wwPDB NMR Structure Validation Summary Report

Nov 20, 2023 – 12:12 PM EST

PDB ID	:	2GG1
BMRB ID	:	6800
Title	:	NMR solution structure of domain III of the E-protein of tick-borne Langat flavivirus (includes RDC restraints)
Authors	:	Mukherjee, M.; Dutta, K.; White, M.A.; Cowburn, D.; Fox, R.O.
Deposited on	:	2006-03-23

This is a wwPDB NMR Structure Validation Summary Report for a publicly released PDB entry.

We welcome your comments at *validation@mail.wwpdb.org*
A user guide is available at
https://www.wwpdb.org/validation/2017/NMRValidationReportHelp
with specific help available everywhere you see the (i) symbol.

The types of validation reports are described at
http://www.wwpdb.org/validation/2017/FAQs#types.

The following versions of software and data (see references (i)) were used in the production of this report:

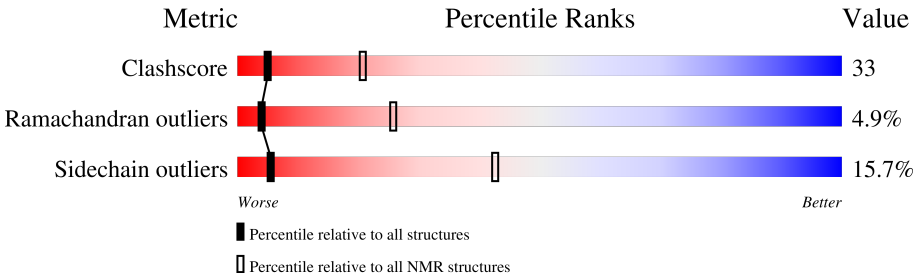
Cyrange	:	Kirchner and Güntert (2011)
NmrClust	:	Kelley et al. (1996)
MolProbity	:	4.02b-467
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
wwPDB-RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
wwPDB-ShiftChecker	:	v1.2
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.36

1 Overall quality at a glance (i)

The following experimental techniques were used to determine the structure:
SOLUTION NMR

The overall completeness of chemical shifts assignment is 78%.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.



Metric	Whole archive (#Entries)	NMR archive (#Entries)
Clashscore	158937	12864
Ramachandran outliers	154571	11451
Sidechain outliers	154315	11428

The table below summarises the geometric issues observed across the polymeric chains and their fit to the experimental data. The red, orange, yellow and green segments indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. A cyan segment indicates the fraction of residues that are not part of the well-defined cores, and a grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <=5%

Mol	Chain	Length	Quality of chain
1	A	102	37% 46% • • • 9%

2 Ensemble composition and analysis

This entry contains 20 models. Model 1 is the overall representative, medoid model (most similar to other models).

The following residues are included in the computation of the global validation metrics.

Well-defined (core) protein residues			
Well-defined core	**Residue range (total)**	**Backbone RMSD (Å)**	**Medoid model**
1	A:303-A:364, A:368-A:395 (90)	0.51	1

Ill-defined regions of proteins are excluded from the global statistics.

Ligands and non-protein polymers are included in the analysis.

The models can be grouped into 5 clusters and 1 single-model cluster was found.

Cluster number	Models
1	1, 4, 5, 9, 10, 15, 16, 18
2	2, 3, 6, 7
3	11, 12, 19
4	13, 17
5	8, 20
Single-model clusters	14

3 Entry composition (i)

There is only 1 type of molecule in this entry. The entry contains 1427 atoms, of which 704 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called Genome polyprotein.

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	S	
1	A	93	1427	457	704	126	134	6	0

There are 6 discrepancies between the modelled and reference sequences:

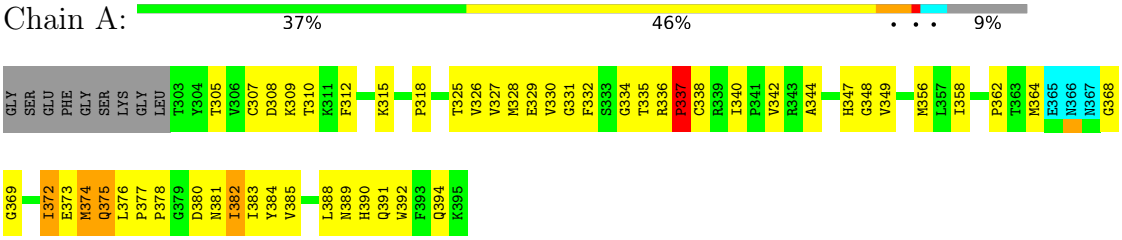
Chain	Residue	Modelled	Actual	Comment	Reference
A	294	GLY	-	cloning artifact	UNP P29838
A	295	SER	-	cloning artifact	UNP P29838
A	296	GLU	-	cloning artifact	UNP P29838
A	297	PHE	-	cloning artifact	UNP P29838
A	298	GLY	-	cloning artifact	UNP P29838
A	299	SER	-	cloning artifact	UNP P29838

