CHAPTER 4

SECONDARY STRUCTURE PREDICTION OF DEN-2 PROTEASE

4.1 Secondary structure prediction

Several studies have been carried out by researchers to gain insights into the protease complex structure of Dengue virus type 2 (DEN-2) (The 2nd International Conference on Dengue and Dengue Haemorrhagic Fever, 2008; Wahab *et al.*, 2007). Until very recently (D'Arcy *et al.*, 2006; Erbel *et al.*, 2006), there was no data on the crystal structure of DEN-2 NS2B-NS3 protease complex. The closest structure to the DEN-2 protease was a homology model of the protease complex built by Brinkworth *et al.* (1999). This homology model used the crystal structure coordinates of the hepatitis C virus NS3-NS4A as template (PDBid: 1JXP). However, the overall identity between the two sequences is only about 14.8% although regions surrounding the putative catalytic residues, as defined by Bazan and Fletterick (1989), indicated a high level of identity. The lack of structural details for the active DEN-2 protease from experiments did not offer substantial insights into its interaction with substrates. Thus, the design for the protease inhibitor was based mainly on either kinetic studies, such as that reported by Tan *et al.* (2006) or theoretical understandings from *in silico* simulations (Brinkworth *et al.*, 1999; Lee *et al.*, 2006).

This chapter describes the secondary structure prediction study of 175 Nterminal amino acids of the DEN-2 NS3 protease using a combination of several prediction tools available over the website. This work was carried out before the experimental attempts to crystallize the protease complex (Chapter 3) and before the report on the DEN-2 NS2B-NS3 crystal structure (D'Arcy *et al.*, 2006; Erbel *et al.*, 2006) was published. The aim of this work was to map out the secondary structure of the different regions in the protease with the knowledge of structurally conserved regions obtained from multiple sequence alignment against NS3 proteases of other viruses from the Flaviviridae family. The structural data obtained from the recently crystallized protease complex had enabled us to make an evaluation of the prediction results and of the predictive power of the *in silico* methods employed. The work described in this chapter has been published (Othman *et al.*, 2007).

4.2 Materials and Methods

The protocols for this study are as illustrated in the flowchart in Figure 4.1.

4.2.1 Multiple Sequence Alignment

Protein sequence alignments and comparisons were carried out using the BLAST (Basic Local Alignment Search Tool) program, blastp, against database specification of non-redundant protein which were available from the National Center for Biotechnology Information (NCBI) Web server, (Altschul *et al.*, 1997); http://www.ncbi.nlm.nih.gov/blast/). Viruses for proteases used in this study were checked against the Universal Virus Database, ICTVdb (International Committee on Taxonomy of Viruses) (Büchen-Osmond, 2003). Amino acid sequences were obtained from NCBI sequence Viewer 2.0, available at http://www.ncbi.nlm.nih.gov. Multiple sequence alignments were done using ClustalW 1.82 (Thompson *et al.*, 1994) available at the European Bioinformatic Institute (EBI) Web server. Consensus of amino acid sequence was obtained from Boxshade available at the European Molecular Biology Network (EMBnet) Web server (http://www.ch.embnet.org).



Figure 4.1 Flowchart of protocols involved in the secondary structure prediction study of DEN-2 NS3 protease.

4.2.2 Secondary Structure Prediction

The 175 amino acid sequence of DEN-2 NS3 protease was submitted for automated prediction of secondary structures to the following programs via their web site interfaces: PSIPRED (Bryson *et al.*, 2005; Jones, 1999; McGuffin *et al.*, 2000); http://bioinf.cs.ucl.ac.uk/psipred/), PROF on PredictProtein Web server (Rost *et al.*, 2004); http://www.predictprotein.org), PHD on PredictProtein Web server (Rost, 1996), APSSP2 (Raghava, 2002); http://www.imtech.res.in/raghava/apssp2/) and Jnet on JpredWeb server (Cuff & Barton, 2000); http://www.compbio.dundee.ac.uk/~www-jpred/jnet/).

