Chapter 4

Modelling of Foot and Mouth Disease Virus 3C and 3D Non-structural Proteins

4.1. Introduction

One of the most important proteases in FMDV is the $3C^{pro}$ and its $3C^{pro}$-containing precursor, 3CD. $3C^{pro}$ is responsible for viral polyprotein cleavage as well as some cleavage of cellular proteins such as eIF4G. The $3C^{pro}$ has been shown to efficiently process ten of the thirteen cleavage sites in the FMDV polyprotein (Bablanian and Grubman, 1993). $3C^{pro}$ is important in virus production as it cleaves the single translated polyprotein into the mature viral proteins needed for virus replication. The specificity of FMDV $3C^{pro}$ differs from its homologue in other picornaviruses like the Poliovirus. In polio $3C^{pro}$ only cleaves between Gln-Gly sites whereas in FMDV cleavage can occur between multiple dipeptides such as Gln-Gly, Glu-Gly, Gln-Leu and Glu-Ser (Palmenberg, 1990; Birtley *et al.*, 2005). Evolutionary studies have shown that the $3C^{pro}$ belongs to the trypsin family of Ser proteinases (Bablanian and Grubman, 1993). This is supported by the $3C^{pro}$ structure from FMDV, which shows a chymotrypsin-like fold (Fig. 4.1) and possesses a Cys-His-Asp catalytic triad in the active site (Birtley *et al.*, 2005). This chymotrypsin-like fold consists of two β-barrels positioned against one another with the active site between the two β-barrels. In FMDV an anti-parallel β-ribbon covers the active site. Sweeney and co-workers (Sweeney *et al.*, 2007) postulated that the β-ribbon is involved in substrate recognition. The β-ribbon is stabilized via hydrophobic contacts with the N-terminal barrel. The N-terminal barrel also contains an invariant region (residues 76-91) with

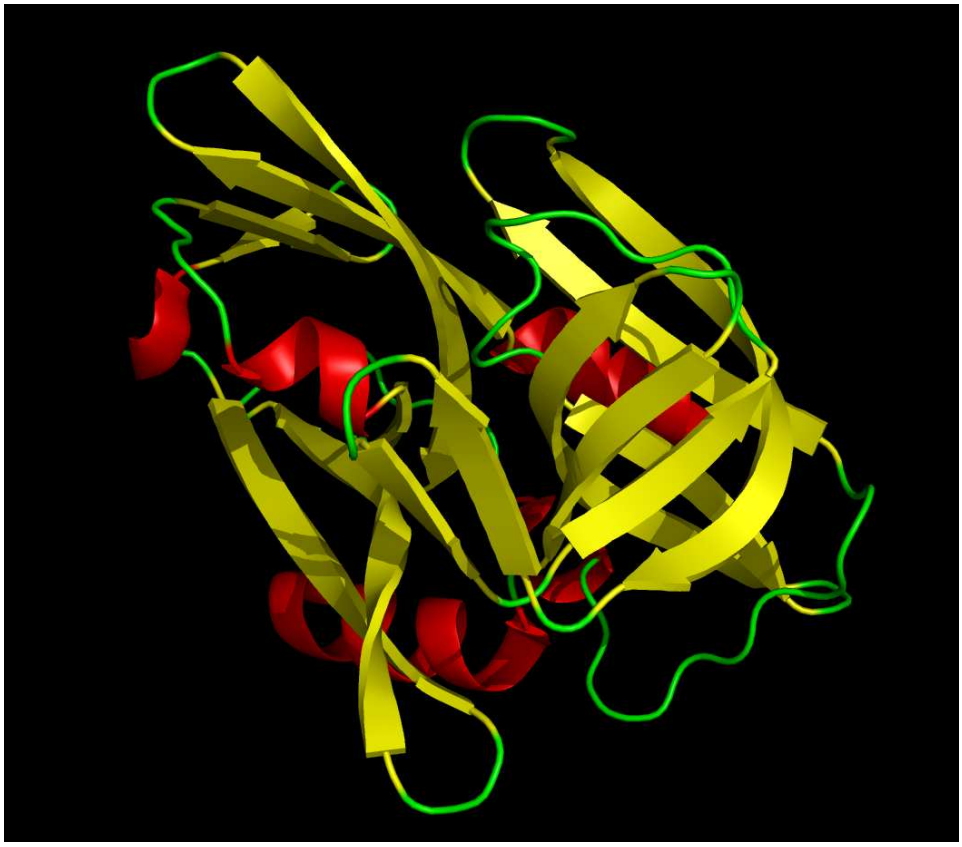

Figure 4.1: The structure of $3C^{pro}$ from FMDV serotype A (Sweeney *et al.*, 2007). Helices coloured red, strands coloured yellow. The β-ribbon can be seen in the foreground covering the active site.

the Asp at position 84 forming part of the catalytic triad (Carrillo *et al.*, 2005). The β-ribbon is quite flexible and very similar to other 14-residue β-ribbons that occur in other bacterial and viral serine proteases (Sweeney *et al.*, 2007). Most of the differences between the different β-ribbons occur neighbouring the turn in the ribbon and all the ribbons seem to be stabilized at the bottom of the ribbon via hydrophobic interactions.

The precursor, $3CD^{pro}$, has some protease activity and also participates in ribonucleoprotein complexes and influences RNA replication and translation by binding to RNA.

The $3D^{pol}$ protein that is produced from the cleavage of 3CD is a RNA dependant RNA polymerase encoded by the viral genome. The $3D^{pol}$ sequence (both RNA and protein) is conserved between the different sub- and serotypes (George *et al.*, 2001). $3D^{pol}$ is responsible for, in collaboration with host proteins, elongation of the nascent RNA chains during replication. The structure of FMDV $3D^{pol}$ is very similar to that of the poliovirus

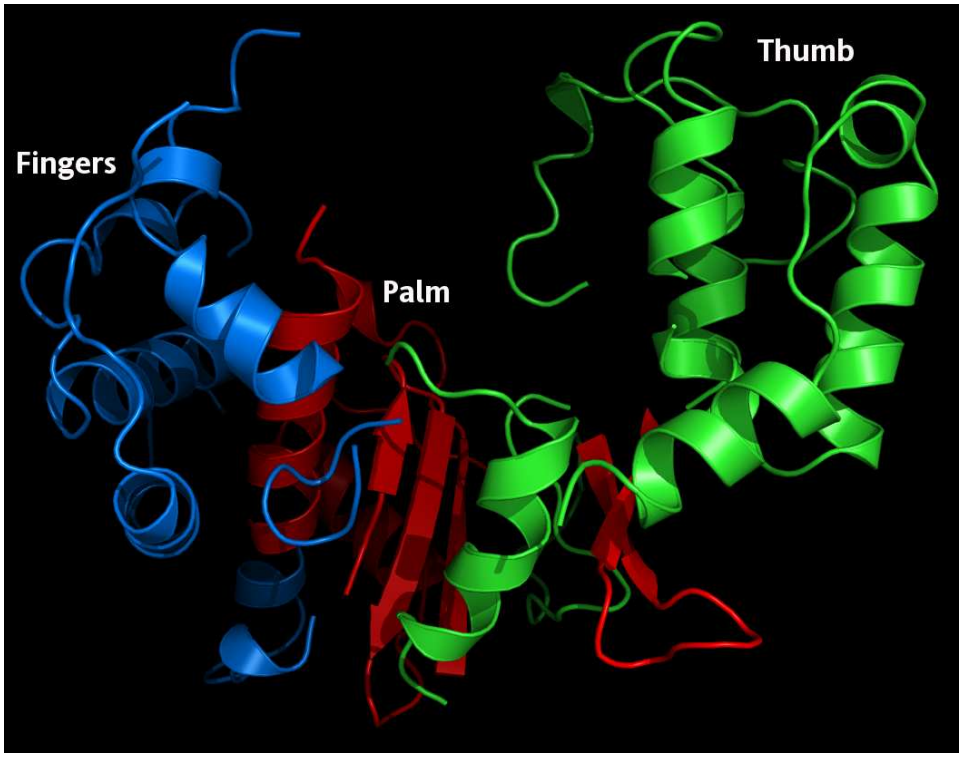

Figure 4.2: The structure of $3D^{pol}$ from the Polio virus (1RDR). Notice the 'palm' (red), 'fingers' (blue) and 'thumb' (green) subdomains (Hansen *et al.*, 1997).