4 Residue-property plots (i)

4.1 Average score per residue in the NMR ensemble

These plots are provided for all protein, RNA, DNA and oligosaccharide chains in the entry. The first graphic is the same as shown in the summary in section 1 of this report. The second graphic shows the sequence where residues are colour-coded according to the number of geometric quality criteria for which they contain at least one outlier: green = 0, yellow = 1, orange = 2 and red = 3 or more. Stretches of 2 or more consecutive residues without any outliers are shown as green connectors. Residues which are classified as ill-defined in the NMR ensemble, are shown in cyan with an underline colour-coded according to the previous scheme. Residues which were present in the experimental sample, but not modelled in the final structure are shown in grey.

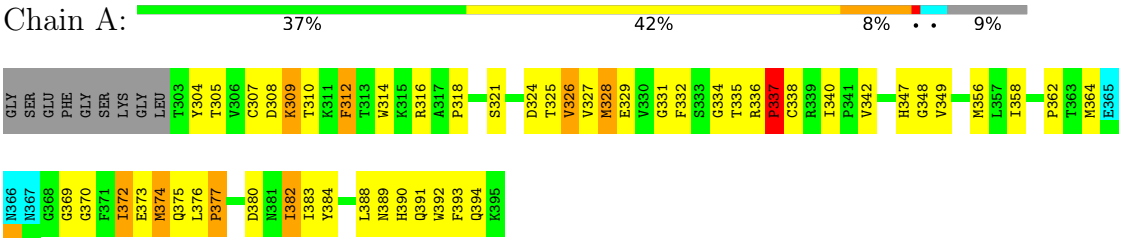
• Molecule 1: Genome polyprotein



4.2 Residue scores for the representative (medoid) model from the NMR ensemble

The representative model is number 1. Colouring as in section 4.1 above.

• Molecule 1: Genome polyprotein



5 Refinement protocol and experimental data overview

The models were refined using the following method: *distance geometry simulated annealing*.

Of the 500 calculated structures, 20 were deposited, based on the following criterion: *structures with the lowest energy*.

The following table shows the software used for structure solution, optimisation and refinement.

Software name	Classification	Version
ARIA	structure solution	2.2
CNS	refinement	1.1

The following table shows chemical shift validation statistics as aggregates over all chemical shift files. Detailed validation can be found in section 7 of this report.

Chemical shift file(s)	working_cs.cif
Number of chemical shift lists	1
Total number of shifts	1048
Number of shifts mapped to atoms	968
Number of unparsed shifts	0
Number of shifts with mapping errors	80
Number of shifts with mapping warnings	0
Assignment completeness (well-defined parts)	78%

6 Model quality

6.1 Standard geometry

There are no covalent bond-length or bond-angle outliers.

There are no bond-length outliers.

There are no bond-angle outliers.

There are no chirality outliers.

There are no planarity outliers.

6.2 Too-close contacts

In the following table, the Non-H and H(model) columns list the number of non-hydrogen atoms and hydrogen atoms in each chain respectively. The H(added) column lists the number of hydrogen atoms added and optimized by MolProbity. The Clashes column lists the number of clashes averaged over the ensemble.

Mol	Chain	Non-H	H(model)	H(added)	Clashes
1	A	698	686	682	45±5
All	All	13960	13720	13640	906

The all-atom clashscore is defined as the number of clashes found per 1000 atoms (including hydrogen atoms). The all-atom clashscore for this structure is 33.

5 of 287 unique clashes are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
1:A:382:ILE:HB	1:A:391:GLN:HA	1.09	1.20	7	20
1:A:376:LEU:HG	1:A:377:PRO:HD2	1.00	1.27	10	10
1:A:332:PHE:H	1:A:368:GLY:HA2	0.88	1.28	9	4
1:A:375:GLN:HE21	1:A:375:GLN:HA	0.86	1.31	16	11
1:A:307:CYS:HB3	1:A:335:THR:HG21	0.85	1.49	7	2

6.3 Torsion angles

6.3.1 Protein backbone

In the following table, the Percentiles column shows the percent Ramachandran outliers of the chain as a percentile score with respect to all PDB entries followed by that with respect to all NMR entries. The Analysed column shows the number of residues for which the backbone conformation was analysed and the total number of residues.

Mol	Chain	Analysed	Favoured	Allowed	Outliers	Percentiles
1	A	88/102 (86%)	72±2 (81±2%)	12±2 (14±2%)	4±1 (5±1%)	4 26
All	All	1760/2040 (86%)	1431 (81%)	243 (14%)	86 (5%)	4 26

5 of 12 unique Ramachandran outliers are listed below. They are sorted by the frequency of occurrence in the ensemble.

