Virus-encoded proteinases of the Togaviridae

Edwin ten Dam,1 Michael Flint2 and Martin D. Ryan1

1 School of Biomedical Sciences, University of St Andrews, Biomolecular Sciences Building, North Haugh, St Andrews KY16 9ST, UK
2 School of Animal and Microbial Sciences, University of Reading, Whiteknights, PO Box 228, Reading RG6 6AI, UK

Introduction

The Togaviridae are enveloped, positive-sense, single-stranded RNA viruses. The family is divided into two genera; the alphaviruses, typified by Sindbis virus (genomic ~ 11.7 kb) and the rubiviruses, typified by rubella virus (genomic ~ 9.7 kb). A major feature of the replication of togaviruses is the synthesis of a single sub-genomic RNA transcript, corresponding to the 3' region of the genome (alphaviruses ~ 41 kb or '26S' RNA; rubiviruses ~ 33 kb; see Fig. 1). This mRNA transcript encodes a polyprotein comprising proteins involved in the encapsidation and envelopment of the virus. Togavirus non-structural (NS) proteins are encoded in the 5' region of genome-length RNA, also in the form of a polyprotein (Fig. 1).

The ability of togaviruses to synthesize a sub-genomic RNA is in contrast with the viruses which were discussed previously in this series of reviews: the Picornaviridae (Ryan & Flint, 1997) and the Flaviviridae (Ryan et al., 1998). In both the Picorna- and Flaviviridae all of their proteins are encoded in a single, long, open reading frame (ORF) with the virus proteins being generated by proteolytic ‘processing’ of the polyprotein. This processing is accomplished by virus-encoded proteinases: either wholly (Picornaaviridae) or in combination with cellular proteinases (Flaviviridae). Whereas for picorna- and flaviviruses control over the biogenesis of virus proteins is solely post-transcriptional (by regulation of proteolytic processing), the togaviruses are able to regulate the biogenesis of encapsidation and replicative proteins in both a temporal and quantitative manner via transcription of the sub-genomic mRNA. Unlike the picorna- and flaviviruses, translation of togavirus replicative proteins does not obligate the translation of encapsidation functions, nor vice versa. The polypeptides derived from either genomic or sub-genomic RNAs are proteolytically processed: in the former by a proteinase domain contained within the NS polyprotein itself, in the latter by a combination of virus and host-cell proteolytic activities (alphavirus) or host-cell proteinases alone (rubivirus).

Reviews are available on other members of the alphavirus ‘super-group': the caliciviruses (Clarke & Lambden, 1997) and the arteriviruses (Snijder & Meulenberg, 1998), together with the alpha- and rubiviruses themselves (Strauss & Strauss, 1994; Frey, 1994). The reader is directed to reviews on virus-encoded proteinases in general (Dougherty & Semler, 1993) and viral cysteine proteinases in particular (Gorbalenya & Snijder, 1996). A further resource is the comprehensive and systematic database of proteinases which classifies proteinases into clans and families ('MEROPS'; accessible on the worldwide web at http://www.bi.bbsrc.ac.uk/Merops/Merops.htm). This database also provides links to sequence databases and 3-D images. The proteinase nomenclature used here is from Barrett et al. (1998). Sequence alignments described in this paper are available from ftp://ftp.st-and.ac.uk/pub/mdr1 together with alignments from previous reviews.