$3D^{pol}$. This structure consists of a 'right-hand' polymerase consisting of 'palm', 'fingers' and 'thumb' subdomains (Fig. 4.2). It contains 17 α-helices and 16 β-strands. The palm subdomain contains some of the most highly conserved features known in all polymerases (Ferrer-Orta *et al.*, 2004). There are five conserved regions designated A-E, which are involved in phosphoryl transfer, nucleotide binding, nucleotide priming and structural integrity. A site in Motif A (Asp240 and Asp 245 in β8) helps motif C with metal ion binding as observed in the 1U09 structure. Motif B is made up of helix α11 that associates with a central β-sheet (β8, β11 and β12). Motif C, consisting of β11-turn-β12, contains the acidic sequence GDD (Gly 337-Asp338-Asp339). This acidic area is almost universally conserved and functions as a metal ion binding site during the nucleotide transfer reaction. Helix α12 forms motif D and β14 and β15 forms motif E. These motifs interact together to form the polymerase catalytic site.

Various studies have indicated the highly conserved nature of 3C and 3D (George *et al.*, 2001, Gorbalenya *et al.*, 1989, Carrillo *et al.*, 2005). In this section, the variation found in

these two proteins of the South African Territories serotypes of FMDV, will be presented. The objective is to identify local variation hotspots within the two proteins. This analysis may also help to identify the 3C-3D interaction site by identifying the most conserved residues based on the structure. Highly conserved patches on the surface may indicate areas that need to be conserved for interaction between 3C and 3D.

4.2. Methods

4.2.1. 3C Protease

Dr. F. Maree (Agricultural Research Council) supplied 21 SAT1, 21 SAT2 and 9 SAT3 sequences (Table 4.1). Alignment was done with ClustalX (Thompson *et al.*, 1997) and due to the high identity the parameters were kept at the default settings. The modelling scripts were generated with the Structural module in FunGIMS and modelling done with Modeller 9v1(Fiser and Sali, 2003) including a fast model refinement step. Models of representative sequences of serotypes SAT1, SAT2 and SAT3 were built based on 2J92 (Sweeney *et al.*, 2007), which is an serotype A virus. For SAT1, KNP/196/91/1 was used with the first five and the last 6 residues removed, for SAT2, ZIM/7/83/2 was used with the first and the last 6 residues removed and for SAT3, KNP/10/90/3 was used with the first and last 6 residues removed. The start and end residues were removed due to no template match for those regions. Another possible template was found (2BHG) but it was decided to use 2J92 as an important loop was crystallized in 2J92 that is not present in the higher resolution of 2BHG (1.90 Å vs 2.20 Å).

4.2.2. 3D RNA Polymerase

Dr. F. Maree (Agricultural Research Council) supplied 9 SAT1, 4 SAT2 and 3 SAT3 sequences (Table 4.1). A FMDV 3D sequence was submitted to a Blastp search against the PDB and it identified two protein structures (1U09 and 2D7S). Both these structures are FMDV 3D structures. It was decided to use 1U09 (Ferrer-Orta *et al.*, 2004) as its resolution was 1.91Å vs 3.00Å of 2D7S. Alignment was done with ClustalX using the

Table 4.1: Top: The SAT serotypes 3C protease sequences used in the variation analysis. Bottom: The SAT serotypes used in the 3D RNA polymerase variation analysis. Provided by Dr. F. Maree of the ARC. The sequences missing a number after the '/' lack a date in the original GenBank entry.

SAT subtype 3C sequences		
SAT1	SAT2	SAT3
SAT1/UGA/3/99 (gi:62362307)	SAT2/ZIM/7/83 (gi:33332022)	SAT3/KNP/10/90 (gi:21434547)
SAT1/UGA/1/97 (gi:15419327)	SAT2/KNP/19/89 (gi:15419331)	SAT3/ZAM/4/96 (gi:62362337)
SAT1/SUD/3/76 (gi:62362303)	SAT2/SAR/16/83 (gi:62362321)	SAT3/ZIM/5/91 (gi:62362339)
SAT1/NIG/15/75 (gi:62362299)	SAT2/ANG/4/74 (gi:62362311)	SAT3/MAL/03/76 (gi:12274987)
SAT1/NIG/5/81 (gi:62362297)	SAT2/KEN/8/99 (gi:62362315)	SAT3/BEC/1/65 (gi:21328275)
SAT1/TAN/37/99 (gi:62362305)	SAT2/ZIM/14/90 (gi:62362331)	SAT3/UGA/2/97 (gi:62362335)
SAT1/TAN/1/99 (gi:15419329)	SAT2/ZIM/17/91 (gi:62362333)	SAT3/KEN/3/ (gi:46810960)
SAT1/KNP/196/91 (gi:15419321)	SAT2/2/ (gi:46810952)	SAT3/BEC/3/ (gi:46810960)
SAT1/SAR/09/81 (gi:62362301)	SAT2/SEN/7/83 (gi:62362325)	SAT3/RSA/2/ (gi:46810956)
SAT1/ZAM/2/93 (gi:62362309)	SAT2/SEN/05/75 (gi:62362323)	
SAT1/NAM/307/98 (gi:62362295)	SAT2/ANG/4/74 (gi:62362311)	
SAT1/MOZ/3/02 (gi:62362341)	SAT2/MOZ/4/83 (gi:15419321)	
SAT1/KEN/5/98 (gi:62362293)	SAT2/RHO/1/48 (gi:62362317)	
SAT1/BOT/1/68 (gi:46810946)	SAT2/KEN/3/57 (gi:6572136)	
SAT1/RSA/5/ (gi:46810940)	SAT2/RWA/2/01 (gi:62362319)	
SAT1/SWA/6/ (gi:46810942)	SAT2/SAU/6/00 (gi:21434553)	
SAT1/RHO/ (gi:46810948)	SAT2/ZAI/1/74 (gi:62362329)	
SAT1/BEC/1/ (gi:46810932)	SAT2/GHA/8/91 (gi:62362313)	
SAT1/SWA/3/ (gi:46810936)	SAT2/UGA/2/02 (gi:62362327)	
SAT1/RHO/4/ (gi:46810938)	SAT2/3KEN/21/ (gi:6810954)	
SAT1/20/ (gi:46810934)	SAT2/RHO/1/48 (gi:46810950)	

SAT subtype 3D sequences		
SAT1	SAT2	SAT3
SAR/09/81 (not yet submitted)	ZIM/7/83 (gi:33332022)	KEN/3/ (gi:46810960)
BOT/1/68 (gi:46810946)	SAT2/2/ (gi:46810952)	RSA/2/ (gi:46810956)
SWA/6/ (gi:46810942)	RHO/1//48 (gi:62362317)	
RSA/5/ (gi:46810940)	3KEN/32/ (gi:6810954)	
RHO/4/ (gi:46810938)		
SWA/3/ (gi:46810936)		
BEC/1/ (gi:46810932)		
RHO/ (gi:46810948)		
SAT1/20/ (gi:46810934)		

default parameters, modelling scripts generated with the Structural module in FunGIMS and modelling done with Modeller 9v1 including a fast model refinement step. SAR/09/81 was used as a representative sequence for SAT1, ZIM/7/83/2 was used for SAT2 and RSA/2/3 was used for SAT3. In all cases the SAT target was 6 residues shorter than the template.

4.3. Results and Discussion

Because the various SAT serotypes are so similar, a representative model was built for each serotype (SAT1, SAT2 and SAT3). The variation for each serotype was then mapped onto the respective model.

4.3.1. 3C Protease

The SAT isolates included in this study are represented across Africa and include isolates from West, East, Central and Southern Africa. All the sequences used to build the respective models for $3C^{pro}$ showed ~85% identity with 2J92. This was to be expected as the conservation of FMDV $3C^{pro}$ is high. The alignments that were used in modelling the $3C^{pro}$ SAT serotypes are shown in Figure 4.3 and the high identity between target and template is indicated.