Threading programs used via their web site interfaces were: 123D+ (Alexandrov et al., 1995); http://123d.ncifcrf.gov/123D+.html), 3D-PSSM (Kelley et al., 2000); http://www.sbg.bio.ic.ac.uk/servers/3dpssm/) and LOOPP (Meller & Elber, 2001; Teodorescu *et al.*, 2004); http://cbsuapps.tc.cornell.edu/loopp.aspx). The secondary structure information of some templates used in the threading programs were obtained DSSP from database available online (Kabsch & Sander, 1983); http://swift.cmbi.ru.nl/gv/dssp/). The templates were chosen based on E-values (E-value < 0.05, highly confident; E-value ≤ 1.00 , worthy of attention) or z-scores (z-score > 3, high confidence).

Two approaches were employed for the one dimensional (1D) secondary structure predictions performed in this study:

Approach 1. Utilising available structure prediction servers only. Alignment of secondary structures obtained from prediction programs was performed to result in a consensus.

77

Approach 2. Gaining information on the secondary structures extracted from structure prediction servers, threading techniques and DSSP database of some of the templates used in the threading techniques. An alignment of all secondary structures obtained was performed to result in another consensus.

The results obtained from the two tool sets (Approach 1 and Approach 2) were then compared to the observed secondary structure of the recently crystallized NS2B-NS3 obtained from the Protein Data Bank (Berman *et al.*, 2000) (PDBid: 2FOM). The percentage accuracy, A, was calculated as follows:

$$\mathbf{A} = \frac{\mathbf{c}}{\mathbf{N}} \ge 100 \%$$

where c is the number of residues predicted correctly, and N is the total number of residues with predicted secondary structures aligned against structures from crystal data.

Percentage difference, D, in secondary structure between both approaches was calculated as follows:

$$\mathbf{D} = \frac{\mathbf{d}}{\mathbf{N}_{t}} \ge 100 \%$$

where d is the number of residues with different secondary structure from both approaches, and N_t is the total number of residues.

The implementation of Approach 2 was carried out to ascertain if a more exhaustive technique of secondary structure prediction would result in a more reliable result.

4.3 Results

Figure 4.2 illustrates the multiple sequence alignment of DEN-2 NS3 protease (EC 3.4.21.91) and other proteases belonging to the genus Flavivirus of the Flaviviridae family. Consensus of sequences obtained from the Boxshade program is shown and regions surrounding the catalytic triads were observed to be highly conserved (Bazan & Fletterick, 1989; Brinkworth et al., 1999). The secondary structure profiles that were built following the two approaches (as stated in the materials and methods section) are illustrated in Figures 4.3 and 4.4. Consensus of the secondary structures is given as E (β -strand), H (α -helix) or C (coil). It can be seen that both approaches yielded mainly β -strands, with Approach 1 resulting in 39.43 % of β -strands from the whole structure, while from Approach 2, 40 % of β -strands was observed from the whole structure. These results are in accordance with the fact that DEN-2 NS3 protease belongs to the fold family of trypsin-like serine-proteases (S07.001; according to MEROPS) (Rawlings *et al.*, 2006), which falls into the all- β proteins class of protein structure (according to SCOP) (Bazan & Fletterick, 1989; Murzin *et al.*, 1995).

The prediction results obtained from the two approaches were compared with the secondary structure of crystallized NS3 protease of DEN-2 (2FOM) to evaluate the predictive power of each approach. Difference in secondary structure between both approaches was calculated to be 7.43 %. Results showed Approach 2 to yield higher accuracy (A = 76 %) compared to the use of prediction servers only (Approach 1; A =

									A	
			10		20		30	40		50
DEN-2	1-49	AGVLWDV	PSPPPV	GKAE-I	EDG	AYRIKÇ	KGILGY	SQIGAGV	YKEGTFHT	М
DEN-1		SGVLWDT	PSPPEVI	ERAV-I	DDGI	LANILO	RGLLGR	sq <mark>v</mark> gvgv	FQEGVENT	М
DEN-3		SGVLWDV	PSPPET	OKAE-I	EEG	/YRIKC	OGIFGK	TOVGVGV	OKEGVENT	М
DEN-4		SGALWDV	PSPAATI	KKAA-I	SEG	/YRTMC	RGLEGK	TOVGVGT	HMEGVEHT	м
KIIN		CCVLWDT	POPKEVI	KRCD-1		IVRIMI	RGLLGS	VOAGAGV	MURCUEHT	т.
TATALY 7		COVINDI				AND TMU		VONCACU	MURCURUM	-
WINV		GGVLWDI	PSPREI	NNGD-1		/IRIMI	RGLLGS	IQAGAGV	MVLGVFHI	ц. т
JEV		GGVFWDT	PSPKPCA	AKGD-1	L.L.L.C.	/YRIMA	RGILGT	YQAGVGV	MYESVEHT	ь
SLEV						I	RGILGT	FQAGVGV	MHEGVEHT	М
YFV		GDVLWDI	PTPKII	EECEHI	EDGI	EYGIFÇ	STFLGA	SQRGVGV	AQGGVFHT	М
Consen	sus	agvlwd p	psp	kae 1	L dgi	yrilo	rgilG	sQ GvGv	egvFHT	m
(Boxsh	lade)									-
					L.			1		
		#	60		70	B #	80	90		100
DEN-2	50-99	WHVTRGA		KRTEPS	WAD	י.ד <u>חאא</u>	SYGGGW	KLEGEWK	EGEEVOVI.	Δ
DDIN 2	30 33	1110 110011				,	010000			
DEN-1		MUUTPCA				ד דחאאז	SACCOM	FOCSMIN	ACEEVOUT	Л
DEN 1		WINTROA					CVCCCM	DI CAOMO	KGEEVQVI.	2
DEN-3		WHVIRGA	V L I I NGI	REEPI	WAS		SIGGGW	RLSAQWQ	NGEEVQVI.	A
DEN-4		WHVTRGS	VICHETO	GRLEPS	SWAL	/RNDM1	SYGGGW	REGDKWD	KEEDVQVL.	A
KUN		WHTTKGA	ALMSGE	GRLDPY	wgs	/KEDRI	CYGGPW	KLQHKWN	GQDEVQMI	V
WNV		WHTTKGA	ALMSGE	GRLDPY	WGS	/KEDRI	CYGGPW	KLQHKWN	GQDEVQMI	V
JEV		WHTTRGA	AIMSGE	GKLTPY	WGS	/KEDRI	AYGGPW	RFDRKWN	GTDDVQVI	V
SLEV		WHATEGA	VLRNGE(GRLDPY	(AGD	/RNDLI	SYGGPW	KLSATWD	GTEEVQMI.	А
YFV		WHVTRGA	FLVRNGI	KKLIPS	WAS	/KEDLV	AYGGSW	KLEGRWD	GEEEVQLI.	А
Consen	isus	WHvTrGa	vlm d	grleP	was\	/k Dli	sygg W	kl gkWn	g eeVQvi	а
(Boxsh	lade)			-	L]	-	
										-
			110		120	C	130	# 140	:	150