Togavirus non-structural polyproteins

The alphavirus 5’-proximal long ORF encodes the replicative functions of the virus and is translated as an [nsP1–3] polyprotein. At low frequency (5–20% – temperature dependent) an amino acid may be incorporated at the site of the ‘leaky’ opal (UGA) stop codon at the end of nsP3 (translational ‘readthrough’), producing [nsP1–4] (de Groot et al., 1990; Li & Rice, 1993; Strauss et al., 1983). Although the identity of the amino acid incorporated at this site has not been confirmed by protein sequencing, it is likely to be arginine, cysteine or tryptophan by analogy with opal codon readthrough in retroviruses (Feng et al., 1990). When readthrough has produced [nsP1–4], proteolytic processing may occur at a site (nsP3/nsP4) six amino acids downstream of the amino acid incorporated at the UGA codon (Fig. 2) reviewed by Strauss & Strauss, 1994). In certain alphaviruses this stop codon is absent altogether and a sense codon [CGA (arginine) or UGU (cysteine)] is found. The polypeptide is processed into the ‘mature’ products: nsP1–nsP4. The rubivirus 5’-proximal long ORF also encodes the replicative functions of the virus. In this case, however, three translation products are observed; a 200 kDa protein (p200), which may then be processed into two products – p150 (1301 aa; mol. mass 140 kDa) and p90 (904 aa; mol. mass 100 kDa) – at the Gly1301/Gly1302 scissile pair (Chen et al., 1996; Fig. 1). The order of the gene functions in the NS polyprotein is different between alpha- and rubiviruses. In alphavirus functions have been ascribed as (N-to C-terminal): methyltransferase (nsP1), helicase/proteinase (nsP2), unknown ‘X’ motif (nsP3) and replicase (nsP4). In
rubiviruses a different order is observed: methyltransferase/‘X’/protease/helicase/replicase. The evolutionary relationships between these (and other positive-strand RNA viruses) is discussed by Koonin & Dolja (1993) and Goldbach & Wellink (1988).

(a) Alphavirus nsP2 protease (clan CA, family C9)

The alphavirus nsP2 protein is thought to be a multifunctional protein. A helicase motif is present in the N-terminal portion of the protein whilst two lines of evidence from early studies showed that protease activity resided in the C-terminal portion of the protein. Firstly, temperature-sensitive (ts) protease mutations (Hahn et al., 1989) and ts mutants deficient in processing of the polyprotein at the nsP1/nsP2, nsP2/nsP3 and nsP3/nsP4 sites (Hardy et al., 1990) were mapped to this region. Secondly, the analysis of in vitro translation profiles derived from truncated or deleted forms of cDNA also showed the proteolytic activity to reside in the C-terminal portion of the molecule (residues 475–728 – Ding & Schlesinger, 1989; Hardy & Strauss, 1989; see Fig. 1). Analysis of sequence alignments predicted the presence of a cysteine protease domain with the catalytic residues being Cys481 (the nucleophile) and His558 (Hardy & Strauss, 1989; Gorbalenya et al., 1991; see Fig. 3A). Subsequent analysis of the proteolytic activities of site-directed nsP2 mutants confirmed that residues Cys481 and His558 were indeed critical for protease activity (Strauss et al., 1992). These characteristics lead to the inclusion of nsP2 in clan CC (cysteine/histidine active site residues; Cys N-terminal to His). In many cysteine proteinases an asparagine residue is thought to play a role in catalysis (see Ryan & Flint, 1997, and references therein) and the possible importance of asparagine for the alphavirus nsP2 protease was examined. Mutation of all the conserved asparagine residues in the protease domain of nsP2 showed that none are absolutely required for proteolytic activity (Strauss et al., 1992), demonstrating that this type of papain-like cysteine protease (PCP) is functional with a catalytic dyad of cysteine and histidine and that an asparagine residue is not apparently required.

The residue immediately C-terminal of His558 is tryptophan. This is in contrast to plant, animal and other virus papain-like
Fig. 2. Cleavage sites of the alphavirus nonstructural proteins. Amino acid sequences at the sites of cleavage by nsP2 are shown. The universally conserved glycine at the P2 position is indicated in bold, amino acids conserved at each site are shaded and asterisks indicate opal stop codons at the nsP3/nsP4 junction. Virus sequences are as in Fig. 3, with chikungunya (CHICK – accession no. L37661), Ockelbo (OCK – accession no. M69205; Shirako et al., 1991), Igbo Ora (IO – accession no. AF079457), Middelburg (MID – Strauss et al., 1983) and Whataroa (WA – see Strauss & Strauss, 1994).