After the KNP/96/91/1 SAT1 $3C^{pro}$ model was built, the variation observed in the SAT1 $3C^{pro}$ alignment was mapped onto the model (Fig. 4.5). There was variation at 45 residue positions (21%) within the 21 SAT sequences. In 76% (35) of the positions, variation was limited to 2 amino acids, 20% (9) of the positions were limited to 3 amino acids and 4% (2) limited to 4 amino acids.

ZIM/7/83/2 was used for the SAT2 model. SAT2 showed 41% more variance between the 21 SAT2 sequences compared to SAT1. Variation was observed in 63 positions (30%) and mapped to a SAT2 3C model (Fig. 4.5). In 76% (48) of the positions, variation was limited to 2 amino acids, 16% (10) of the positions was limited to 3 amino acids, 6% (4) limited to 4 amino acids and 2% (1) limited to 5 amino acids.

A.

```
2J92            1 ---QKMVMGNTKPVELILDGKTVAICCATGVFGTAYLVPRHLFAEQYDKIMLDGRAMTDS
SAT1KNP196-91   1 TDLQKMVMANVKPVELILDGKTVALCCATGVFGTAYLVPRHLFAEKYDKIMLDGRALTDS

2J92           58 DYRVFEFEIKVKGQDMLSDAALMVLHRGNKVRDITKHFRDTARMKKGTPVVGVVNNADVG
SAT1KNP196-91  61 DFRVFEFEVKVKGQDMLSDAALMVLHSGNRVRDLTGHFRDTMKLSKGSPVVGVVNNADVG

2J92          118 RLIFSGEALTYKDIVVSMDGDTMPGLFAYKAATRAGYAGGAVLAKDGADTFIVGTHSAGG
SAT1KNP196-91 121 RLIFSGDALTYKDLVVCMDGDTMPGLFAYRAGTKVGYCGAAVLAKDGAKTVIVGTHSAGG

2J92          178 NGVGYCSCVSRSMLQKMKAHV-
SAT1KNP196-91 181 NGVGYCSCVSRSMLLQMKAHID
```

B.

```
2J92            1 --QKMVMGNTKPVELILDGKTVAICCATGVFGTAYLVPRHLFAEQYDKIMLDGRAMTDSD
SAT2ZIM7-83     1 DLQKMVMANVKPVELILDGKTVALCCATGVFGTAYLVPRHLFAEKYDKIMLDGRALTDSD

2J92           59 YRVFEFEIKVKGQDMLSDAALMVLHRGNKVRDITKHFRDTARMKKGTPVVGVVNNADVGR
SAT2ZIM7-83    61 FRVFEFEVKVKGQDMLSDAALMVLHSGNRVRDLTGHFRDTMKLSKGSPVVGVVNNADVGR

2J92          119 LIFSGEALTYKDIVVSMDGDTMPGLFAYKAATRAGYAGGAVLAKDGADTFIVGTHSAGGN
SAT2ZIM7-83   121 LIFSGDALTYKDLVVCMDGDTMPGLFAYRAGTKVGYCGAAVLAKDGAKTVIVGTHSAGGN

2J92          179 GVGYCSCVSRSMLQKMKAHV-
SAT2ZIM7-83   181 GVGYCSCVSRSMLLQMKAHID
```

C.

```
2J92            1 --QKMVMGNTKPVELILDGKTVAICCATGVFGTAYLVPRHLFAEQYDKIMLDGRAMTDSD
SAT3KNP10-90    1 DLQKMVMANVKPVELILDGKTVALCCATGVFGTAYLVPRHLFAEKYDKIMLDGRALTDGD

2J92           59 YRVFEFEIKVKGQDMLSDAALMVLHRGNKVRDITKHFRDTARMKKGTPVVGVVNNADVGR
SAT3KNP10-90   61 FRVFEFEVKVKGQDMLSDAALMVLHSGNRVRDLTGHFRDTMKLSKGSPVVGVVNNADVGR

2J92          119 LIFSGEALTYKDIVVSMDGDTMPGLFAYKAATRAGYAGGAVLAKDGADTFIVGTHSAGGN
SAT3KNP10-90  121 LIFSGDALTYKDLVVCMDGDTMPGLFAYRAGTKVGYCGAAVLAKDGAKTVIVGTHSAGGN

2J92          179 GVGYCSCVSRSMLQKMKAHV-
SAT3KNP10-90  181 GVGYCSCVSRSMLLQMKAHID
```

Figure 4.3: The alignments used in the modelling of $3C^{pro}$. A: KNP/96/91/1. B: ZIM/7/82/2. C: KNP/10/90/3 with 2J92 being the template sequence (serotype A10).

KNP/10/90/3 was used as a representative for the SAT3 serotype. SAT3 showed 35% less variation than SAT1 and 54% less variation than SAT2 in the 9 sequences analyzed. There was variation in 29 positions (14%) of which 93% (27 positions) varied by 2 amino acids and 7% (2 positions) varied by 3 amino acids (Fig. 4.5). An important residue position was Asp 84 that is part of the catalytic triad. In ZIM/5/91/3 this Asp was replaced by a Tyr. This is the only occurrence in all the analyzed sequences where a mutation was present in the active site. There are 2 reasons for less variation in SAT3: SAT3 is not well represented in this study and it has a geographical distribution limited to Southern and Central Africa.

A.

```
1U09          1 -GLIVDTRDVEERVHVMRKTKLAPTVAHGVFNPEFGPAALSNKDPRLNEGVVLDEVIFSK
SAR09-81-1    1 EGLVVDTREVEERVHVMRKTKLAPTVAYGVFQPEFGPAALSNNDKRLNEGVVLDEVIFSK

1U09         60 HKGDTKMSAEDKALFRRCAADYASRLHSVLGTANAPLSIYEAIKGVDGLDAMEPDTAPGL
SAR09-81-1   61 HKGDAKMSEADKKLFRLCAADYASHLHNVLGTANSPLSVFEAIKGVDGLDAMEPDTAPGL

1U09        120 PWALQGKRRGALIDFENGTVGPEVEAALKLMEKREYKFACQTFLKDEIRPMEKVRAGKTR
SAR09-81-1  121 PWALQGKRRGALIDFENGTVGPEIEQALKLMEKKEYKFTCQTFLKDEIRPLEKVKAGKTR

1U09        180 IVDVLPVEHILYTRMMIGRFCAQMHSNNGPQIGSAVGCNPDVDWQRFGTHFAQYRNVWDV
SAR09-81-1  181 IVDVLPVEHIIYTRMMIGRFCAQMHSNNGPQIGSAVGCNPDVDWQRFGCHFAQYRNVWDI

1U09        240 DYSAFDANHCSDAMNIMFEEVFRTEFGFHPNAEWILKTLVNTEHAYENKRITVEGGMPSG
SAR09-81-1  241 DYSAFDANHCSDAMNIMFEEVFREEFGFHPNAVWILKTLINTEHAYENKRITVEGGMPSG

1U09        300 CSATSIINTILNNIYVLYALRRHYEGVELDTYTMISYGDDIVVASDYDLDFEALKPHFKS
SAR09-81-1  301 CSATSIINTILNNIYVLYALRRHYEGVELSHYTMISYGDDIVVASDYDLDFEALKPHFKS

1U09        360 LGQTITPADKSDKGFVLGHSITDVTFLKRHFHMDYGTGFYKPVMASKTLEAILSFARRGT
SAR09-81-1  361 LGQTITPADKSDKGFVLGQSITDVTFLKRHFHLDYGTGFYKPVMASKTLEAILSFARRGT

1U09        420 IQEKLISVAGLAVHSGPDEYRRLFEPFQGLFEIPSYRSLYLRWVNAVCGDAAALEHH
SAR09-81-1  421 IQEKLISVAGLAVHSGPDEYRRLFEPFQGTFEIPSYRSLYLRWVNAVCGDA------
```