Mol	Chain	Res	Type	Models (Total)
1	A	337	PRO	20
1	A	334	GLY	16
1	A	309	LYS	14
1	A	305	THR	13
1	A	377	PRO	6

6.3.2 Protein sidechains (i)

In the following table, the Percentiles column shows the percent sidechain outliers of the chain as a percentile score with respect to all PDB entries followed by that with respect to all NMR entries. The Analysed column shows the number of residues for which the sidechain conformation was analysed and the total number of residues.

Mol	Chain	Analysed	Rotameric	Outliers	Percentiles
1	A	77/86 (90%)	65±2 (84±3%)	12±2 (16±3%)	5 42
All	All	1540/1720 (90%)	1298 (84%)	242 (16%)	5 42

5 of 36 unique residues with a non-rotameric sidechain are listed below. They are sorted by the frequency of occurrence in the ensemble.

Mol	Chain	Res	Type	Models (Total)
1	A	307	CYS	20
1	A	382	ILE	20
1	A	375	GLN	17
1	A	337	PRO	16
1	A	372	ILE	15

6.3.3 RNA (i)

There are no RNA molecules in this entry.

6.4 Non-standard residues in protein, DNA, RNA chains (i)

There are no non-standard protein/DNA/RNA residues in this entry.

6.5 Carbohydrates (i)

There are no monosaccharides in this entry.

6.6 Ligand geometry (i)

There are no ligands in this entry.

6.7 Other polymers (i)

There are no such molecules in this entry.

6.8 Polymer linkage issues (i)

There are no chain breaks in this entry.

7 Chemical shift validation

The completeness of assignment taking into account all chemical shift lists is 78% for the well-defined parts and 78% for the entire structure.

7.1 Chemical shift list 1

File name: working_cs.cif

Chemical shift list name: *assigned_chem_shift_list_1*

7.1.1 Bookkeeping

The following table shows the results of parsing the chemical shift list and reports the number of nuclei with statistically unusual chemical shifts.

Total number of shifts	1048
Number of shifts mapped to atoms	968
Number of unparsed shifts	0
Number of shifts with mapping errors	80
Number of shifts with mapping warnings	0
Number of shift outliers (ShiftChecker)	0

The following assigned chemical shifts were not mapped to the molecules present in the coordinate file.

- No matching atom found in the structure. First 5 (of 80) occurrences are reported below.

List ID	Chain	Res	Type	Atom	Shift Data		
					Value	Uncertainty	Ambiguity
1	A	294	GLY	H	8.253	0.02	1
1	A	294	GLY	HA2	3.943	0.02	2
1	A	294	GLY	HA3	3.846	0.02	2
1	A	294	GLY	C	173.957	0.4	1
1	A	294	GLY	CA	45.283	0.4	1
1	A	294	GLY	N	110.961	0.2	1
1	A	295	SER	H	8.087	0.02	1
1	A	295	SER	HA	4.456	0.02	1
1	A	295	SER	HB2	3.818	0.02	2
1	A	295	SER	C	174.454	0.4	1
1	A	295	SER	CA	58.28	0.4	1
1	A	295	SER	CB	63.771	0.4	1
1	A	295	SER	N	116.003	0.2	1
1	A	296	GLU	H	8.721	0.02	1

Continued on next page...

Continued from previous page...

List ID	Chain	Res	Type	Atom	Shift Data		
					Value	Uncertainty	Ambiguity
1	A	296	GLU	HA	4.21	0.02	1
1	A	296	GLU	HB2	1.891	0.02	2
1	A	296	GLU	HB3	1.813	0.02	2
1	A	296	GLU	HG2	2.128	0.02	2
1	A	296	GLU	HG3	2.066	0.02	2
1	A	296	GLU	C	176.358	0.4	1
1	A	296	GLU	CA	56.942	0.4	1
1	A	296	GLU	CB	29.857	0.4	1
1	A	296	GLU	CG	35.901	0.4	1
1	A	296	GLU	N	123.185	0.2	1
1	A	297	PHE	H	8.328	0.02	1
1	A	297	PHE	HA	4.601	0.02	1
1	A	297	PHE	HB2	3.162	0.02	2
1	A	297	PHE	HB3	3.003	0.02	2
1	A	297	PHE	C	176.475	0.4	1
1	A	297	PHE	CA	58.01	0.4	1
1	A	297	PHE	CB	39.431	0.4	1
1	A	297	PHE	N	121.0	0.2	1
1	A	298	GLY	H	8.254	0.02	1
1	A	298	GLY	HA2	3.933	0.02	2
1	A	298	GLY	HA3	3.856	0.02	2
1	A	298	GLY	C	174.318	0.4	1
1	A	298	GLY	CA	45.441	0.4	1
1	A	298	GLY	N	110.945	0.2	1
1	A	299	SER	H	8.146	0.02	1
1	A	299	SER	HA	4.401	0.02	1
1	A	299	SER	HB2	3.858	0.02	2
1	A	299	SER	C	175.083	0.4	1
1	A	299	SER	CA	58.502	0.4	1
1	A	299	SER	CB	63.78	0.4	1
1	A	299	SER	N	115.649	0.2	1
1	A	300	LYS	H	8.374	0.02	1
1	A	300	LYS	HA	4.242	0.02	1
1	A	300	LYS	HB2	1.816	0.02	2
1	A	300	LYS	HB3	1.748	0.02	2
1	A	300	LYS	HG2	1.404	0.02	2
1	A	300	LYS	HD2	1.654	0.02	2
1	A	300	LYS	HE2	2.98	0.02	2
1	A	300	LYS	C	177.181	0.4	1
1	A	300	LYS	CA	56.892	0.4	1
1	A	300	LYS	CB	32.649	0.4	1