DEN-2	100-148	LEPGKNPI	RAVQTKI	PGLFKT	rn a g-	-TIGAV	SLDFSP	GTSGSPI	IDKKGKVV	G
DEN-1		VEPGKNPI	KNVOTA	PGTFKI	PEG-	-EVGAI	ALDFKP	GTSGSPI	VNREGKIV	G
DEN-3		VEPGKNPI	KNFOTMI	PGTFOT	rTTG-	-ETGAT	ALDEKP	GTSGSPT	TNREGKVV	G
DEN-4		TEPRKNP	KHVOTKI	PSLEK	T.TG-	ETGAL	TLDFKP	GTSGSPT	TNRKGKVT	G
KIIN		VEPCKNU	XNVOTK	PCVERT	PPC-	FICAL		GTSGSPI	VDKNCDVT	ĉ
WNIV		VEPCKNW		DCVERT	PPC-			CTRCSPT	VDKNCDVT	c
757		VEPCKAA						CTRCCPT	IDSNCDIT	c
CIEV		VERGINAN						CTRCODI	INVERT	c
VEV		VALGREA.			TDNCC			GISGSFI	INNIGHT	G
IFV		AVPGRNV		PSLEAN	RNGC		ALDIPS	GISGSPI	VNRNGEVI	G
Consen	isus	vergkii	KIIVQIKI	PGIFKI	-p G	elGAV	storp	GISGSPI	TUK GUVI	G
(Boxsh	lade)									—
		_								
DEN-2	140-175		TPSCAV		170 \TTTT					
DDN Z	149 175	LIGNOVV.	INDOMI	Voning	2 т ш тс					
DEN-1		TYCNOW	rmeenv		אאת					
DEN 1		LIGNGVV	TISGII	VOAIAG			1			
DEN-3		LIGNGVV	INGGI	VOGIA		TODD2				
DEN-4		LYGNGVV	TKSGDY	VSALT	JAER-	-IGEPL)			
KUN		LYGNGVI	MPNGSY	ISAIV(ACCURENTS N		,			
WNV		TVCNCUT			JGERP					
		LIGNGVI	MPNGSY	ISAIVÇ)GERN)GERN	ADEPIE)			
JEV		LYGNGVE	MPNGSY: LGDGSY	ISAIV(VSAIV(QGERN QGERN QGDR(ADEPIE QEEPIE	5 5			
JEV SLEV		LYGNGVI LYGNGVE LYGNGVL	MPNGSY: LGDGSY LGQG-Y	ISAIV(VSAIV(VSGII(QGERN QGERN QGDR(QGERJ	ADEPIE QEEPIE CEEPIE				
JEV SLEV YFV		LYGNGVI LYGNGVE LYGNGVL LYGNGIL	MPNGSY: LGDGSY IGQG-Y VGDNSFY	ISAIV VSAIV VSGII VSGII VSAIS	QGERN QGERN QGDR(QGER1 QTEVN	ADEPIE QEEPIE CEEPIE KEEGKE				
JEV SLEV YFV Consen	ISUS	LYGNGVI LYGNGVI LYGNGVI LYGNGII LYGNGVV	MPNGSY: LGDGSY IGQG-Y VGDNSF n gty	ISAIV(VSAIV(VSGII(VSAIS(VSAIS(QGERN QGDR(QGDR(QGER1 QTEVN Qaer	ADEPIE QEEPIE TEEPIE KEEGKE de				

