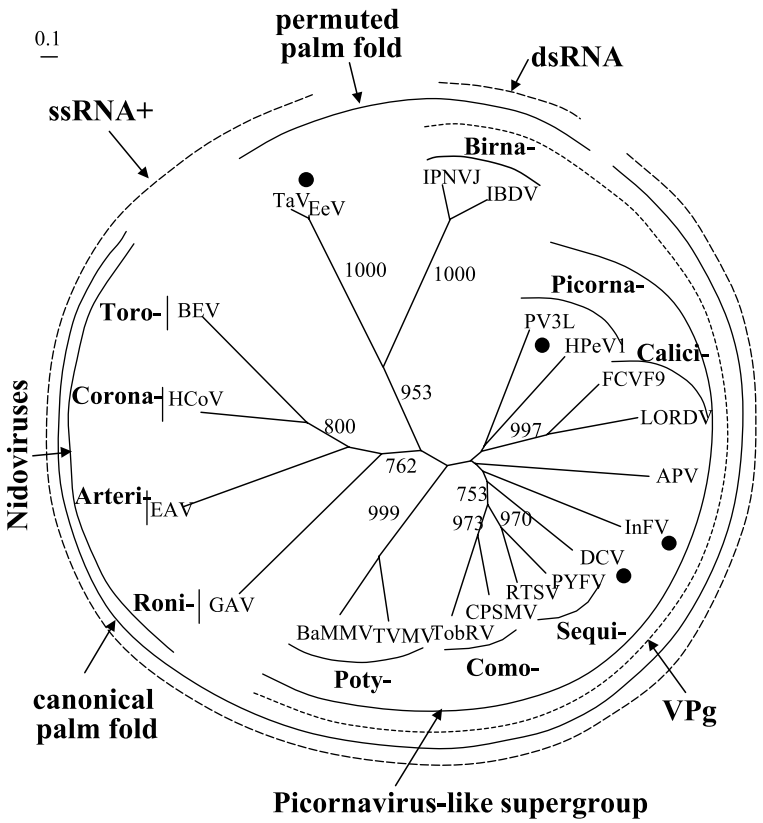


Figure 6. Phylogenetic analysis of a selected set of RdRps. Using an extended, gap-free version of the Figure 3 alignment, an unrooted neighbor-joining tree was inferred by the ClustalX1.81 program. All bifurcations with support in >700 out of 1000 bootstraps are indicated. A similar tree topology was inferred from analysis of the 328 parsimonious informative characters of the alignment using an heuristic search and parsimonious criterion (not shown). Different groups of viruses are highlighted. Only selected picornaviruses that do not include PV and HPeV1 employ a 2A protein (●) of the NPGP family. For the virus names, see the legend to Figure 4.

analysis. This conservation was originally uncovered in the course of the database searches (see above) and is evident in the alignment shown in Figure 3(b). The actual distance between RdRps of TaV/EeV/birnaviruses and those of other viruses must be even greater than that depicted in this tree, given that the motif C relocation in these quasi-canonical RdRps (Figure 4) has artificially increased the genuine similarity between enzymes of TaV/EeV/birnaviruses and those of other viruses. However, the precise evolutionary weight of the motif permutation remains unknown.

The large distance between the permuted and canonical RdRps is also evident in the active site replacements in the permuted RdRps, which are not observed elsewhere in ssRNA + viruses. Thus, birnaviruses have accepted Asp-to-Glu and Asp-to-Asn mutations of the second Asp in motifs A ($DX_{4-5}D$) and C (GDD), respectively, and TaV/EeV have an accepted Asn-to-Asp mutation in motif B ($GX_{2-3}TX_3N$) (Figure 4; see also O'Reilly & Kao[11]). Some of these substitutions were shown to be compatible with the RdRp activity of the PV enzyme[15] and one of them, GDD-to-GDN, resulted in a change of metal specificity.[39]

Discussion

The palm-based polymerases form the major family of the TDPPs and are universally used in all kingdoms of life. By fixation of mutations at selected positions in the palm subdomain active site motifs and elsewhere, four types of palm-based polymerases, RdRp, RdDp, DpDp and DdRp, could have evolved from the common ancestor which likely inhabited the RNA–protein world.[6] We demonstrate here, for the first time, that within the RdRps there occurred a bifurcation involving an otherwise unique permutation of the palm sequence motifs that yielded a new RdRp lineage. This permutation must be compatible with the RdRp activity, since birnaviruses, encoding a permuted RdRp, are known to be replicatively competent.[29,30] The transfer of TaV and EeV between insect hosts and the presence of replicase genes in the virus genomes suggest that the yet-to-be-characterized TaV and EeV are also non-defective.

Discovery of internally permuted replicases in the TaV/EeV and IPNV/IBDV lineages

A fraction of known proteins have been shown to have evolved from ancestors by migration of the N and C termini into a loop region (circular permutation). Proteins with these rearrangements have been identified by either the analysis of tertiary structures or bioinformatics analysis of the canonical and permuted homologs.[36,40,41] The latter approach involves sequence comparisons and protein modeling and was employed here. Like other studies concerned with the bioinformatics identification of permutations,[42] we showed that: (i) the reversion of the identified permutation significantly and selectively increases similarity of the

affected sequence with other canonical homologs; and (ii) a canonical architecture could accommodate the sequence permutation through changes in the backbone connectivity in loop regions.