Continued on next page...

Continued from previous page...

List ID	Chain	Res	Type	Atom	Shift Data		
					Value	Uncertainty	Ambiguity
1	A	300	LYS	CG	24.74	0.4	1
1	A	300	LYS	CD	29.078	0.4	1
1	A	300	LYS	CE	42.189	0.4	1
1	A	300	LYS	N	123.174	0.2	1
1	A	301	GLY	H	8.28	0.02	1
1	A	301	GLY	HA2	3.733	0.02	2
1	A	301	GLY	HA3	3.563	0.02	2
1	A	301	GLY	C	174.167	0.4	1
1	A	301	GLY	CA	45.383	0.4	1
1	A	301	GLY	N	109.102	0.2	1
1	A	302	LEU	H	7.856	0.02	1
1	A	302	LEU	HA	4.291	0.02	1
1	A	302	LEU	HB2	1.536	0.02	2
1	A	302	LEU	HB3	1.413	0.02	2
1	A	302	LEU	HG	1.536	0.02	1
1	A	302	LEU	HD11	0.836	0.02	2
1	A	302	LEU	HD12	0.836	0.02	2
1	A	302	LEU	HD13	0.836	0.02	2
1	A	302	LEU	C	177.3	0.4	1
1	A	302	LEU	CA	55.334	0.4	1
1	A	302	LEU	CB	42.54	0.4	1
1	A	302	LEU	CG	27.067	0.4	1
1	A	302	LEU	CD1	24.833	0.4	2
1	A	302	LEU	CD2	23.447	0.4	2
1	A	302	LEU	N	121.037	0.2	1

7.1.2 Chemical shift referencing

The following table shows the suggested chemical shift referencing corrections.

Nucleus	# values	Correction ± precision, *ppm*	Suggested action
$^{13}C_{\alpha}$	100	-0.10 ± 0.20	None needed (< 0.5 ppm)
$^{13}C_{\beta}$	88	-0.08 ± 0.23	None needed (< 0.5 ppm)
$^{13}C'$	92	0.12 ± 0.20	None needed (< 0.5 ppm)
^{15}N	93	0.02 ± 0.61	None needed (< 0.5 ppm)

7.1.3 Completeness of resonance assignments

The following table shows the completeness of the chemical shift assignments for the well-defined regions of the structure. The overall completeness is 78%, i.e. 937 atoms were assigned a chemical shift out of a possible 1203. 0 out of 13 assigned methyl groups (LEU and VAL) were assigned

stereospecifically.

	Total	^{1}H	^{13}C	^{15}N
Backbone	423/443 (95%)	174/181 (96%)	168/180 (93%)	81/82 (99%)
Sidechain	486/654 (74%)	314/428 (73%)	162/203 (80%)	10/23 (43%)
Aromatic	28/106 (26%)	14/52 (27%)	12/46 (26%)	2/8 (25%)
Overall	937/1203 (78%)	502/661 (76%)	342/429 (80%)	93/113 (82%)

7.1.4 Statistically unusual chemical shifts (i)

There are no statistically unusual chemical shifts.

7.1.5 Random Coil Index (RCI) plots (i)

The image below reports *random coil index* values for the protein chains in the structure. The height of each bar gives a probability of a given residue to be disordered, as predicted from the available chemical shifts and the amino acid sequence. A value above 0.2 is an indication of significant predicted disorder. The colour of the bar shows whether the residue is in the well-defined core (black) or in the ill-defined residue ranges (cyan), as described in section 2 on ensemble composition.If well-defined core and ill-defined regions are not identified then it is shown as gray bars.

Random coil index (RCI) for chain A: