wwPDB NMR Structure Validation Summary Report (i)

Apr 27, 2016 – 05:45 AM BST

PDB ID	:	4D4W
Title	:	Solution structure of human MBD1 CXXC1 domain
Authors	:	Thomson, R.; Smith, B.O.
Deposited on	:	2014-10-31

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
http://wwpdb.org/validation/2016/NMRValidationReportHelp
with specific help available everywhere you see the (i) symbol.

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

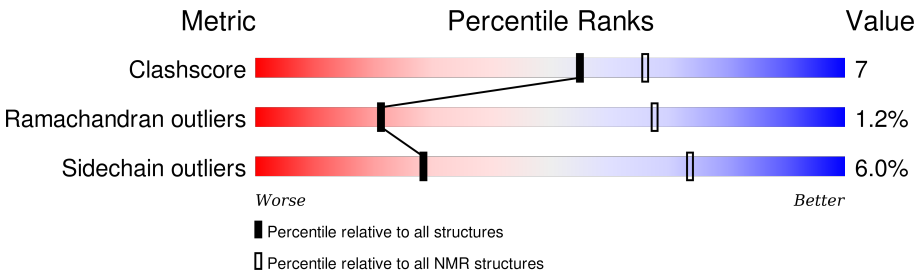
Cyrange	:	Kirchner and Güntert (2011)
NmrClust	:	Kelley et al. (1996)
MolProbity	:	4.02b-467
Mogul	:	unknown
Percentile statistics	:	20151230.v01 (using entries in the PDB archive December 30th 2015)
RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
ShiftChecker	:	rb-20027457
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	rb-20027457

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 53%.

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	114402	11133
Ramachandran outliers	111179	9975
Sidechain outliers	111093	9958

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	61	43% 8% 49%

2 Ensemble composition and analysis (i)

This entry contains 25 models. Model 2 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:15-A:35, A:48-A:57 (31)	0.44	2

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 4 clusters and 4 single-model clusters were found.

Cluster number	Models
1	1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 15, 16, 18, 24
2	6, 20, 25
3	14, 23
4	17, 21
Single-model clusters	4; 13; 19; 22

3 Entry composition (i)

There is only 1 type of molecule in this entry. The entry contains 910 atoms, of which 452 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called METHYL-CPG-BINDING DOMAIN PROTEIN 1.

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	S	
1	A	61	910	274	452	89	85	10	0

There are 5 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	1	GLY	-	EXPRESSION TAG	UNP Q9UIS9
A	2	PRO	-	EXPRESSION TAG	UNP Q9UIS9
A	3	LEU	-	EXPRESSION TAG	UNP Q9UIS9
A	4	GLY	-	EXPRESSION TAG	UNP Q9UIS9
A	5	SER	-	EXPRESSION TAG	UNP Q9UIS9

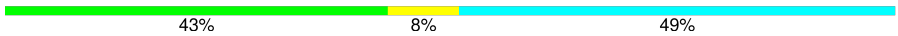
4 Residue-property plots (i)

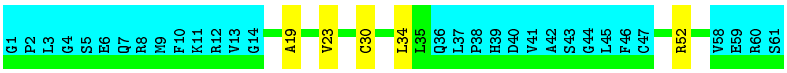
4.1 Average score per residue in the NMR ensemble

These plots are provided for all protein, RNA and DNA 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: METHYL-CPG-BINDING DOMAIN PROTEIN 1

Chain A:



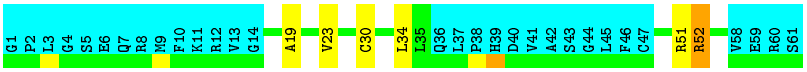


4.2 Residue scores for the representative (medoid) model from the NMR ensemble

The representative model is number 2. Colouring as in section 4.1 above.

- Molecule 1: METHYL-CPG-BINDING DOMAIN PROTEIN 1

Chain A:



5 Refinement protocol and experimental data overview (i)

The models were refined using the following method: *SIMULATED ANNEALING*.

Of the 200 calculated structures, 25 were deposited, based on the following criterion: *NOE ENERGY*.

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

Software name	Classification	Version
CNS	refinement	
TOPSIN	structure solution	1.3
CNS	structure solution	1.2
CCPNMR ANALYSIS	structure solution	2.4
CCPNMR ANALYSIS	structure solution	2.1
CCPNMR ANALYSIS	structure solution	2.2
CCPNMR ANALYSIS	structure solution	1.0
ARIA	structure solution	2.3

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)	4d4w_cs.str
Number of chemical shift lists	1
Total number of shifts	434
Number of shifts mapped to atoms	404
Number of unparsed shifts	30
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Assignment completeness (well-defined parts)	53%

No validations of the models with respect to experimental NMR restraints is performed at this time.

6 Model quality (i)

6.1 Standard geometry (i)

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 (i)

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	227	220	228	3±1
All	All	5675	5500	5700	74

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 7.