Figure 4.2 Multiple sequence alignment of DEN-2 NS3 protease and other proteases belonging to the genus Flavivirus. The catalytic triad residues: His51, Asp75 and Ser135, are labelled with the symbol # and are found in boxes labelled A, B and C. (The numbers in the diagram may not represent the exact positions of residues in the actual protein due to the insertion of gaps during alignment). The boxes labelled A, B, C and D identify regions of significant similarity surrounding the catalytic triad residues and residues that might form the substrate-binding pocket (Bazan & Fletterick, 1989). Consensus of residues throughout the different viruses is indicated by Boxshade, with those having 100% similarity written in the upper case font. Sequences used in the analyses: DEN-1 (Dengue virus type 1), DEN-3 (Dengue virus type 3), DEN-4 (Dengue virus type 4), KUN (Kunjin virus), WNV (West Nile virus), JEV (Japanese encephalitis virus), SLEV (St. Louis encephalitis virus) and YFV (Yellow fever virus).

		10		20	30		40		50 #		60		70	#	80		90
DEN-2 1-50	AGVLWDV	/PSPPPVC	GKAELE	DGAYRI	KQKGIL	GYSQI	G <mark>A</mark> GVY	KEGTF.	HTMWH	VTRGA	VLMH	KGKRIE	PSWAD	KKDL	SYGGO	WKLE	GEWK
Consensus (Boxshade)	agvlwd	psp	kael	dgvyri	lqrgil	G sQ	GvGv	egvF	HTm W H	vTrGa	vlm	grle	eP was∖	/k Dl:	LsYGG	Wkl	gk₩n
PHD	LL EE I	LLLL	L	ΕE	EEEEE	EEE	EEEE	EEEEE	EEEEE	L	EE	LLLLL	LLL		LLL	LLLL	LLLL
Psipred	CCCCCCC	cccccc	ccccc	CCEEEE	EECCCC	CCEEE	EEEEE	ECCEEI	EEEEC	ccccc	ннсс	CCCEEE	EECCCC	CCCEE	EECCC	сссс	CCCC
PROF	LLLLLL	LLLLLI	LLLLL	LLEEEE	EEEEEE	LLEEE	EEEEE	ELLEE	EEEEEI	LLLLL	EEEL	LLLEEL	LLLLL	LLLEE	ELLLL	LLLL	LLLL
APSSP2	CCEECCC	СССССНЕ	ннссс	CCCCCE	CCCCCE	CCCCE	CCEEE	ECCEEI	EECEEF	ECCCC	CCCC	cccccc	ccccc	CCCEE	EECCC	CCEC	ECCC
Jnet	CCCEEEC		ccccc	CCEEEE	EEEEEE	EEEEE	EEEEE	EEEEE	EEEEE	ECCCC	EEEC	CCCCEE	EEECCC	CCEEE	ECCCC	CCCC	CCCC
Consensus 2° struct	CCCECCC		ccccc	CCEEEE	EEEEEE	CCEEE	EEEEE	ECCEEI	EEEEE(ccccc	EECC	CCCCEC	cccccc	CCCEE	ECCCC	cccc	CCCC
DEN-2 91-175	5 EGEE <mark>V</mark> QV	100 LALEPGF	NPRAV	110 QTKPGL	120 FKTNAG	TIGAV	130 SLDFS	# PGTSG:	140 SPIIDE	KKGK V	150 VGLY	gng <mark>vv</mark> i	160 RSGAY	VSAIA	170 QTEKSI	EDN	
Consensus (Boxshade)	g eeVQv	viavePgŀ	Kn knv	QTkPgl	.Fktp G	eiGAv	sLDfp	GTSG	SPIin	k Gdv	iGLY	GNGvvn	ı gtyv	'SaI ()aer d	е	
PHD	LLLEEEE	EEE LLI	LL	LLLLE	EEELLL	E EE	LLL	LLLLI	LL LI	LLL E	EEE	EEE	EEEE		LLLLL	LLL	
Psipred	CCCCEEE	EEECCCC	CEEEE	EEECCE	EECCCC	CCCCE	ECCCC	ccccc	CCEEC	CCCCE	EEEE	CCCEEE	CCCCEE	CEEEEE	EEEEC	CCC	
PROF	LLLLEEE	EEEELLI	LEEEE	ELLLLE	EEELLL	EEEEE	EELLL	LLLL	LLEELI	LLLLE	EEEE	ELLEEE	EELEEE	EEEEE	ELLLLI	LLL	
APSSP2	CCCCEEE	ECCCECC	ccccc	CCCCEC	CEECCC	EECCC	CCCCC	CCCCC	CCEEE	CCCCE	EEEE	EEEEEC	CCCCEE	EEEEE	сннннс	CCC	
Jnet	CCCCEEE	EEECCCC	CCEEEE	EEECCC	cccccc	CCCCC	CCCCC	CCCCC	CCEEC	CCCCE	EEEE	EEEEEE	CCCEEE	EEEEE	EECCC	CCC	
Consensus 2° struct	CCCCEEE	EEECCCC	CEEEE	ECCCCE	EEECCC	CECCE	ccccc	CCCCC	CCEEC	CCCCE	EEEE	ECCEEE	CCCEEE	EEEEE	ccccc	CCC	

Profile of 1D secondary structure prediction of DEN-2 NS3 protease based on automated prediction programme. Programs used for Figure 4.3 secondary structure prediction are: PHD on PredictProtein Web server (Rost, 1996), PSIPRED (Bryson et al., 2005; Jones, 1999; McGuffin et al., 2000); http://bioinf.cs.ucl.ac.uk/psipred/), PROF on PredictProtein Web server (Rost et al., 2004); http://www.predictprotein.org), APSSP2 (Raghava, 2002); http://www.imtech.res.in/raghava/apssp2/) and Jnet on JpredWeb server (Cuff & Barton, 2000); http://www.compbio.dundee.ac.uk/~wwwjpred/jnet/). Consensus of secondary structure are given as $E = \beta$ -strand, $H = \alpha$ -helix or C = coil.

	10 20 30 40 50π 60 70π 80 90
DEN-2 1-90	AGVLWDVPSPPPVGKAELEDGAYRIKQKGILGYSQIGAGVYKEGTFHTMWHVTRGAVLMHKGKRIEPSWADVKKDLISYGGGWKLEGEWK
Consensus	agvlwd psp kael dgvyrilqrgilG sQ GvGv egvFHTmWHvTrGavlm grleP wasVk DlisYGG Wkl gkWn
(Boxshade)	
PHD	LL EE LLLLL L E EEEEEEE EEE EEEEEEEEEE
Psipred	CCCCCCCCCCCCCCCCCCCCEEEEEEEEEEEEEEEEEE
PROF	LLLLLLLLLLLLLLLLLEEEEEEEEEEEEEEEEEEEEE
APSSP2	CCEECCCCCCCCHHHHCCCCCCCCCCCCCCCCCCCCCCC
Jnet	CCCEEECCCCCCCCCCCCCCEEEEEEEEEEEEEEEEEE
123D+	EE E EHHHHHHHHHHH EE E E EE EEEEHHHHHHHH
3D-PSSM	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEE
LOOPP	CCCEECCCCCCCCCCCCCCEEEEEEEEEEEEEEEEEEEE
1 d. 0	
layp	
15Vp	
1 vcp	
10n5	R C FFFFF FTT FFFFFFFFFFTT CFULL FFFCCCFFF LTT R R
	D 5 EEBEE EIT EEEEBEEEETT SEMMI EEEGGGEEE MIT D D HUHHHHHHHHHHFFFTTFFFFFFFFFFFFFFFFFFFFFF
тсдд	
CONSENSUS	
2 FOM	
DEN-2 1-90	ACVI.WDVPSPPPVCKAEI.EDCAYRIKOKCII.CYSOICACVYKECTEHTMWHVTRCAVI.MHKCKRIEPSWADVKKDI.ISYCCCWKI.ECEWK
	$10 \qquad 20 \qquad 30 \qquad 40 \qquad 50^{\text{\text{H}}} \qquad 60 \qquad 70 \qquad \text{\text{H}} \qquad 80 \qquad 90$

ш

Figure 4.4 Profile of 1D secondary structure prediction of DEN-2 NS3 protease based on combinations of results from automated prediction programme, threading programme and secondary structure of templates used in threading method obtained from DSSP database. Template sequences used in the analyses: 1dy9 (Hepatitis C virus NS3 protease/helicase), 1svp (Sindbid virus capsid protein), 1vcp (Semiliki Forest virus capsid protein), 1jxp (Human Hepatitis C virus NS3 protease), 1ep5 (Venezuelan Equine Encephalitis virus capsid protein), and 1cqq (Type 2 Rhinovirus 3C protease). Codes for secondary structures are given by: $E = extended \beta$ -strand, B = residue in isolated β -bridge, $H = \alpha$ -helix, G = 3 / 10 helix, T = hydrogen-bonded turn, S = bend, L = loop and C = coil (Kabsch and Sander, 1983). Regions with no codings are noted as coils. Consensus of secondary structure is given as E, H or C. 2FOM: secondary structure of crystallized DEN-2 NS3 protease available on PDB.