The permuted replicative proteins described here are different in two respects from circular permutants described elsewhere.[36,40–42] They have evolved through permutation of extremely small structures (~22 aa) with upstream structures (~110 aa) in large replicative proteins (~850–1200 aa). This domain reshuffling involved changes of the backbone connectivity in three loops rather than one loop and two terminal regions.[40–42] A complex protocol of sequence comparisons was introduced (see Materials and Methods) to uncover these unprecedented permutations, which must be self-evident in the tertiary structures, which are not currently available. Due to a large evolutionary distance between the permuted and canonical RdRps, the three positions that delimit two adjacent permuted subsequences in each replicase have been identified with small margins which are expected to decrease when more related replicases could be analyzed. Work is in progress (B.T.L. & A.E.G., unpublished results) to extend our approach to the identification of internal permutations in other structurally uncharacterized proteins in sequence databases that should clarify the extent of the contribution of this type of permutation to protein evolution.

RNA viruses with canonical and permuted replicases

This study was initiated to gain insight into the replicase of TaV by its sequencing and bioinformatics analysis. Contrary to the current capsid-based classification of TaV,[23] and the closely related EeV, within the Tetraviridae family, the replicases of TaV and EeV proved to be significantly different in the domain organization and overall similarity from those of the known tetraviruses HaSV and NβV (Figures 1 and 2). These two groups of viruses therefore employ very different replicative machineries that have diverged relatively early in evolution. With their shared capsid architecture and divergent replicases, TaV/EeV and the well-established tetraviruses enjoy a mosaic relationship (Figure 1) resembling that between Picornavirus-like potyviruses and Alphavirus-like potexviruses.[43] On the basis of this parallel and the phylogeny of RdRps (Figure 6), we suggest to re-examine the taxonomic position of TaV and EeV with regard to the Tetraviridae. These viruses may be prototypes for a new family distinct from tetraviruses and not belonging to any existing virus supergroup. A formal proposal to make these changes is to be submitted to the International Virus Taxonomy Committee (T.N.H., K.H.J.G. *et al*., unpublished results).

The discovery of a distantly related group of permuted replicases in birnaviruses was also very surprising, since other replicases from true ssRNA + and dsRNA viruses are not interleaved.[19] The clustering of TaV/EeV and birnaviruses indicates that these viruses may share important characteristics of RNA synthesis not common to their respective classmates that are involved in contrasting, i.e. virion-independent and dependent, respectively, modes of replication.[44] The unique, intermediate position of birnaviruses between other dsRNA and ssRNA+ viruses was also evident with other virus properties.[45]

The observed replicase conservation covers approximately 550 aa and includes the RdRp domain, which is flanked by other uncharacterized domains. It may be relevant to this conservation that the genomic RNAs of viruses of these two groups have unique 3′-ends[24,23] which do not have the poly(A) or tRNA-like structures common in many other ssRNA + and dsRNA viruses.[46] The 3′-end is crucial for the initiation of minus RNA synthesis in RNA viruses.[47] In birnaviruses, a fraction of replicase molecules are known to be covalently linked to the 5′-end of genomic RNAs; these molecules are likely to be used to prime RNA synthesis.[24] The 5′-ends of the TaV/EeV genomic RNAs, which remain to be characterized, might also have a similar structure.

The CDs in replicases of TaV, EeV and birnaviruses have remote similarities to the RdRps of Nidoviruses[34] and Picornavirus-like viruses[22] that are not limited to the four palm motifs and include one new G motif. These two large virus supergroups comprise approximately one fourth of about 50 currently known RNA virus families and groups.[48] Although TaV, EeV, and birnaviruses do not have sequence characteristics that would classify them with either Nidoviruses or Picornavirus-like viruses, the observed clustering of their RdRps is correlated with other similarities. Protein priming of RNA synthesis with a special viral protein (VPg) was originally discovered in Picornaviruses,[49] and all Picornavirus-like viruses, as well as birnaviruses, may use this mechanism[24,46] (VPg curve in Figure 6). Some viruses from the Picornavirus-like supergroup and TaV/EeV employ a 2A or 2A-like protein of the NPGP family for proteolytic autoprocessing (dots in Figure 6).[23,26,50] These correlations are yet to be rationalized in structure–function terms.

Viruses with a permuted palm fold form a minor lineage that includes approximately 4% of known RNA virus families/groups. The deep rooting of the permuted RdRps branch in the RdRp tree and the striking genome diversity of the few known viruses of this branch both indicate that these viruses are being significantly underrepresented. The identification of new viruses employing the permuted RdRps should be assisted by the results reported here. (After this manuscript was prepared, we found that the motif C permutation is also conserved in a newly sequenced replicase of *Drosophila* X virus, an insect birnavirus[51] (unpublished observation).

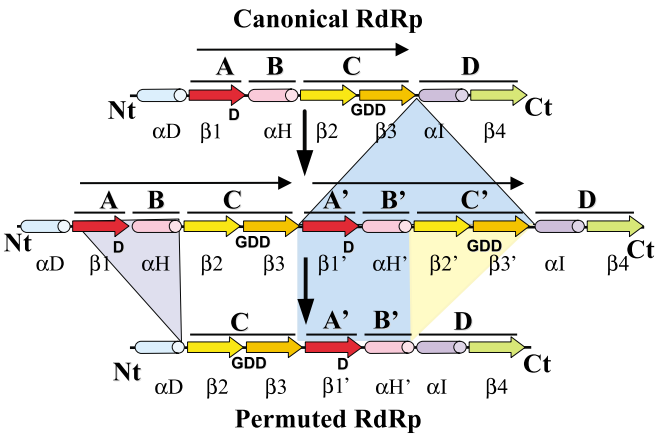


Figure 7. A tentative scenario for evolution of the permuted RdRps of TaV/EeV and birnaviruses. For labels, see the legend to Figure 5.

The emergence of permuted RdRps

To derive a permuted motif organization from the canonical one, a tandem duplication of motifs A, B and C with a subsequent deletion of the original motifs A and B and the duplicated motif C′ must have taken place (Figure 7). A reverse scenario is equally possible. Genetic rearrangements of this or greater complexity have been observed in evolution of contemporary RNA viruses (e.g. pestiviruses) and are linked to the high rate of RNA virus recombination.[52] Although these observations indicate that there would seem to be no mechanistic barriers to permutation of the palm subdomain occurring in different lineages at different time points, all the palm permutations that we have identified in the present study are likely to be descendants of a common ancestral permutation fixed early in the evolution of RNA viruses.

Five important characteristics of the palm subdomain permutation — involvement of the *catalytic* core of the *ancient* domain encoded by *diverged RNA* viruses of a *distinct* lineage, all indicate that the structural diversification of the palm subdomain may have happened at the primitive stage of evolution of the enzyme. Compared to its permuted relative, the canonical organization of the palm subdomain is in overwhelming dominance among contemporary TDPPs of DNA and RNA origin. This pattern of the fold utilization suggests that the canonical organization may have originated in the RNA–protein world from the permuted ancestor and was later selected as the basis for the DNA-involved TDPPs. Further comparative characterization of RdRps of these two palm folds may give unique insight into the major forces that determined the profound disparity in the utilization of the two folds among organisms and identify a key property that directed the early bifurcation of the palm fold evolution.

Do permuted RdRps link different folds?

Protein permutations commonly involve structural rearrangements that preserve the fold type. In naturally evolved and artificially engineered circularly permuted proteins, the N and C termini migrate from the original to new positions between either α/α, or β/β or α/β units.[36,41] The results of this study show that the protein folding and function can also sustain an internal permutation that changes a fold type. These observations indicate that divergent evolution has contributed to the origin of the structurally diverse folds. Particularly, it might have generated variants of the ferredoxin-like fold that were identified in numerous proteins with a variety of functions, no significant sequence similarity and several backbone connectivities.[6,53,54]

In the light of our observations, the intriguing question emerges whether the permuted or other deviant structural form of the palm subdomain could have evolved further to give rise to the palm subdomain of structurally different eukaryotic DNA polymerase β, which employs a nucleotidyltransferase-like fold.[8,55] Studies of the permuted RdRps might also be useful for understanding the relationship between palm-based RdRps and those involved in RNA silencing,[56] enzymes that are currently considered unrelated.

Materials and Methods

Cloning and sequencing of the TaV genome

TaV was purified from frozen infected *Setothosea asigna* larvae supplied by Dr Bernhard Zelazny, Integrated Coconut Pest Control Project, Jakarta, Indonesia. Virus purification and RNA extraction were as described.[23] A TaV cDNA library was prepared and a 2200 nt clone containing a portion of the TaV RdRp was isolated previously.[23] Here, the plasmid library was screened by colony blotting using the original clone as a probe, and by PCR of the clone library to isolate the remainder of the replicase gene using RdRp-specific primers and universal forward or reverse primers. All nt sequences were confirmed on two separate clones or by sequencing of RT-PCR products derived from viral genomic RNA.

General bioinformatics analyses

Genpeptides, CD[57] and protein family (Pfam)[37] databases were used here. Amino acid (aa) sequence alignments were generated using ClustalX1.81[58] and Dialign2[59] programs assisted by Blosum position-specific matrices,[60] and were processed for presentation using GeneDoc.[61] An alignment of the RdRps from 58 viruses, representing 13 ssRNA + virus families and groups of the Picornavirus-like supergroup and *Nidovirales*, was termed 58RdRp. Protein alignments were sent as input for the Jpred server to generate consensus prediction of secondary structures over several methods.[62,63] Secondary structures were also predicted using a single sequence as input for the PSIPRED server.[64,65] Multiple sequence alignments were converted into Hidden Markov Model (HMM) profiles using HMMER2.01 software[33] or used to build profiles using the Profileweight program.[66] Sequence databases were searched in default mode, unless otherwise stated, using the HMMER2.01 package[33,37] and a family of Blast programs.[67] The expectation values of similarity (E) of 0.05 or lower for Blast searches and 0.1 or lower for HMMER-mediated searches were considered to be statistically significant.[68] The Profileweight profiles were compared in pairs by sliding a window of a selected length along each possible register, and matches above a threshold were recorded using the Proplot program.[66] The Plotsimilarity routine of the GCG-Wisconsin package (Genetics Computer Group, Madison, USA) was used to visualize the conservation in sequence alignments. Cluster phylogenetic trees were reconstructed using the neighbor-joining algorithm of Saitou & Nei[69] with the Kimura correction[70] and were evaluated with 1000 bootstrap trials, as implemented in the ClustalX1.81 program. Parsimonious trees were generated using heuristic search and evaluated with bootstrap analysis using a UNIX version of the PAUP* 4.0.0d55 program[71] that is included in the GCG–Wisconsin Package programs. The resulting trees were visualized using the TreeView program.[72] Protein modeling and structure visualization were performed using Insight II (Accelrys Inc.) and Whatif4.99 packages.[73]

Computational analysis of sequence permutations: approach and application to TaV/EeV and IPNVJ/IBDV

To identify and validate sequence permutations, a multi-step protocol was introduced that is briefly described below along with results of its application to replicases of RNA viruses. The identification of a genuine sequence permutation is straightforward, provided an analyzed sequence returns non-linear, permuted matches with other, canonical homologs upon scanning a sequence database using Blast or other search engine.[42] None of these conditions were apparent upon analysis of replicases of TaV, EeV and birnaviruses that differ from the canonical homologs through a permutation of a short internal sub-sequence and profound divergence elsewhere. To meet the challenge of identifying permutations of this complexity, we decided to analyze large spaces of the computer-generated replicase permutants using HMMER2.01[33] and rps-BLAST-mediated[67] database searches. We made use of an observation that alignment with a highest score between permuted and canonical homologs is produced when permutation is reversed.[40,42,74] In other words, if two protein families have diverged through permutation of a sub-sequence in the ancestor of one of two families, then back-permutation in the proteins of the affected family produces sequences that outscore the parental sequences upon comparison with the other protein family. It is reasonable to assume further that this back-permutation must also outscore any other possible permutations as they, at the best, can only approach the similarity between the back-permutant and the other protein family.