Proteases where the side-chain of the residue immediately C-terminal of the active site histidine is small/aliphatic. It has been suggested that Trp^59 in alphavirus nsP2 could function in place of asparagine by interaction with Trp^182 (adjacent to the active site Cys^181) to maintain the conformation of the active site residues. Indeed, Trp^59 has been shown by site-directed mutagenesis to be essential for catalytic activity (Strauss et al., 1992). By analogy, an interaction between these large, aromatic, residues could be present in the cysteine protease domains of other virus groups (rubi-, arteri-, tymo- and coronaviruses).

If one considers the extent of the alphavirus nsP2 protease domain to be comparable with other cysteine protease (domains), then the C-terminal region of nsP2 would contain an additional domain of some 210 amino acids. This region may possess yet another biochemical function, but analysis of the proteolytic properties of C-terminally deleted forms of nsP2 suggests that this region also has a role in proteolysis (Hardy & Strauss, 1989). By analogy, an interaction between these large, aromatic, residues could be present in the cysteine protease domains of other virus groups (rubi-, arteri-, tymo- and coronaviruses).

Sequences flanking the NS polyprotein cleavage sites are shown in Fig. 2. Residues in the P3–P2–P1 positions are thought to be important for cleavage. The glycine residue at position P2 is completely conserved at all sites and was shown to be required for cleavage to occur (Shirako & Strauss, 1990). The alphavirus nsP2 protease domain has, therefore, a similar cleavage specificity to most other viral PCPs in cleaving between two short aliphatic residues. For residues immediately C-terminal of the cleavage site (P1^P2–P4^), similarities are observed between different virus polyproteins at any given cleavage site. This region is, however, dissimilar between the different cleavage sites for any particular polyprotein (see Fig. 2). Detailed kinetic data for the proteolysis at each of the different sites are unavailable. The presence or absence of the nsP1 and nsP3 domains in either target or cleaving polypeptide may, however, be highly influential in determining the rate of cleavage at a specific site (see below).

Much data are, however, available for the regulation of Sindbis virus nsP2 protease and its role in virus replication. A number of observations on the alphavirus polyprotein processing have been combined to account for aspects of virus replication (Hardy & Strauss, 1989; Shirako & Strauss, 1990; de Groot et al., 1990; Sawicki & Sawicki, 1993, 1994; Lemm & Rice, 1993a, b; Lemm et al., 1994; Wang et al., 1994; Dé et al., 1996).

(i) All nsP2-containing polypeptides are proteolytically active
(ii) The kinetics of cleavage of each of the nsP sites is different and may require particular nsP2-containing forms for processing in trans
Fig. 3. For legend see opposite.
(iii) All nsP2-containing polypeptides were able to cleave the nsP1/2 site in trans
(iv) Proteinase forms containing nsP1 were unable to cleave the nsP3/4 site
(v) Only nsP3-containing proteinase forms were able to cleave the nsP3/4 site
(vi) Whilst nsP123pro can cleave the nsP1/2 site, it cannot cleave the nsP2/3 site
(vii) nsP23pro can cleave the nsP2/3 site very efficiently
(viii) Cleavage activity is not influenced by the presence or absence of nsP4

Using these proteolytic processing ‘rules’ an elegant model has been developed to account for the regulation of transcription of positive- and negative-sense RNA (de Groot et al., 1990; Lemm et al., 1994; Strauss & Strauss, 1994; Dé et al., 1996). The presence of nsP1 and/or nsP3 sequences within the proteolytically active form serves to alter the substrate specificity of the proteinase form and, thereby, the different biochemical activities which are generated by the alternative processing pathways. Initiation complexes containing an intact nsP123 or nsP23 have an high activity in minus-strand synthesis, whereas upon cleavage of the nsP23 junction plus-strand synthesis is preferred. An in vitro replication system consisting of nsP4, an RNA template and various nsP123 mutants showed minus-strand production. Replication was reduced in constructs that increased processing at the nsP23 junction, whereas mutations that abolished processing at this site stimulated the formation of minus strands. Processing at the nsP12 site was allowed, as shown by replication in a reaction containing nsP1, P23 (protease-disabled) and nsP4 as separate peptides (Lemm et al., 1998).

(b) Semliki Forest virus nsP4 protein

In Semliki Forest (SF), o’nyong-nyong and Igbo Ora viruses a sense (arginine) codon is found at the site corresponding to the opal stop codon delineating nsP3 and nsP4 in other alphaviruses. In these viruses, therefore, the entire ORF translated without suppression of a stop codon; the nsP4 protein is translated in equivolumar ratios to nsP1–3. The SFV nsP3/nsP4 junction is Ala1817/Tyr1818, determined by N-terminal sequencing (Fig. 2; Kalkkinen et al., 1981). Analyses of in vitro translation profiles derived from cDNA clones encoding SFV NS proteins were taken to show that cleavage of the SFV nsP3/4 site is not performed by the nsP2 proteinase. It was proposed that: (i) nsP4 is produced by a cleavage of nascent nsP1234, mediated by nsP4 and, (ii) proteolytic cleavages at the nsP1/2 and nsP2/3 sites were catalysed by the nsP2 proteinase (Takkinen et al., 1991). The aspartate residue of a conserved – DTG – motif (characteristic of aspartyl proteinases), close to the N terminus of nsP4, was suggested to be the active-site residue of the putative proteinase activity associated with nsP4 mediating the nsP3/nsP4 cleavage (Takkinen et al., 1990). Translation profiles from constructs encoding nsP4 with the N-terminal 102 residues deleted were, apparently, not subject to cleavage in vitro. Cleavage at the nsP3/nsP4 junction could not be mediated in trans; proteolysis could be inhibited by pepstatin A, but not by Zn2+ or by PMSF – consistent with the suggestion of an aspartyl proteinase mechanism (Takkinen et al., 1991). It should be noted, however, that this scheme would necessitate the interaction of two nsP4 molecules to form the ‘classical’ acid proteinase active site and that such bimolecular interactions (tri-, including interaction with the substrate) are difficult to observe using in vitro translation systems due to the extremely low amounts of translation products synthesized. Interestingly, the nsP3/nsP4 junction maintains good homology with the viruses using readthrough (see Fig. 2) and should therefore be a substrate for the nsP2 proteinase.

(c) Rubivirus proteinase (clan CA, family C27)

Note: the numbering scheme used here refers to the corrected rubella virus sequence (Pugachev et al., 1997). The analysis of sequence alignments predicted the NS polyprotein to contain a PCP domain with the catalytic residues being Cys1152 and His1273 (Gorbalenya et al., 1991; Figs 1 and 3A). Subsequent site-directed mutagenesis experiments confirmed that these residues were critical for proteolytic activity (Chen et al., 1996; Marr et al., 1994; Yao et al., 1998). The rubella virus proteinase domain mediates a single cleavage in the NS polyprotein at its own C terminus cleaving the p200 precursor into the p150 and p90 products (Fig. 1; Forng & Frey, 1995; Chen et al., 1996). By deletion mutagenesis the proteinase domain has been shown to map between amino acids 1006 and 1508 of the non-structural polyprotein (Marr et al., 1994).

Amino-terminal amino acid sequencing of the p90 cleavage product and site-directed mutagenetic studies showed the rubivirus NS polyprotein to be cleaved between residues Gly1301 and Gly1306 (Chen et al., 1996; Yao et al., 1998). This
Fig. 4. For legend see facing page.
cleavage-site specificity is similar to other viral papain-like proteinases [mouse hepatitis virus PLP-1 – Gly/Val (Hughes et al., 1995; Bonilla et al., 1997); equine arteritis virus – Gly/Gly (Snijder et al., 1994); potyvirus HC-Pro – Gly/Gly (Carrington et al., 1989); hypovirulence associated virus p29 protease – Gly/Gly (Choi et al., 1991; Shapira & Nuss, 1991). Characterization of the cleavage site by mutagenesis suggested that the P1 residue is the most important for cleavage, but that the P2 and P1' residues also contribute to recognition of the cleavage site (Chen et al., 1996).

The p150/p90 cleavage occurs in cis, although a recent analysis showed that the protease domain could also act in trans. Using an in vitro translation system co-translation studies showed cleavage in trans (Liu et al., 1998) of a proteolytically active recombinant polyprotein (in which the mutated cleavage site was not capable of being cleaved) on another polyprotein active recombinant polyprotein (in which the mutated cleavage site could also have affected structural integrity and inactivated the protease (Liu et al., 1998).