Chapter 4

	100	1:	0	120	130 #	140	150	160	170		
DEN-2 91-175 EGEEVQVLALEPGKNPRAVQTKPGLFKTNAGTIGAVSLDFSPGTSGSPIIDKKGKVVGLYGNGVVTRSGAYVSAIAQTEKSIEDN											
Consensus	g eeVQviave	PgKn knvQ	TkPglFktp	GeiGAvsL	Dfp GTSG	SPIink G	dviGLYGNGvv	m gtyvSaI	Qaer de		
(Boxshade)											
PHD	LLLEEEEEEE	LLLL	LLLLEEEL	LL E EE 🗄	LLLLLLL	LL LLLL	EEEE EE	E EEEE	LLLLLLL	٦	
Psipred	CCCCEEEEEC	CCCCEEEEE	EECCEEECC	CCCCCEEC	ccccccc	CCEECCCC	CEEEEECCCEE	ECCCCEEEEE	EEEEEECCCC		2° struct.
PROF	LLLLEEEEEEE	LLLLEEEEE	LLLLEEEL	LLEEEEEEE	LLLLLLL	LLEELLLL	LEEEEEELLEE	EEELEEEEEE	EELLLLLLL	Ļ	prediction
APSSP2	CCCCEEEECCC	ECCCCCCCC	CCCECCEEC	CCEECCCCC	ccccccc	CCEEECCC	CEEEEEEEEE	CCCCCEEEEE	ЕЕННННСССС	(
Jnet	CCCCEEEEEC	CCCCEEEEE	EECCCCCCC	cccccccc	ccccccc	CCEECCCC	CEEEEEEEEE	ECCCEEEEE	EEEECCCCCC		
)	
123D+	нннннененн	I EE	EE	EEEEEE	EE	EE	EEEEEE E	се ееннн	іннн ннн н	٦	
3D-PSSM	CCCEEEEEEC	CCCCEEEEE	ECCCEEECC	CCCEEEEC	ccccccc	CCEECCCC	CEEEEECCCEE	ECCCCEEEEE	EECCCCCCCC	Ļ	Threading
LOOPP	CCCEEEEEEC	CCCCCEEEE	ECCCCEEEC	CCCEEEEC	ccccccc	CCEEECCC	CEEEEEECCEE	ECCCEEEEE	EECCCCCCCC	ſ	
										J	
1dy9	S EEEEE	TT EEEEE	EEE TTE	EEEEEEG	GGTTT TT	EEE TTS	EEEEEEEE	ETTEE G		~	
1svp	GGGSTT B B	S SEEEE	ETTEEEEE	TTEEEETT	S TT TT	EEE TTS	EEEEEEEE	SSEEEEEE	EEEE TT		DSSP
1vcp	TTTTEEEEE	S SEEEE	ETTEEEEE	TTEEEETT	S TT TT	EEE TTS	EEEEEEEE	SSEEEEEE	EE SS TT		Template
1jxp	E EEEE	TT EEEEE	E S SS	EEEEEEG	GGGTT TT	EEEETTT	EEEEEEEEE	E TT EEEE	E	7	2° struct.
1ep5		SETTEEEE	EETTEEEEE	ТТ	TT TT	EEE TTS	EEEEEEEE	SSEEEEEE	EE TTT B		
1cqq	EEEEEEE T	SSS EEEE	E EEETT	E SEEEE	TT TT	EEETT	EEEEEE S	S EEEEE	GGG		
)	
CONSENSUS	CCCCEEEEEEC	CCCCEEEEE	ECCCEEEEC	CCCEEEEEC	ccccccc	CCEECCCC	CEEEEEEEEE	ECCCEEEEEE	EECCCCCCCC		
2FOM	CCCCEEEEECC	CCCCCEEEE	ECCEEEECC	CCEEEEECC	сснннннс	CEEECCCC	CEEEECCCEEE	ECCCCCEEEEE	EC		
DEN-2 91-175	5 EGEEVQVLALE	PGKNPRAVÇ	TKPGLFKTN.	AGTIGAVSL	DFSPGTSG	SPII <mark>D</mark> KKG	KVVGLYGNGVV	TRSGAYVSAI	AQTEKSIEDN		
	100	11	0	120	130 #	140	150	160	170		

Figure 4.4 (continue)

Chapter 4

72.67 %). This showed that Approach 2 gave a more reliable 1D secondary structure profile.

4.4 Discussion

In contrast to the suggestions made by Russell (2002), the techniques used in this study do not involve prior knowledge of protein structure to reach the consensus. With only a 'palmful' of knowledge on the DEN-2 NS3 protein structure during the early part of this study, the accuracy of the prediction results to the true secondary profile of the system was questionable. However, the recently published crystal model of DEN-2 NS2B-NS3 protease (Erbel *et al.*, 2006), which resolved to 1.5 Å, assured the quality of the prediction results (from Approach 2) at 76% accuracy, which is reasonably high. Presumably, this approach for building secondary structure profile of virus, i.e. Flaviviridae (which can further be classified into 3 genus). Nevertheless, the percentage of accuracy of the prediction results compared with the true structure will also depend on the prediction tools chosen. The selection of the appropriate prediction tools can be determined by referring to CASP (the Critical Assessment of Structure Prediction) (Bourne, 2003) experiments.

The crystal structure of DEN-2 NS2B-NS3 revealed that the protease adopt a β barrel fold (Erbel *et al.*, 2006). It was also reported that residues 51–57 of the cofactor NS2B contributed one β -strand to the N-terminal β -barrel. Furthermore, the crystal structure of West Nile virus NS2B-NS3, complexed with inhibitor, showed direct interactions of NS2B with active site of NS3, underlining the work by Yusof *et al.* (2000), which indicated the dependence of the protease on the cofactor for cleavage of substrates with dibasic amino acids. However, this study only concentrated on the secondary structure of the protease domain (NS3) and it was assumed, at this stage, that the involvement (or not) of NS2B cofactor in the prediction will not affect the 1D secondary structure profile of the NS3 protease, since alignment of the protein was made with other proteins of the same domain (when using prediction tools).