Technically, the back-permutation is equivalent to a permutation of the parental, permuted sequence. To denote a particular permutation, three cut-points (I, J, and L) need to be chosen, where each index represents the position before the residue. For example, if $I = 5$, the first cut-point lies between residues 4 and 5. If I and J represent the beginning and end of the region being moved and L represents the position where this region is inserted (so residues I through $J - 1$ are placed between residues $L - 1$ and L), and if:

$$1 \leq I < J < L \leq N + 1$$

where N is length of a parental sequence, then relocating the I–J region to L is identical to relocating J–L to I. Values of three indexes vary in the following ranges: for I from 1 to $N - 1$, for J from $I + 1$ to N, and for L from $J + 1$ to $N + 1$, for each index with a stride of S. Three indexes can be ordered by $3! = 6$ ways to yield the same permutation. This means that the number of all possible permuted sequences (permutants) derived from the parental sequence ($S = 1$) is equal to $(N + 1)N(N - 1)/6 = (N^3 - N)/6$. Since the replicases of IPNVJ/IBDV and TaV/EeV contain from 845 aa to 1257 aa, the number of permutants that can be generated is on the order of 10^8. This number is approximately two orders of magnitude larger than the number of sequences in the current version of the National Center for Biotechnology Information (NCBI) non-redundant protein database. To routinely manage the databases of this scale, extensive computational resources would be required (see also below).

To reduce the computational requirements of this search over permuted sequences, a two-step procedure of the permutant database generation was employed. In the first step, the possible values of I, J, and L were chosen with a non-unit stride S rather than with the $S = 1$ as when a complete permutant database is generated. This reduces the number of permuted sequences that are generated by a factor of approximately S^3. The stride length, S, should be odd so that unique sequences could be easily generated in the second step (see below), and, in practice, a stride length of 9 aa that is significantly smaller than sizes of expected permutations was used. Using this stride, from 2.6×10^5 to 6.9×10^5 shuffled sequences were generated from replicases of TaV/EeV and IPNVJ/IBDV.

To offset differences in the sizes of the databases of permutants used here and, thus, make direct comparisons between results of different HMMER2.01-mediated database scans possible, the database sizes were set equal (10^5). Each 9 aa stride database was searched with the 58RdRp HMM. The ratio of the HMMER E-values for the original (E_o) *versus* shuffled (E_s) sequences was used to rank the permutations in descending order. Though many permutations resulted in the same E_s value, a plot of the average value of E_o/E_s over the best K permutations (ordinate) *versus* K values (abscissa) had the general appearance of a decaying exponential (not shown). Using a central difference method, the place where the slope of this curve stayed below 0.1 for 100