The rubella virus proteinase domain was found not to function following translation in a rabbit reticulocyte lysate in vitro translation system, although all of the other viral PCPs do so. However, following the supplementation of the translation system with divalent cations such as Zn²⁺, Cd²⁺ and Co²⁺, the rubella virus proteinase functioned efficiently indicating that these cations are required either as direct cofactors in catalytic activity or as a structural component of the protease. The picornavirus 2A and hepatitis C virus NS3 proteinases have a zinc atom as a structural feature, the metal not playing a direct role in catalysis (Sommergruber et al., 1994; Voss et al., 1995; Kim et al., 1996; Love et al., 1996; Yan et al., 1998). A putative zinc-binding motif has been identified in the sequence of the rubella virus proteinase domain (Liu et al., 1998; Fig. 3 A).

(d) Sequence alignments

The amino acid similarities/identities between disparate PCPs are very low although the presence of protease domains can be suspected by the presence of short motifs (including the Cys/His active site residues) and their spacing. Fig. 3 shows that the alphavirus nsP2 PCPs are conserved and are easily aligned. The rubivirus proteinase domain aligns with the alphavirus sequences very poorly. Indeed, a domain in the N-terminal region of the p90 protein (Gly₁₁₂₁-Pro₁₆₃₆) can be aligned with the alphavirus sequences very nearly as well as the ‘authentic’ proteinase domain (Fig. 3 B). Such PCP domains are found in a wide range of positive-strand RNA virus polyproteins and are discussed in the excellent review of Gorbalenya & Snijder (1996).

Structural polyprotein

The togavirus structural polyprotein comprises the nucleocapsid protein (C) and glycoproteins (Fig. 1). In such enveloped viruses the nucleocapsid protein must interact with both the RNA genome (packaging) and the cytoplasmic domains of the glycoproteins (particle morphogenesis) in addition to C–C interactions to form the capsid structure. In both alpha- and rubiviruses the nucleocapsid protein is N-terminal of the polyprotein. Early work showed that the translation products derived from alpha- and rubivirus sub-genomic mRNAs differ: the alphavirus structural polyprotein was very rapidly processed at the C/E3 site into two products whilst the rubivirus polyprotein was uncleaved. Supplementation of translation systems with microsomes produced full processing in both systems (reviewed by Strauss & Strauss, 1994; Frey, 1994). Processing of the rubivirus structural polyprotein is mediated, therefore, by host-cell signalases alone.

Alphavirus chymotrypsin-like proteolytic capsid protein (clan PA, family S3)

The first ~100 residues of the capsid protein are not highly conserved amongst alphaviruses. This region is rich in basic/proline residues (Fig. 4) and is thought to protrude into the interior of the capsid where it interacts with the genomic RNA via electrostatic interactions.

The activity which brings about the microsome-independent C/E3 polyprotein cleavage of the alphavirus structural polyprotein was identified as a proteolytic activity of the C protein itself. Site-directed mutagenetic experiments identified two components of the catalytic triad, His¹¹¹ and Ser²¹⁵ (Melancon & Garoff, 1987; Hahn & Strauss, 1990), the third member (Asp¹⁶₃) latterly being identified by resolution of the atomic structure of the protein (Choi et al., 1991; Tong et al., 1993). This work confirmed earlier suggestions of a serine proteinase-like activity of the nucleocapsid protein (Boege et al., 1981). Substitution of Ser²¹⁵ by cysisteine did not inhibit C/E3 processing, but did prevent formation of infectious Sindbis virus particles (Hahn & Strauss, 1990). The atomic structure of the capsid protein of another alphavirus, SFV, showed very similar features to the Sindbis structure (Choi et al., 1997).