5 of 18 unique clashes are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
1:A:52:ARG:HA	1:A:52:ARG:NE	0.59	2.12	2	5
1:A:52:ARG:NE	1:A:52:ARG:HA	0.57	2.14	19	7
1:A:19:ALA:O	1:A:23:VAL:HG23	0.57	1.99	22	19
1:A:30:CYS:O	1:A:34:LEU:HD13	0.54	2.03	15	17
1:A:15:CYS:SG	1:A:16:GLY:N	0.47	2.86	21	1

6.3 Torsion angles (i)

6.3.1 Protein backbone (i)

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	31/61 (51%)	28±1 (89±4%)	3±1 (10±5%)	0±1 (1±2%)	21 68
All	All	775/1525 (51%)	690 (89%)	76 (10%)	9 (1%)	21 68

All 4 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	57	ILE	6
1	A	16	GLY	1
1	A	24	THR	1
1	A	15	CYS	1

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	26/51 (51%)	24±1 (94±3%)	2±1 (6±3%)	28 74
All	All	650/1275 (51%)	611 (94%)	39 (6%)	28 74

5 of 7 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	52	ARG	18
1	A	32	THR	8
1	A	27	CYS	5
1	A	26	ASP	3
1	A	15	CYS	2

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 carbohydrates 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 (i)

The completeness of assignment taking into account all chemical shift lists is 53% for the well-defined parts and 48% for the entire structure.

7.1 Chemical shift list 1

File name: 4d4w_cs.str

Chemical shift list name: *assigned_chem_shift_list*

7.1.1 Bookkeeping (i)

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	434
Number of shifts mapped to atoms	404
Number of unparsed shifts	30
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Number of shift outliers (ShiftChecker)	0

The following errors were found when reading this chemical shift list.

- Chemical shift has been reported more than once. First 5 (of 24) occurences are reported below.

Shift ID	Chain	Res	Type	Atom	Shift Data		
					Value	Uncertainty	Ambiguity
23	A	7	GLN	HE21	7.596	0.007	1
25	A	7	GLN	HE22	6.89	0.01	1
30	A	7	GLN	NE2	112.915	0.055	1
48	A	10	PHE	HE2	7.302	?	3
173	A	31	SER	H	9.008	0.006	1

- Entity instance (chain) must be specified. First 5 (of 6) occurences are reported below.

Shift ID	Chain	Res	Type	Atom	Shift Data		
					Value	Uncertainty	Ambiguity
108	?	21	CYO	H	8.541	?	1
113	?	21	CYO	N	120.995	?	1
186	?	33	CYO	H	8.935	0.003	1
191	?	33	CYO	N	126.406	0.003	1

Continued on next page...

Continued from previous page...

Shift ID	Chain	Res	Type	Atom	Shift Data		
					Value	Uncertainty	Ambiguity
326	?	49	CYO	H	8.599	?	1

7.1.2 Chemical shift referencing (i)

The following table shows the suggested chemical shift referencing corrections.

Nucleus	# values	Correction ± precision, *ppm*	Suggested action
$^{13}C_{\alpha}$	0	—	—
$^{13}C_{\beta}$	0	—	—
$^{13}C'$	0	—	—
^{15}N	57	-1.70 ± 0.76	Should be applied

7.1.3 Completeness of resonance assignments (i)

The following table shows the completeness of the chemical shift assignments for the well-defined regions of the structure. The overall completeness is 53%, i.e. 189 atoms were assigned a chemical shift out of a possible 358. 0 out of 4 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	^{1}H	^{13}C	^{15}N
Backbone	93/155 (60%)	62/62 (100%)	0/62 (0%)	31/31 (100%)
Sidechain	96/203 (47%)	93/120 (78%)	0/69 (0%)	3/14 (21%)
Aromatic	0/0 (—%)	0/0 (—%)	0/0 (—%)	0/0 (—%)
Overall	189/358 (53%)	155/182 (85%)	0/131 (0%)	34/45 (76%)

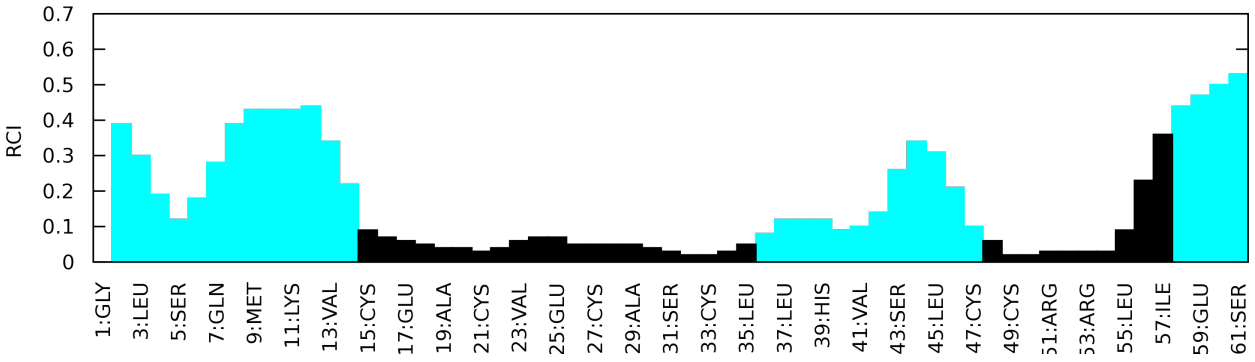
7.1.4 Statistically unusual chemical shifts (i)

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.

Random coil index (RCI) for chain A:


RCI
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1:GLY
3:LEU
5:SER
7:GLN
9:MET
11:LYS
13:VAL
15:CYS
17:GLU
19:ALA
21:CYS
23:VAL
25:GLU
27:CYS
29:ALA
31:SER
33:CYS
35:LEU
37:LEU
39:HIS
41:VAL
43:SER
45:LEU
47:CYS
49:CYS
51:ARG
53:ARG
55:LEU
57:ILE
59:GLU
61:SER