B.

```
1U09          1 -GLIVDTRDVEERVHVMRKTKLAPTVAHGVFNPEFGPAALSNKDPRLNEGVVLDEVIFSK
ZIM-7-83-2    1 EGLVVDTREVEERVHVMRKTKLAPTVAHGVFQPEFGPAALSNNDKRLSEGVVLDEVIFSK

1U09         60 HKGDTKMSAEDKALFRRCAADYASRLHSVLGTANAPLSIYEAIKGVDGLDAMEPDTAPGL
ZIM-7-83-2   61 HKGDAKMSEADKRLFRLCAADYASHLHNVLGTANSPLSVFEAIKGVDGLDAMEPDTAPGL

1U09        120 PWALQGKRRGALIDFENGTVGPEVEAALKLMEKREYKFACQTFLKDEIRPMEKVRAGKTR
ZIM-7-83-2  121 PWALRGKRRGALIDFENGTVGSEIEAALKLMEKKEYKFTCQTFLKDEIRPLEKVKAGKTR

1U09        180 IVDVLPVEHILYTRMMIGRFCAQMHSNNGPQIGSAVGCNPDVDWQRFGTHFAQYRNVWDV
ZIM-7-83-2  181 IVDVLPVEHIIYTRMMIGRFCAQMHSNNGPQIGSAVGCNPDVDWQRFGTHFAQYKNVWDI

1U09        240 DYSAFDANHCSDAMNIMFEEVFRTEFGFHPNAEWILKTLVNTEHAYENKRITVEGGMPSG
ZIM-7-83-2  241 DYSAFDANHCSDAMNIMFEEVFREEFGFHPNAVWILKTLINTEHAYENKRITVEGGMPSG

1U09        300 CSATSIINTILNNIYVLYALRRHYEGVELDTYTMISYGDDIVVASDYDLDFEALKPHFKS
ZIM-7-83-2  301 CSATSIINTILNNIYVLYALRRHYEGVELSHYTMISYGDDIVVASDYDLDFEALKPHFKS

1U09        360 LGQTITPADKSDKGFVLGHSITDVTFLKRHFHMDYGTGFYKPVMASKTLEAILSFARRGT
ZIM-7-83-2  361 LGQTITPADKSDKGFVLGQSITDVTFLKRHFHLDYETGFYKPVMASKTLEAILSFARRGT

1U09        420 IQEKLISVAGLAVHSGPDEYRRLFEPFQGLFEIPSYRSLYLRWVNAVCGDAAALEHH
ZIM-7-83-2  421 IQEKLISVAGLAVHSGQDEYRRLFEPFQGTFEIPSYRSLYLRWVNAVCGDA------
```

C.

```
1U09          1 -GLIVDTRDVEERVHVMRKTKLAPTVAHGVFNPEFGPAALSNKDPRLNEGVVLDEVIFSK
RSA-2-3       1 EGLVVDTREVEERVHVMRKTKLAPTVAHGVFQPEFGPAALSNNDKRLNEGVVLDEVIFSK

1U09         60 HKGDTKMSAEDKALFRRCAADYASRLHSVLGTANAPLSIYEAIKGVDGLDAMEPDTAPGL
RSA-2-3      61 HKGDAKMSEADKKLFRLCAADYASHLHNVLGTANSPLSVFEAIKGVDGLDAMEPDTAPGL

1U09        120 PWALQGKRRGALIDFENGTVGPEVEAALKLMEKREYKFACQTFLKDEIRPMEKVRAGKTR
RSA-2-3     121 PWALQGRRRGALIDFENGTVGPEIEQALKLMEKKEYKFTCQTFLKDEIRPLEKVKAGKTR

1U09        180 IVDVLPVEHILYTRMMIGRFCAQMHSNNGPQIGSAVGCNPDVDWQRFGTHFAQYRNVWDV
RSA-2-3     181 IVDVLPVEHIIYTRMMIGRFCAQMHSNNGPQIGSAVGCNPDVDWQRFGCHFAQYKNVWDI

1U09        240 DYSAFDANHCSDAMNIMFEEVFRTEFGFHPNAEWILKTLVNTEHAYENKRITVEGGMPSG
RSA-2-3     241 DYSAFDANHCSDAMNIMFEEVFREEFGFHPNAVWVLKTLINTEHAYENKRITVEGGMPSG

1U09        300 CSATSIINTILNNIYVLYALRRHYEGVELDTYTMISYGDDIVVASDYDLDFEALKPHFKS
RSA-2-3     301 CSATSIINTILNNIYVLYALRRHYEGVELSHYTMISYGDDIVVASDYDLDFEALKPHFKS

1U09        360 LGQTITPADKSDKGFVLGHSITDVTFLKRHFHMDYGTGFYKPVMASKTLEAILSFARRGT
RSA-2-3     361 LGQTITPADKSDKGFVLGQSITDVTFLKRHFHLDYETGFYKPVMASKTLEAILSFARRGT

1U09        420 IQEKLISVAGLAVHSGPDEYRRLFEPFQGLFEIPSYRSLYLRWVNAVCGDAAALEHH
RSA-2-3     421 IQEKLISVAGLAVHSGQDEYRRLFEPFQGTFEIPSYRSLYLRWVNAVCGDA------
```

Figure 4.4: The alignments used in the modelling of 3D. A: SAR/09/81/1. B: ZIM/7/83/2. C: RSA/2/3.

Table 4.2: The changes observed in the SAT serotypes as compared to the invariant region from residue 76-91 identified by Carillo *et al.* (2005). A structural representation of the invariant region can be seen in figure 4.8.

Subtype	Variation (aa71-86)	Effect
Invariant region	VKGQDMLSDAALMVLH	-
SAT1/UGA/1/97	VKGQDMLSDAALMVL*N*	Maintains backbone H-bond and side-chain H-bond
SAT1/UGA/3/99	VKGQDMLSDAALMVL*N*	Maintains backbone H-bond and side-chain H-bond
SAT1/NIG/15/75	VKGQ*E*MLSDAALMVLH	Maintains backbone H-bond and side-chain H-bond
SAT2/ZIM/17/91	VKG*P*DMLSDAALMVLH	Maintains backbone H-bond. Might distort the loop slightly
SAT2/KNP/19/89	VKGQDMLSDAALM*G*LH	Maintains backbone H-bond
SAT2/SEN/7/83	VKGQDM*M*SDAALMVL*N*	Maintains backbone H-bond and side-chain H-bond
SAT2/SEN/05/75	VKGQDM*M*SDAALMVL*N*	Maintains backbone H-bond and side-chain H-bond
SAT2/GHA/8/91	VKGQDM*M*SDAALMVL*N*	Maintains backbone H-bond and side-chain H-bond
SAT2/UGA/2/02	VKGQDMLSDAALMVL*N*	Maintains backbone H-bond and side-chain H-bond
SAT3/ZIM/5/91	VKGQDMLS*Y*AAL*I*VLH	This includes a mutation in the active site.
SAT3/UGA/2/97	VKGQDMLSDAALMVL*N*	Maintains backbone H-bond and side-chain H-bond

Most of the variation in the SAT $3C^{pro}$ seems to occur at one end of the C-terminal β-barrel (Fig. 4.6). This region is surface-exposed and can potentially accommodate more variation without influencing the activity of the enzyme. Another interesting observation was that the inner β-sheet in the C-terminal β-barrel contained very little variation and is conserved, whereas the N-terminal β-barrel contains significantly more variation.