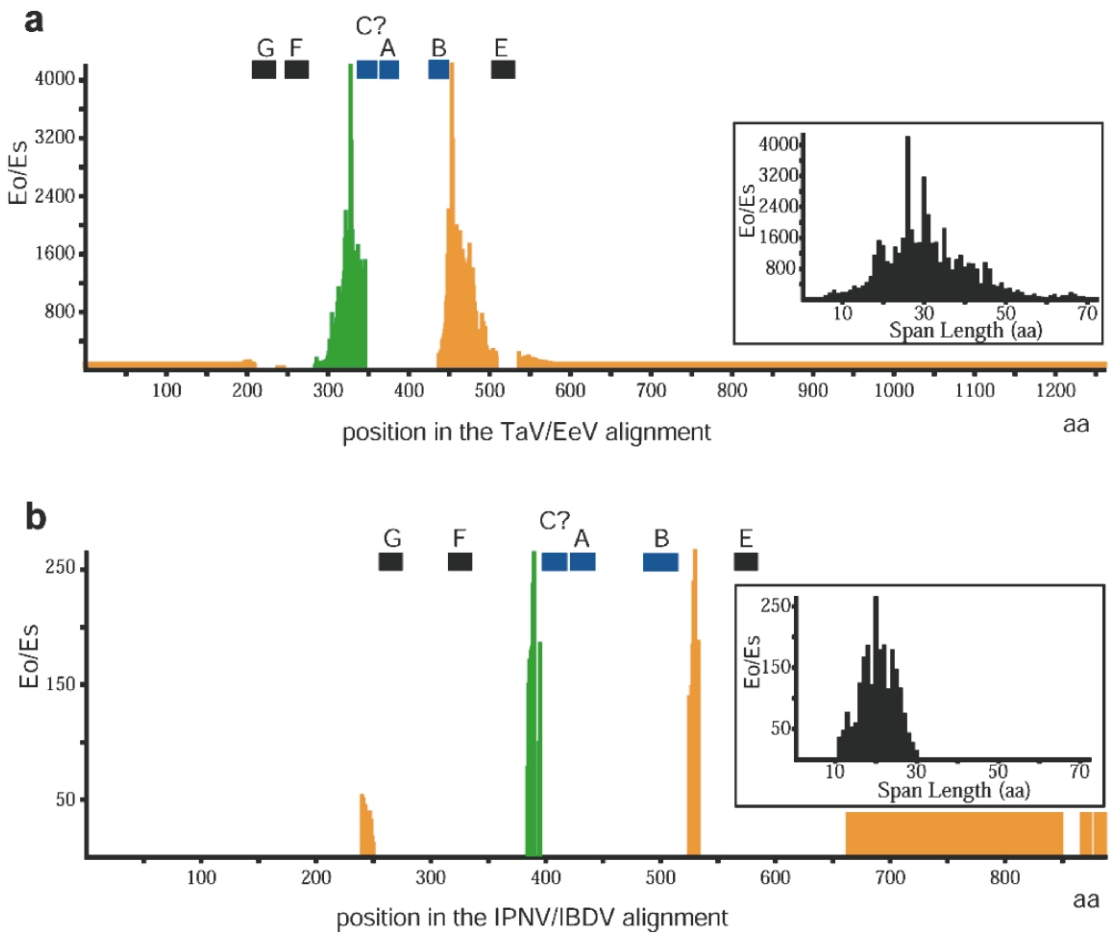


Figure 8. Distributions of three sequence characteristics of artificially permuted replicases of tetraviruses and birnaviruses. The 1 aa stride databases of artificially permuted replicases of TaV, EeV, IPNVJ and IBDV were scanned with the 58RdRp HMM and those high-scoring permutants that involved relocation of homologous sub-sequences in pairs TaV and EeV (a), and IPNVJ and IBDV (b), respectively, were identified. Distributions of the sum of E_o/E_s scores for each pair of viruses that are associated with the positions of the origin (I; green) and destination (L; ochre) of permutations in respective alignments, and the highest scores for all sizes of relocated sub-sequences (I–J; insets) were plotted. Positions of six conserved sequence motifs are indicated.

contiguous K-values (say K_o) was located. All sequences that ranked higher than K_o were considered top-scoring permutants and were selected for subsequent analysis.

In the second step of the permutant database generation, each of the permutations chosen above were taken to represent seed points about which a more detailed analysis was performed. If a selected permutation was represented by the triad (I', J', L'), then I was varied from $I'-4$ to $I'+4$, J from $J'-4$ to $J'+4$ and L from $L'-4$ to $L'+4$ in strides of 1 aa, subject to the inequalities given above. This procedure produced a set of up to 729 permutations around each of the K_o permutations selected above to generate a new database of at most $729K_o$ permutants. The sizes of these databases for analyzed replicases were on the same order as the sizes of the original 9 aa stride databases. These 1 aa stride databases were searched with the 58RdRp HMM.

To check the validity of the above two-step procedure, the IPNVJ replicase was also examined using a 1 aa stride, all-inclusive permutant database whose size was about 28GB. The top-scoring permutant identified through the searching of this database with the 58RdRp HMM was the same motif C permutant that was found with the two-step procedure (not shown).

At the final stage, some biologically irrelevant permutants were removed through selection of only those high-scoring permutants that were generated by relocation of homologous sub-sequences in pairs of related sequences: TaV/EeV and IPNVJ/IBDV, respectively. We considered these alignment-filtered permutations evolutionarily conserved, namely, that each pair of such permutations may have descended from a permutation fixed in a common ancestor of the virus pair. The E_o/E_s values of each pair of evolutionarily conserved permutations of two viruses were summed, ranked and plotted. Among thousands of shuffled sequences that outscored the parental sequences during the database scans, 63,419 and 9147 involved relocation of homologous regions of proteins in the TaV/EeV (Figure 8(a)) and IPNVJ/IBDV (Figure 8(b)) pairs, respectively. All top-scoring sequences contained permutation of a 20–30 aa stretch (insets in Figure 8(a) and (b)) that either overlapped or encompassed motif C? (green graphs in Figure 8(a) and (b)). This motif was relocated into a region normally occupied by motif C (red graphs in Figure 8(a) and (b)). Furthermore, no other large peaks, which could be linked to the relocation of other sequences (e.g. motif D), were evident in Figure 8(a) and (b), indicating