Analysis of conservation in the protein families is effective in secondary structure predictions performed before the knowledge of the protein structure was obtained (Cuff & Barton, 1999). A strong correlation can be made between structurally conserved regions (Figure 4.4) and regions with highly conserved sequences across the different proteases (Figure 4.2), particularly the regions surrounding the catalytic triad. However, two regions were missed by the prediction methods; i.e., where 2FOM defined residues Trp50-Arg54 and Ser131-Ser135 (each carrying residues comprising the catalytic triad, i.e. His51 and Ser135, respectively) to be α -helices. On the other hand, it is quite unexpected for α -helix to span the region Ser131-Ser135 since this region also included Pro132, and proline is well-known to be a 'helix-breaker' due to its rigid ring conformation (Garrett & Grisham, 1997a). Looking at the prediction results in Figure 4.4 for the region Ser131-Ser135, there is 100% consensus for this region to be coiled. Perhaps the α -helices could be recognised after the secondary structure profile is put through fold recognition procedures.

4.5 Conclusion

This study illustrated the application of both automated prediction servers and information of secondary structures from threading programs, to build a consensus of secondary structure (1D) of DEN-2 protease. In the early stages, the secondary structure

predictions were carried out prior to the PDB deposition of the crystal structure (2FOM) for NS2B-NS3. Soon after the published report of the 3D-structure in the year 2006, a validation of the predictive power of the tools was conducted in this study. The conclusions were drawn by comparing a combination of tools (Approaches 1 and 2) against the observed secondary structure of 2FOM. The consensus obtained in the comparison studies showed higher similarity in Approach 2 than Approach 1 to 2FOM (Figure 4.5). In the present case and possibly other cases of low homology sequence relationships, a significantly better prediction could be attributed to Approach 2 since this approach comprises heuristic (similarity of amino acid properties), probabilistic (from structural data collections) and more sophisticated overlaying techniques, i.e. threading the query into the template backbone to detect hidden phylogenetic resemblance.

	1	10	20	30	40	50#	60	70	4 80	90
DEN-2 1-90	AGVLWDV	PSPPPVGKA	ELEDGAYRIF	(QKGILGYSQ	IGAGVYKEG	TFHTMWHVTRO	AVLMHKGKRI	PSWADVKKI	DLISYGGGWK	LEGEWK
Approach 1	CCCECCC	cccccccc	CCCC <u>C</u> EEEEE	EEEE <u>EECC</u> EE	EEEEEECC	EEEEE <u>EECCC</u> (CCEECCCCCEC	ccccccccc	CEEE <u>C</u> CCCCC	cccccc
Approach 2	cccccc		CCCCCEEEEE	ECCCCCCEE	EEEEEECC	EEEEEECCCC	CEEECCCEEEC	cccccccc	CEEECCCCCC	cccccc
2FOM			CCCEEEEEE	EEECCEEEE	EEEEEECC	ЕЕЕЕЕНННННО	CCEEECCEEEC	EEEEECCCO	CEEEECCCCC	cccccc
	91	100	110	120	130	# 140	150	160	170	
DEN-2 91-175	EGEEVQV	LALEPGKNE	RAVQTKPGLE	KTNAGTIGA	VSLDFSPGT	SGSPIIDKKGP	VVGLYGNGVV	RSGAYVSA	IAQTEKSIED	N
Approach 1	CCCCEEE	EEECCCCCE	EEEECCCCEE	EEECCCCECC	ECCCCCCCC	CCCCEECCCC	CEEEEEECCEEE	ECCCEEEEEI	EEECCCCCCC	C
Approach 2	CCCCEEE	EEECCCCCE	EEEEECCCEE	EECCCCEEE	EECCCCCCC	CCCCEECCCC	EEEEEEEEEE	CCCEEEEEI	EEECCCCCCC	С
2FOM	CCCCEEE	EECCCCCC	EEEEECCEEE	ECCCCEEEE	ЕССССНННН	HCCEEECCCCC	CEEEECCCEEEC	CCCCEEEEI	EEC	

Figure 4.5 Alignment of secondary structure prediction consensus obtained from Approaches 1 and 2, against the secondary structure of 2FOM. Underlined secondary structures in the consensus lines refer to regions which differ from 2FOM. The catalytic triad residues: His51, Asp75 and Ser135, are labelled with the symbol #.