An invariant section (residues 76-91, VKGQDMLSDAALMVLH) in $3C^{pro}$ identified by Carillo and co-workers (Fig. 4.8), was shown to contain variation within the SAT serotypes. Table 4.2 shows the aa changes for each isolate compared to the invariant region. Eleven isolates showed variation in the invariant region. The invariant region is located on two consecutive β-strands of which the second β-sheet (residues 85-91) contains one of the catalytic triad residues (Asp). A reason for this conservation of the

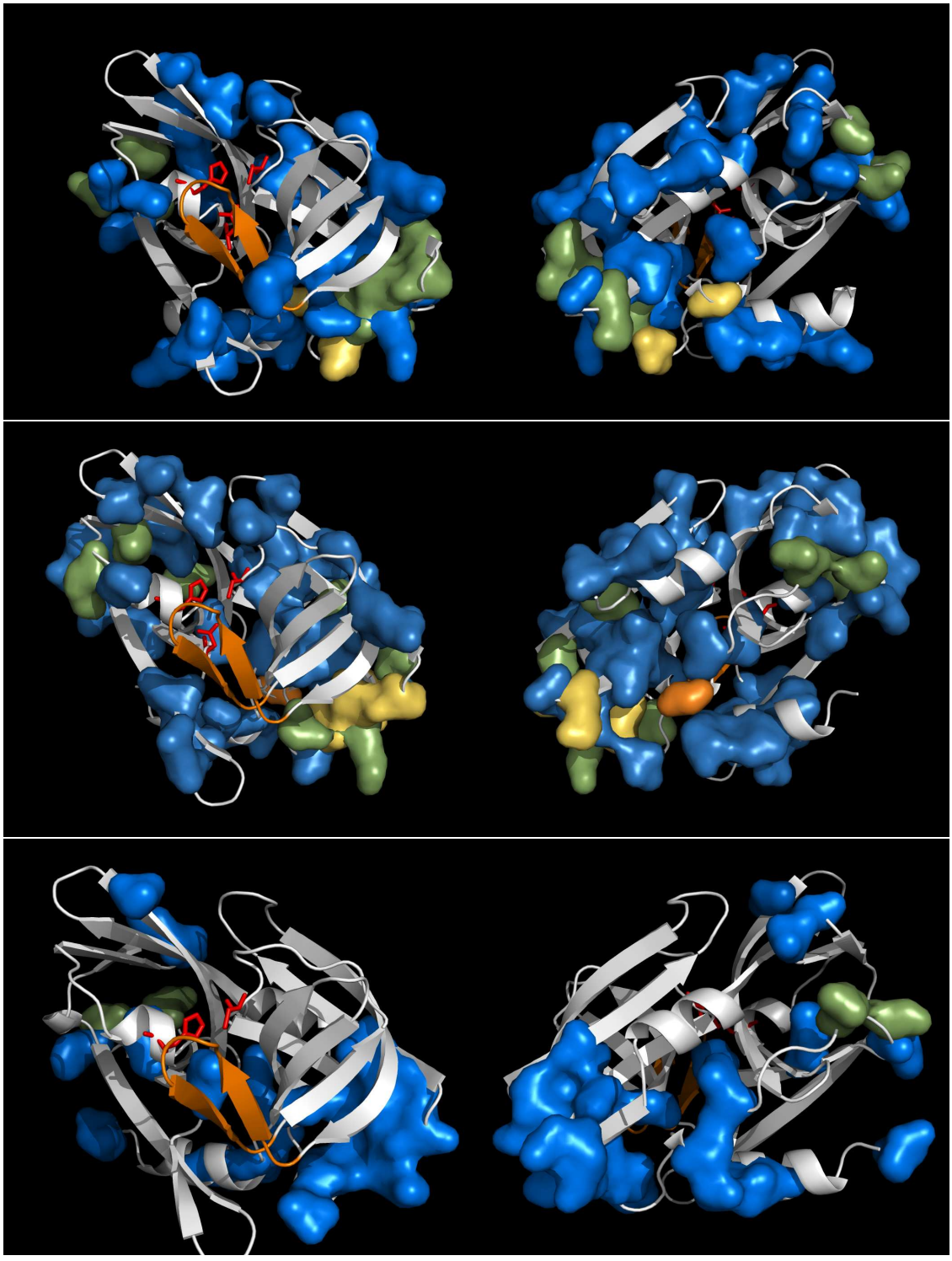

Figure 4.5: SAT $3C^{pro}$ variation mapped onto a SAT $3C^{pro}$model. Views from both sides of the enzyme are shown. Top: SAT1, middle: SAT2, bottom: SAT3. White indicates conserved positions across all the sequences analyzed, blue indicates 2 different residues found at that position, green indicates 3 different residues found at that position and yellow indicates the presence of 4 different residues. The active site catalytic triad is coloured red and the β-ribbon is coloured orange.

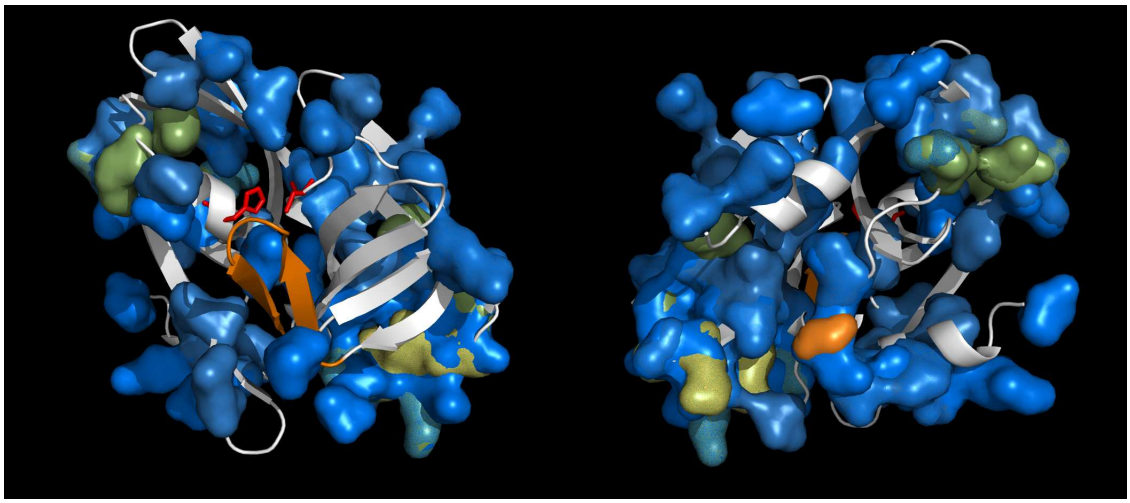

Figure 4.6: The variation seen in the $3C^{pro}$ protease as mapped to a cartoon representation of the enzyme. Both sides of the enzyme are shown. White indicates conserved positions across all the serotype sequences analyzed, blue indicates 2 different residues found at that position, green indicates 3 different residues found at that position and yellow indicates the presence of 4 different residues.

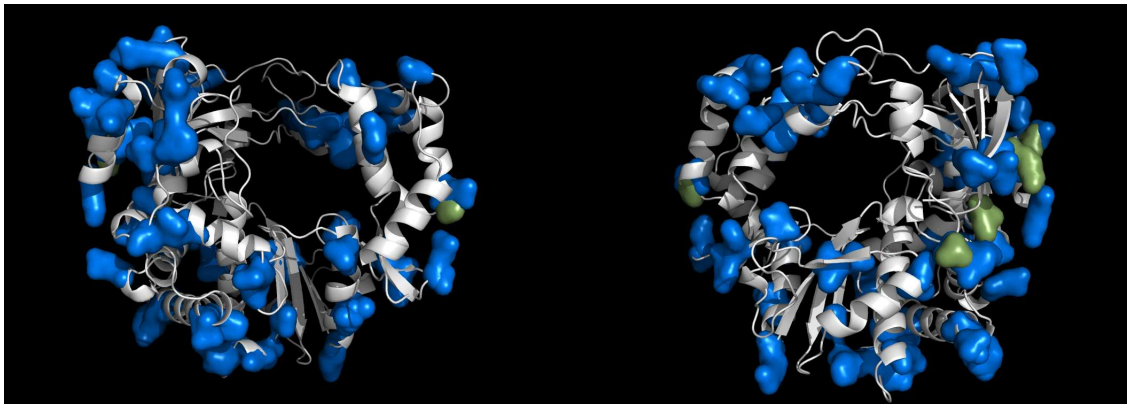

Figure 4.7: The variation seen in the 3D protease as mapped to a cartoon representation of the enzyme. Views from both sides are shown. White indicates conserved positions across all the serotype sequences analyzed, blue indicates 2 different residues found at that position and green indicates 3 different residues found at that position.