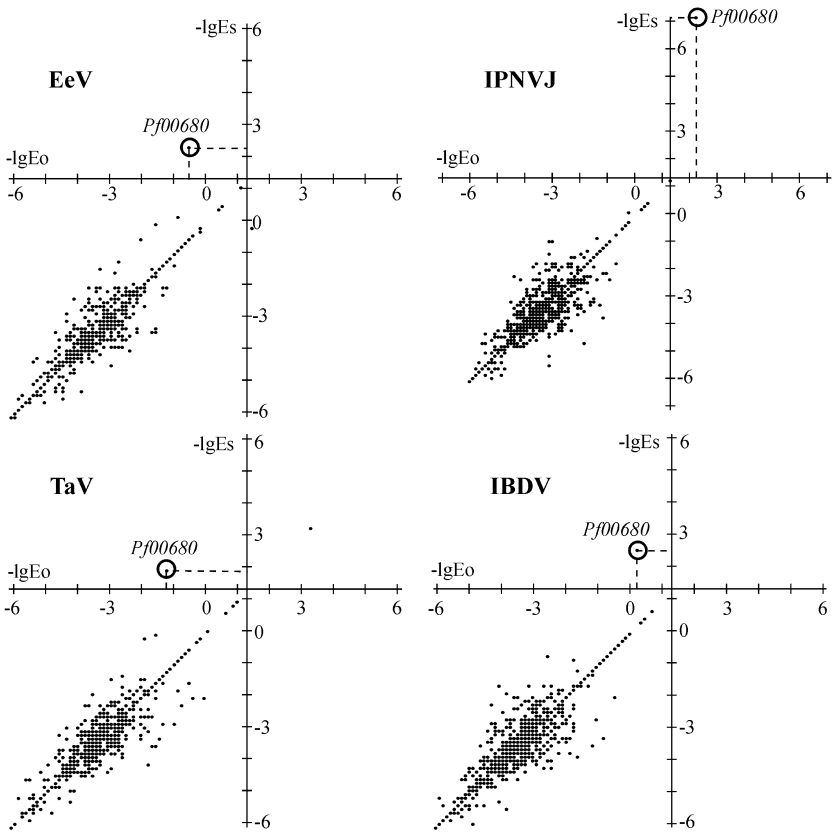


Figure 9. Distributions of maximum similarity scores for comparison of the original and shuffled replicases of four viruses with known protein families. The original replicase sequences and top-scoring permutants (see Figure 8) of TaV, EeV, IPNVJ, and IBDV were used to scan a CD-database containing ~3500 entries using rps-Blast.[67] The $-\ln E$ scores for similarities of each profile with a pair of the original (abscisa) and shuffled (ordinate) sequences were recorded in the quadrant plots labeled according to virus. The intersection of the axes was set at the 1.3 score. Low-left quadrant, statistically insignificant hits of the original and shuffled sequences; upper-left quadrant, statistically insignificant hits of the original sequence and statistically significant hits of the shuffled sequence; upper-right quadrant, statistically significant hits of the original and shuffled sequences; lower-right quadrant, statistically significant hits of the original sequences and statistically insignificant hits of the shuffled sequences. The magnitude of effect of shuffling on the similarity between a replicase and a profile is estimated by the deviation of the hits' position from the imaginary 45° diagonal running through the intersection of the axes. The position of the PF00680 profile hit is highlighted and projected to the axes for each virus.

that motif-C-related peaks are very specific. Thus, the selected top-scoring sequences are *bona fide* quasi-canonical replicases.

For every analyzed virus, the most top-scoring permutant and its parental replicase sequence were then compared in a special test to assess the statistical significance and specificity of the selected permutations. This test included rps-Blast-mediated[67] comparisons of the pair of sequences with the ABCC in-house copy of a CD-database curated at the NCBI,[57] and results were plotted for each virus. The rps-Blast *E*-values were converted into the negative logarithm scores ($-\ln E$) with the ($E = 0.05$) threshold being 1.3. Unlike the respective parents, the quasi-canonical replicases of EeV, TaV and IBDV reached a statistically sound level of similarity with a profile of RdRps from Picornavirus-like viruses (Pfam database accession number PF00608[37]) (Figure 9, compare the PF00608 scores projected on the $-\ln E_s$ *versus* $-\ln E_o$ axes in the EeV, TaV and IBDV plots). The relocation of motif C? of the IPNVJ replicase also increased the already statistically significant similarity of the parental replicase and the PF00608 profile by five orders of magnitude (Figure 9, compare the PF00608 scores projected on the $-\ln E_s$ *versus* $-\ln E_o$ in the IPNVJ plot). Although shuffling also increased the similarity of the replicases with some other protein families from a pool of approximately 3500 profiles (Figure 9 and data not shown), these effects were statistically insignificant and could be stochastic in origin. It is worth noting that the profile database contains, in addition to the PF00608 family, several other RdRp families (e.g. PF00946, PF00972, PF00978, PF00998, PF02123) that were not significantly similar to the quasi-canonical replicases, indicating that the observed increase in similarity was very specific.

Atomic coordinates

The TaV replicase sequence was deposited in GenBank (accession number AF82930).

Acknowledgements

We are grateful to Andy Ball and Ellie Ehrenfeld for critical reading of early versions of the manuscript, Karol Miaskiewicz and the staff of ABCC for assistance with computer resources and software. B.T.L. & A.E.G. were partly supported with funds from the National Cancer Institute, National Institutes of Health, under contracts no. NO1-CO-56000 and NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade

names, commercial products, or organization imply endorsement by the US Government.

References

1. Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriol. Rev.* **35**, 235–241.
2. Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G. & Steitz, T. A. (1985). Structure of large fragment of *Escherichia coli* DNA polymerase I complexed with dTMP. *Nature*, **313**, 762–766.
3. Bressanelli, S., Tomei, L., Roussel, A., Incitti, I., Vitale, R. L., Mathieu, M. *et al.* (1999). Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proc. Natl Acad. Sci. USA*, **96**, 13034–13039.
4. Butcher, S. J., Grimes, J. M., Makeyev, E. V., Bamford, D. H. & Stuart, D. I. (2001). A mechanism for initiating RNA-dependent RNA polymerization. *Nature*, **410**, 235–240.
5. Delarue, M., Poch, O., Tordo, N., Moras, D. & Argos, P. (1990). An attempt to unify the structure of polymerases. *Protein Eng.* **3**, 461–467.
6. Hansen, J. L., Long, A. M. & Schultz, S. C. (1997). Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure*, **5**, 1109–1122.
7. Lesburg, C. A., Cable, M. B., Ferrari, E., Hong, Z., Mannarino, A. F. & Weber, P. C. (1999). Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nature Struct. Biol.* **6**, 937–943.
8. Steitz, T. A. (1999). DNA polymerases: structural diversity and common mechanisms. *J. Biol. Chem.* **274**, 17395–17398.
9. Ng, K. K., Cherney, M. M., Vazquez, A. L., Machin, A., Alonso, J. M., Parra, F. & James, M. N. (2002). Crystal structures of active and inactive conformations of a caliciviral RNA-dependent RNA polymerase. *J. Biol. Chem.* **277**, 1381–1387.
10. Poch, O., Sauvaget, I., Delarue, M. & Tordo, N. (1989). Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J.* **8**, 3867–3874.
11. O'Reilly, E. K. & Rao, C. C. (1998). Analysis of RNA-dependent RNA polymerase structure and function as guided by known polymerase structures and computer predictions of secondary structure. *Virology*, **252**, 287–303.
12. Villarreal, L. P. & DeFilippis, V. R. (2000). A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. *J. Virol.* **74**, 7079–7084.
13. Xiong, Y. & Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353–3362.
14. Wang, J., Sattar, A. K., Wang, C. C., Karam, J. D., Konigsberg, W. H. & Steitz, T. A. (1997). Crystal structure of a pol alpha family replication DNA polymerase from bacteriophage RB69. *Cell*, **89**, 1087–1099.
15. Gohara, D. W., Crotty, S., Arnold, J. J., Yoder, J. D., Andino, R. & Cameron, C. E. (2000). Poliovirus RNA-dependent RNA polymerase (3Dpol): structural, biochemical, and biological analysis of conserved structural motifs A and B. *J. Biol. Chem.* **275**, 25523–25532.
16. Argos, P. (1988). A sequence motif in many polymerases. *Nucl. Acids Res.* **16**, 9909–9916.
17. Bruenn, J. A. (1991). Relationships among the positive strand and double-strand RNA viruses as viewed through their RNA-dependent RNA polymerases. *Nucl. Acids Res.* **19**, 217–226.
18. Koonin, E. V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *J. Gen. Virol.* **72**, 2197–2206.
19. Zanotto, P. M., Gibbs, M. J., Gould, E. A. & Holmes, E. C. (1996). A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *J. Virol.* **70**, 6083–6096.
20. Goldbach, R. W. (1986). Molecular evolution of plant RNA viruses. *Annu. Rev. Phytopathol.* **24**, 289–310.
21. Strauss, J. H. & Strauss, E. G. (1988). Evolution of RNA viruses. *Annu. Rev. Microbiol.* **42**, 657–683.
22. Gorbalenya, A. E. & Koonin, E. V. (1993). Comparative analysis of the amino acid sequences of the key enzymes of the replication and expression of positive-strand RNA viruses. Validity of the approach and functional and evolutionary implications. *Sov. Sci. Rev. D Physicochem. Biol.* **11**, 1–84.
23. Pringle, F. M., Gordon, K. H. J., Hanzlik, T. N., Kalmakoff, J., Scotti, P. D. & Ward, V. K. (1999). A novel capsid expression strategy for *Thosea asigna* virus (Tetraviridae). *J. Gen. Virol.* **80**, 1855–1863.
24. Dobos, P. (1995). The molecular biology of infectious pancreatic necrosis virus (IPNV). *Annu. Rev. Fish Dis.* **5**, 25–54.
25. Kibenge, F. S., Dhillon, A. S. & Russell, R. G. (1988). Biochemistry and immunology of infectious bursal disease virus. *J. Gen. Virol.* **69**, 1757–1775.
26. Pringle, F. M., Kalmakoff, J. & Ward, V. K. (2001). Analysis of the capsid processing strategy of *Thosea asigna* virus using baculovirus expression of virus-like particles. *J. Gen. Virol.* **82**, 259–266.
27. Gorbalenya, A. E. & Koonin, E. V. (1988). Birnavirus RNA polymerase is related to polymerases of positive strand RNA viruses. *Nucl. Acids Res.* **16**, 7735.
28. Duncan, R., Mason, C. L., Nagy, E., Leong, J. A. & Dobos, P. (1991). Sequence analysis of infectious pancreatic necrosis virus genome segment B and its encoded VP1 protein: a putative RNA-dependent RNA polymerase lacking the Gly-Asp-Asp motif. *Virology*, **181**, 541–552.
29. Mundt, E. & Vakharia, V. N. (1996). Synthetic transcripts of double-stranded Birnavirus genome are infectious. *Proc. Natl Acad. Sci. USA*, **93**, 11131–11136.
30. Yao, K. & Vakharia, V. N. (1998). Generation of infectious pancreatic necrosis virus from cloned cDNA. *J. Virol.* **72**, 8913–8920.
31. Gordon, K. H. J., Williams, M. R., Hendry, D. A. & Hanzlik, T. N. (1999). Sequence of the genomic RNA of *Nudaurelia* beta virus (Tetraviridae) defines a novel virus genome organization. *Virology*, **258**, 42–53.
32. Gordon, K. H. J., Johnson, K. N. & Hanzlik, T. N. (1995). The larger genomic RNA of *Helicoverpa armigera* stunt tetravirus encodes the viral RNA polymerase and has a novel 3′-terminal tRNA-like structure. *Virology*, **208**, 84–98.
33. Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.
34. Gorbalenya, A. E. (2001). Big nidovirus genome: when count and order of domains matter. *Advan. Exp. Med. Biol.* **494**, 1–17.
35. Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. & Blinov, V. M. (1989). Coronavirus genome: prediction of putative functional domains in the non-structural