The capsid protein of alphaviruses cleaves only at its own C terminus in an intramolecular, co- translational, fashion (in cis) at a conserved Trp/Ser scissile pair. The C-terminal region in

---

Fig. 4. Sequence alignment of alphavirus capsid proteins. The atomic structure was determined from residue Arg¹¹⁴ (Choi et al., 1991; Tong et al., 1993). The active site Asp, His and Ser residues are indicated (*). Note: the Tyr residue in place of the active site His in the published AURA sequence is a sequencing error and, in fact, is His (J. H. Strauss, personal communication). The assignment of the β-sheets is that given in Tong et al. (1993). Virus sequences are as in Figs 1 and 2, with Venezuelan equine encephalitis (VEEV – Kinney et al., 1989) and Western equine encephalitis (WEEV – Hahn et al., 1988). The numbering scheme refers to polyprotein positions.
chymotrypsin-like serine proteinases is an α-helix some distance from the active site. The C-terminal region of the alphavirus capsid protein is, however, not helical and the structure terminates within the active site, the ultimate tryptophan residue remaining in the P1 position. This accounts for the absence of catalytic activity following the C/E3 cleavage in cis. Neither in the original, nor refined, atomic structures was it possible to determine whether, following proteolysis, a covalent bond was present between the carboxyl group of Trp<sup>264</sup> and the side-chain of the nucleophile Ser<sup>215</sup> – equivalent to an unresolved acyl-enzyme intermediate form. Amino acid sequencing of tryptic peptides, however, suggests that the C terminus of the protein is not covalently linked to the catalytic serine (Boege et al., 1980). The active site residues have a similar topology to other serine proteinases. The unusual folding of the C-terminal region into the active site exposes the catalytic Asp<sup>163</sup> to the solvent. In other serine proteinases this active site residue is buried: shielding from the solvent may well be a requisite for the catalytic mechanism. During the co-translational cleavage another structural form may exist in which Asp<sup>163</sup> is buried. The residue at the bottom of the substrate binding pocket, thought to mediate a large part of the substrate specificity, is valine whilst other residues surrounding the substrate binding site include a number of glycines. This creates a relatively hydrophobic pocket for the recognition of (and the space to accommodate) the bulky hydrophobic C-terminal tryptophan residue.

**Concluding remarks**

Proteolytic processing is an extremely common strategy amongst animal and plant positive-strand RNA viruses, only a few such plant viruses apparently generating all of their proteins from sub-genomic mRNAs. In many plant viruses (i.e. comov-, nepo- and banyuviruses) genomes are multi-partite, each strand encoding a polypeptide. Some of the ‘primary’ polypeptide cleavages observed in the (single) polypeptides of monopartite viruses are, patentently, not required in these viruses (see Ryan & Flint, 1997). In this review we have remarked upon two other strategies: synthesis of sub-genomic mRNA and readthrough of leaky stop codons. In general, non-enveloped positive-strand RNA viruses do not require the involvement of host proteinases in the processing of their polypeptides, whilst the enveloped viruses recruit host enzymes for the processing of precursors comprising encapsidation functions (i.e., proteolytic processing in the lumen of the endoplasmic reticulum), and, in some cases, for the processing of the non-structural protein precursor.

In this series of reviews we have discussed one of the major strategies adopted by positive-strand RNA viruses for the biogenesis of proteins during replication – polypeptide processing by virus-encoded proteinases. For the limited number of atomic structures available for the virus proteinases, remarkable structural similarities to their cellular counterparts are observed, although sequence similarities may be very low. In considering virus and cellular proteinases as a whole, however, the striking feature is the conservation of the catalytic mechanism and the active site stereochemistry within a (virus or cellular) proteinase type.

Perhaps the most noticeable difference between virus and cellular proteinases is a theme now developing in the literature: the subtle regulation of the activity and specificity of the virus enzymes by binding other virus proteins or virus RNA, and, the differences in activity which are observed between larger, uncleaved, forms containing proteolytic domains. These may be metastable, proteolytically active, and not necessarily precursor forms as such – i.e. 3CD<sup>pro</sup> (picornaviruses) and the various nsP2-containing products described above.

Virus-encoded proteinases are an attractive target for antiviral drugs. Relatively simple assays may be developed which are suitable for automated high-throughput screening of combinatorial chemical libraries. The study of the regulation of these enzymes can, however, give major insights into the control of virus replicative processes and provide alternative strategies for future drug development.

**References**