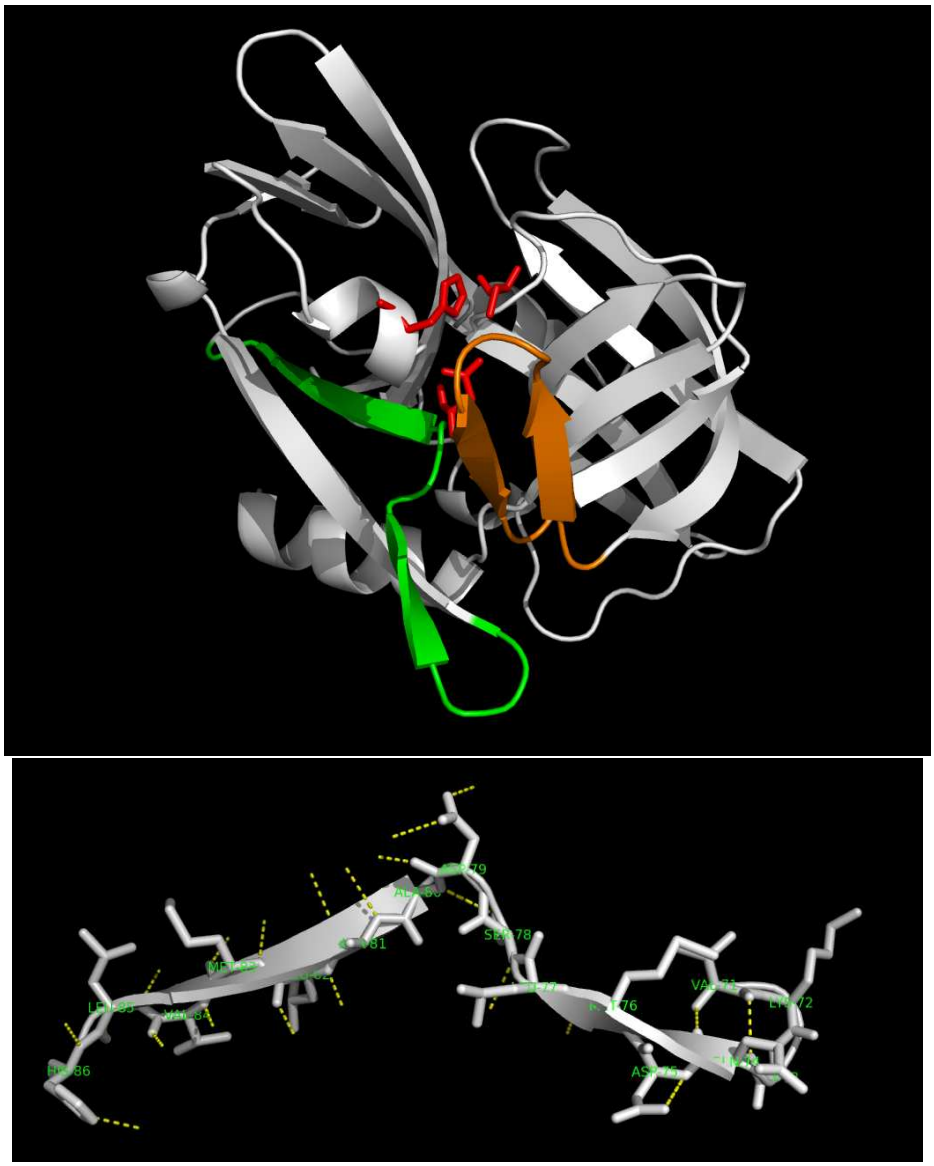

Figure 4.8: Top: The location of the invariant region identified by Carillo *et al.* in the $3C^{pro}$ structure. The numbers are the residue numbers used in the model and correspond to $3C^{pro}$ residues 76-91. Bottom: The hydrogen bond network for the invariant region. All residues are labeled according to the SAT1/KNP/96/91. Hydrogen bonds are indicated in yellow, dashed lines.

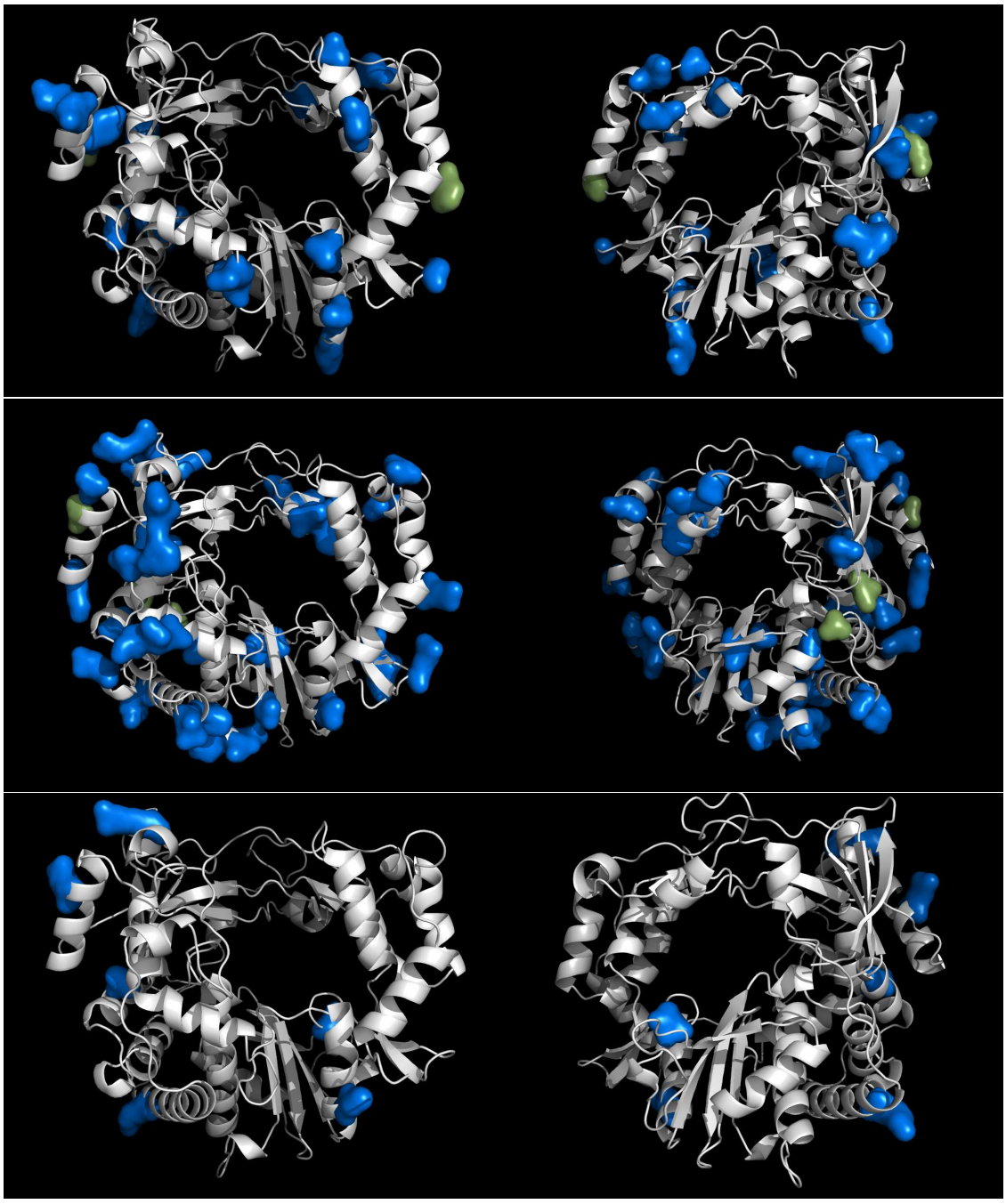

Figure 4.9: SAT 3D variation mapped onto a SAT 3D model. Views from both sides of the enzyme are shown. Top: SAT1, middle: SAT2, bottom: SAT3. White indicates conserved positions across all the sequences analyzed, blue indicates 2 different residues found at that position and green indicates 3 different residues.