polyprotein by comparative amino acid sequence analysis. *Nucl. Acids Res.* **17**, 4847–4861.
36. Lindqvist, Y. & Schneider, G. (1997). Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.* **7**, 422–427.
37. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260–262.
38. Goldenberg, D. P. (1989). Circularly permuted proteins. *Protein Eng.* **2**, 493–495.
39. Jablonski, S. A. & Morrow, C. D. (1995). Mutation of the aspartic acid residues of the GDD sequence motif of poliovirus RNA-dependent RNA polymerase results in enzymes with altered metal ion requirements for activity. *J. Virol.* **69**, 1532–1539.
40. Uliel, S., Fliess, A. & Unger, R. (2001). Naturally occurring circular permutations in proteins. *Protein Eng.* **14**, 533–542.
41. Jung, J. & Lee, B. (2001). Circularly permuted proteins in the protein structure database. *Protein Sci.* **10**, 1881–1886.
42. Russell, R. B. & Ponting, C. P. (1998). Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**, 364–371.
43. Morozov, S., Yu, ., Dolja, V. V. & Atabekov, J. G. (1989). Probable reassortment of genomic elements among elongated RNA-containing plant viruses. *J. Mol. Evol.* **29**, 52–62.
44. Ball, L. A. (2001). Replication strategies of RNA viruses. In *Fields Virology* (Knipe, D. M., Howley, P. M., Griffin, D. E., Lamb, R. A., Martin, M. A., Roizman, B. & Straus, S. E., eds), pp. 105–118, Lippincott Williams & Wilkins, Philadelphia.
45. Birghan, C., Mundt, E. & Gorbalenya, A. E. (2000). A non-canonical lon proteinase lacking the ATPase domain employs the Ser-Lys catalytic dyad to exercise broad control over the life cycle of a double-stranded RNA virus. *EMBO J.* **19**, 114–123.
46. Buck, K. W. (1996). Comparison of the replication of positive-stranded RNA viruses of plants and animals. *Advan. Virus Res.* **47**, 159–251.
47. Dreher, T. W. (1999). Functions of the 3′-untranslated regions of positive strand RNA viral genomes. *Annu. Rev. Phytopathol.* **37**, 151–174.
48. van Regenmortel, M. H. V., Fauquet, C. M., Bishop, D. H. L., Carstens, E. B., Estes, M. K., Lemon, S. M., *et al.* (2000). Virus taxonomy. *Classification and Nomenclature of Viruses* Seventh Report of the International Committee on Taxonomy of viruses, Academic Press, San Diego pp. 1–1162.
49. Wimmer, E., Hellen, C. U. T. & Cao, X. (1993). Genetics of poliovirus. *Annu. Rev. Genet.* **27**, 353–435.
50. Donnelly, M. L., Hughes, L. E., Luke, G., Mendoza, H., ten Dam, E., Gani, D. & Ryan, M. D. (2001). The "cleavage" activities of foot-and-mouth disease virus 2A site-directed mutants and naturally occurring "2A-like" sequences. *J. Gen. Virol.* **82**, 1027–1041.
51. Shwed, P. S., Dobos, P., Cameron, L. A., Vakharia, V. N. & Duncan, R. (2002). Birnavirus VP1 proteins form a distinct subgroup of RNA-dependent RNA polymerases lacking a GDD motif. *Virology*, **296**, 241–250.
52. Meyers, G. & Thiel, H. J. (1996). Molecular characterization of pestiviruses. *Advan. Virus Res.* **47**, 53–118.
53. Delarue, M., Poterszman, A., Nikonov, S., Garber, M., Moras, D. & Thierry, J. C. (1994). Crystal structure of a prokaryotic aspartyl tRNA-synthetase. *EMBO J.* **13**, 3219–3229.
54. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
55. Holm, L. & Sander, C. (1995). DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily. *Trends Biochem. Sci.* **20**, 345–347.
56. Ahlquist, P. (2002). RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science*, **296**, 1270–1273.
57. Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. & Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucl. Acids Res.* **30**, 281–283.
58. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* **25**, 4876–4882.
59. Morgenstern, B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
60. Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574–578.
61. Nicholas, K. B., Nicholas, N. H. B., Jr & Deerfield, D. W. (1997). GeneDoc: analysis and visualization of genetic variation. *EMBNET News*, **4**, 1–4.
62. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
63. Cuff, J. A. & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Struct. Funct. Genet.* **40**, 502–511.
64. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.
65. McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
66. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.* **10**, 19–29.
67. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
68. Ponting, C. P., Schultz, J., Copley, R. R., Andrade, M. A. & Bork, P. (2000). Evolution of domain families. *Advan. Protein Chem.* **54**, 185–244.
69. Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
70. Kimura, M. (1983). *The neutral theory of molecular evolution*, Cambridge University Press, Cambridge, NY.
71. Swofford, D. L. (2000). *PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4*, Sinauer Associates, Sunderland, MA.
72. Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357–358.
73. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52–56.

74. Uliel, S., Fliess, A., Amir, A. & Unger, R. (1999). A simple algorithm for detecting circular permutations in proteins. *Bioinformatics*, **15**, 930–936.
75. Koonin, E. V., Gorbalenya, A. E., Purdy, M. A., Rozanov, M. N., Reyes, G. R. & Bradley, D. W. (1992). Computer-assisted assignment of functional domains in the nonstructural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses. *Proc. Natl Acad. Sci. USA*, **89**, 8259–8263.
76. Meng, X. J., Purcell, R. H., Halbur, P. G., Lehman, J. R., Webb, D. M., Tsareva, T. S. *et al.* (1997). A novel virus in swine is closely related to the human hepatitis E virus. *Proc. Natl Acad. Sci. USA*, **94**, 9860–9865.
77. Cavanagh, D. (1997). Nidovirales: a new order comprising coronaviridae and arteriviridae. *Arch. Virol.* **142**, 629–633.
78. Cowley, J. A. & Walker, P. J. (2002). The complete genome sequence of gill-associated virus of *Penaeus monodon* prawns indicates a gene organization unique among nidoviruses. *Arch. Virol.* **147**, 1977–1987.

Edited by J. Karn

(Received 17 July 2002; received in revised form 11 September 2002; accepted 20 September 2002)