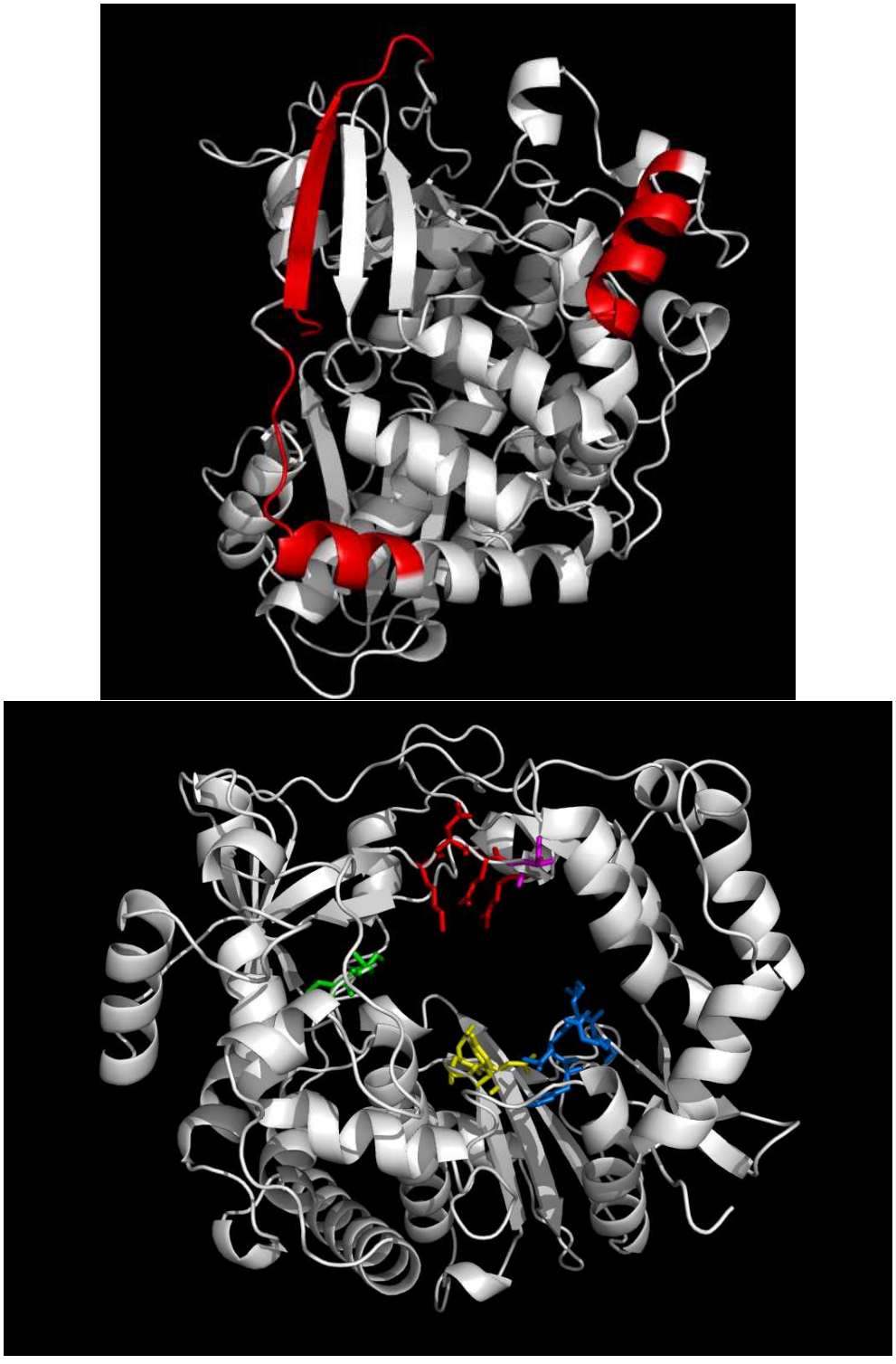

Figure 4.10: Top: The three hypervariable regions previously identified in 3D (George *et al.*, 2001). The regions coloured red and are residues 1-12 (β-strand), 64-76 (half α-helix and part of loop) and 143-153 (α-helix). Bottom: The four highly conserved motifs in 3D (Doherty *et al.*, 1999). The motifs are coloured as follows: red: KDELR; green: PSG; blue: FLKR; yellow: YGDD. The residue involved in mutation in the KDELR motif is coloured pink.

invariant region appears to be the orientation of the active site residues. The second β-strand (residues 85-91) in the invariant region associates with an adjacent β-strand (residues 40-45). This β-strand is followed by a a very short α-helix which is the location of the a second catalytic triad residue (His 46). It is involved in an extensive hydrogen bond network with two surrounding β-strands as well as with nearby residues. Figure 4.8 shows the hydrogen bond network in the region. The majority of the variable sites are involved in protein backbone hydrogen bonds. Thus, if the residue change does not involve a big physiochemical property change, it will not affect the backbone as much as the hydrogen bond network stays intact. This supports the hypothesis that the invariant region serves as an anchor region for the 3C protease. Thus, by conserving the invariant region's two β-strands, most of the active site residue orientation is also conserved.

SAT3/ZIM/5/91 showed a mutation in the active site where the Asp is converted to a Tyr. It has been previously proposed that a similar virus, Hepatitis A (HAV), may utilize a two-residue active site in 3C, which used only the Cys and His residues for catalysis (Bergmann *et al.*, 1997) but this has since been refuted (Yin *et al.*, 2005) and shown that HAV also uses a catalytic triad. This Asp-Tyr mutation has not yet been confirmed with resequencing.

In all 54 SAT 3C sequences analyzed, only one active site mutation occurred (D84Y in ZIM/5/91/3). In all the other sequences the catalytic triad and the residues surrounding them had very little, if any, variation. The analysis of the sequences showed that SAT2 3C had the most variation and that SAT3 had the least amount of variation.

4.3.2. 3D RNA Polymerase

The 3D RNA polymerase is highly conserved as mentioned before. The general sequence identity was 92% between the target and the template. This varied by no more that 1% between the three targets. The alignments used for each of the representative models are shown in Figure 4.4 and the high identity between target and template is indicated.

SAR/09/81/1 was used as the representative model for the SAT1 serotype. In the 9 SAT1 sequences provided there were 20 positions (91%) that had either one of two residues and

2 positions (9%) which had one of three residues (Fig. 4.9). The variation seemed to be limited to the outer edges of the protein.

ZIM/7/83/2 was used as the representative model for the SAT2 serotype (Fig. 4.9). SAT2 3D showed more variation compared to SAT1 and SAT3 3D. SAT2 3D had 38 positions (8%) with either one of two residues and three positions (0.8%) which had a three residue difference. This is almost double the variation seen in half the number of proteins when compared to SAT1 3D. This indicates that the 3D protein of SAT2 is more variable than that of SAT1 even though isolates from the same broad geographical region was included for both serotypes.

RSA/2/3 was used as the representative model for the SAT3 serotype (Fig. 4.9). A limited number of sequences made this serotype difficult to compare with SAT1 and SAT2. The three supplied proteins differed by two residues only in 6 positions (1.6%). The rest of the sequence was conserved.

3D variation did not seem to be limited to certain areas as seen for the 3C variation (Fig. 4.7). The results presented here suggests an average of 5% variable residues for 3D in each serotype. This is much lower than the other reported variability studies which reported variation as high as 26% variable residues (Carrillo *et al.*, 2005). This difference might be explained by the number of isolates in each serotype included in the studies as well as the geographical distribution. Intra and inter-serotype comparisons can also influence this value.

Three hypervariable regions in 3D have been identified previously (Fig. 4.10; George *et al.*, 2001). These areas did show some variability in the proteins analyzed here but it was mostly two residue differences between the proteins. The 3D hypervariable region, between residues 143-153, showed the most variability with four positions being variable. This area corresponds to a surface exposed α-helix. As can be expected, the variability are located on the exposed side of the α-helix. An α-helix important in inter-protein dimer interaction was identified from residue 68-89 (Ferrer-Orta *et al.*, 2004). The alignment of SAT 3D sequences revealed four residue positions that contained either one of two residues. The changes were located in two variable hot spots occurring at the ends of

the α-helix (two mutations per site), which still conserves the important central region involved in 3D dimer interaction.

Previously four conserved motifs were described in 3D polymerases of FMDV (Doherty *et al.*, 1999; Carrillo *et al.*, 2005). These four motifs are: KDELR (residues 159-163), PSG (residues 289-291), YGDD (residues 324-327) and FLKR (residues 371-374). The location of the conserved motifs can be seen in figure 4.10. Three of the motifs were also conserved in the SAT 3D sequences used here. However, the first motif, KDELR was present in the SAT sequences as either KDEIR or KDEVR. KDEIR was found to be conserved in all the SAT 3D sequences used except for SAT2/3KEN/21 that used the KDEVR motif. When looking at the orientation and location of the KDELR/KDEIR motif on the structure (Fig. 4.10) it is evident that the variable residue (L) is pointing away from the active site. The two mutations seen here (Leu->Ile, Leu->Val) are both similar in size and hydrophobicity, which maintain the physiochemical properties probably required for a residue in this location.

In comparison, the sequences used here showed that 3D also has less variation than $3C^{pro}$. The SAT 3D variation followed the trend seen in SAT $3C^{pro}$ where SAT2 had the most variation. This is explained by the fact that SAT2 is more prevalent in wildlife in Africa and has caused the most outbreaks. This results in an increased chance for variation accumulation in the genome, which can possibly be an indication of the age of the SAT2 serotype. If SAT2 was the ancestral SAT serotype, it would have acquired more variation over time. But without a detailed phylogenetic study of the relationship between the SAT types, this is pure speculation.

4.4. Conclusion

The replication of FMDV is dependent on several factors, including cell entry via receptors, replication of the RNA genome, translation, the correct polyprotein processing by viral encoded proteases, and packaging of the RNA into virions. A recent study investigated possible factors involved in the replication of SAT isolates which presented with diverse growth kinetics. The implication of this is in the implementation of engineered

virus to be used as custom-made vaccine specific for a geographic region. In principle infectious cDNA technology can be used to produce foot-and-mouth disease viruses with improved biological properties if the antigenic determinants of the outer capsid of a good vaccine strain with the desirable biological properties in a production plant are substituted by that of an outbreak isolate (Zibert *et al.*, 1990; Rieder *et al.*, 1993; Almeida *et al.*, 1998; Beard and Mason, 2000; van Rensburg *et al.*, 2004; Storey *et al.*, 2007). In practice we have found that the resulting chimera virus mostly took on the growth performance of the parental field isolate, although some improvement was observed by the presence of the better genetic background of the vaccine strain. Even with improvement of the cell entry pathway by introduction of alternative receptor entry mechanisms the growth performance was not significantly enhanced (Blignaut et al., unpublished; Maree, personal communication). To investigate whether these amino acid differences impact on the ability of the $3C^{pro}$ to recognise different cleavage sites within the P1 polyprotein, several chimeric viruses were engineered and the analysis of these are underway. In this study we investigated the amount of variation within the $3C^{pro}$ responsible for ten of the twelve proteolytic processing events of the FMDV polyprotein to support a present study on the amount of variation within the 3C cleavage sites and the activity of the enzyme within the cleavage site variation.

A study of the heterogeneity of the FMDV $3C^{pro}$ revealed 32% variant amino acid positions, whilst 57%, 65% and 75% variant amino acids were observed for the external capsid proteins (1B to 1D) (van Rensburg *et al.*, 2004). Similar to other picornaviral $3C^{pro}$, FMDV $3C^{pro}$ belongs to an unusual family of chymotrypsin-like cysteine proteases, containing a serine protease fold, as confirmed by the recently solved FMDV $3C^{pro}$ crystal structure (Birtley *et al.*, 2005). The catalytic mechanism of $3C^{pro}$ involves a Cys-His-Asp triad which has a very similar conformation to the Ser-His-Asp triad found in serine proteases. It is important to note that the third member of the triad is also an Asp residue in HAV, but a Glu in HRV (Curry *et al.*, 2007). The FMDV 3C ^{pro} cleavage specificity exhibits great heterogeneity, but similar to other picornaviral $3C^{pro}$, the enzyme requires a hydrophobic residue at P4 (Curry *et al.*, 2007). Whereas other picornavirus 3C proteases accept only Gln at the P1 position, the FMDV $3C^{pro}$ differs in that it is able to accept

both Gln and Glu in this position. It has been suggested that correlations between the different sub-sites in the substrate binding pocket of $3C^{pro}$ exist. By analysing FMDV sequences (Carrillo *et al.*, 2005), Curry and co-workers (2007) suggested correlations between P1, P2 and P1'. For instance, if P1 is a Gln, P2 would usually be a Lys and P1' a hydrophobic residue. Small amino acids (Gly or Ser) are however present in the P1' position for all the viruses analysed when P1 is Glu. Important roles for P2 and P4' have also been implicated (Birtley *et al.*, 2005).

In addition to processing of the viral polyprotein, $3C^{pro}$ has been shown to cleave host cell proteins in cell culture. Cleavage of histone H3, resulting in a down-regulation of transcription, has been demonstrated (Falk *et al.*, 1990; Tesar and Marquardt, 1990), although an unusual cleavage site was suggested. The enzyme has also been reported to cleave host cell translation initiation proteins, eIF4G and eIF4A (Belsham and Sonenberg, 2000; Li *et al.*, 2001; Strong and Belsham, 2004). These cleavage events occur rather late in the infection cycle and their role in viral replication is unclear. A recent report indicated that PTB, eIF3a,b and PABP RNA-binding proteins are cleaved during FMDV infection in cell culture, although no evidence for $3C^{pro}$ involvement was established (Pulido *et al.*, 2007).

Mapping the variation found within 53 SAT viruses representative across Africa onto the $3C^{pro}$ structure reveals that these are almost entirely peripheral to the substrate-binding site, supportive to previous finding by Birtley *et al.* (2005). There was some variation close-by the active site in the invariant region but all the variation still preserved the backbone hydrogen bond structure needed to keep the catalytic triad in the correct conformation for catalysis. This emphasizes the highly conserved nature of $3C^{pro}$ and the likeliness that chimeric viruses containing the outer capsid region of a disparate virus within the genetic background of an existing SAT2 genome-length clone (van Rensburg *et al.*, 2004) will be processed by the SAT2 $3C^{pro}$. The rate of processing might however be influenced by the sequence variation within the 3C cleavage sites in the P1 polyprotein.

The 3D RdRp is extremely conserved and is needed for virus replication. All of the variation were seen to occur outside of the binding cavity (Fig. 4.9) in the central part of the enzyme. Some of the variation may influence the activity of 3D but this study

found that the majority of the differences are natural variation. The few differences in the invariant regions (KDEI/V/LR) were found not to significantly influence the overall activity as they have similar physiochemical properties. Another factor was that the side chains of the different residues in the invariant regions pointed away from the active site. All the variation seen in the different serotypes may have a small effect on the activity of the enzymes or on interaction cellular proteins, and this in turn could affect the replication speed of the virus. The variation may simply be a result of natural variation in SAT serotype enzymes. After analysis of the models and variation, there does not appear to be a reasonable site where 3C-3D interaction occurs. Although 3C presents an area on the C-terminal β-barrel where there is almost no variation, it does not necessarily imply an interaction site. 3D has a flattish area on the protein which, although it is sometimes used in protein-protein interaction, is not conclusive proof of an interaction site. The crystal structure of polio 3CD has been published (Marcotte *et al.*, 2007) but upon analysis it was found that the crystal structure provides no evidence for the interaction between 3C and 3D as they are separated by a 7-residue linker region. Further studies into co-variation was not done as it falls outside the scope of this specific study. The variation seen in 3C confirms the conserved nature of 3C yet it highlights that the variation that does occur, are limited to certain areas. Chapter 5 investigates the effect of variation on the capsid protein stability and its structure.