wwPDB NMR Structure Validation Summary Report

May 28, 2020 – 10:47 pm BST

PDB ID	:	2L1G
Title	:	RDC refined solution structure of the THAP zinc finger of THAP1 in complex with its 16bp RRM1 DNA target
Authors	:	Campagne, S.; Gervais, V.; Saurel, O.; Milon, A.
Deposited on	:	2010-07-28

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 following versions of software and data (see references (1)) were used in the production of this report:

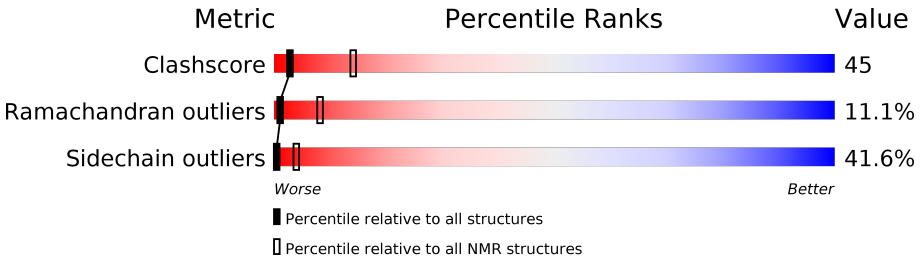
Cyrange	:	Kirchner and Güntert (2011)
NmrClust	:	Kelley et al. (1996)
MolProbity	:	4.02b-467
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
ShiftChecker	:	2.11
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.11

1 Overall quality at a glance

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

The overall completeness of chemical shifts assignment was not calculated.

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	87	23% 52% 16% • 8%
2	B	16	88% 6% 6%
3	C	16	13% 81% 6%

2 Ensemble composition and analysis

This entry contains 17 models. Model 11 is the overall representative, medoid model (most similar to other models). The authors have identified model 1 as representative, based on the following criterion: *lowest energy*.

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:3-A:82 (80)	0.14	11

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 2 clusters and 3 single-model clusters were found.

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

3 Entry composition

There are 4 unique types of molecules in this entry. The entry contains 2447 atoms, of which 1078 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called THAP domain-containing protein 1.

Mol	Chain	Residues	Atoms						Trace
1	A	87	Total	C	H	N	O	S	0
			1434	455	719	127	128	5	

There are 7 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	62	SER	CYS	ENGINEERED MUTATION	UNP Q9NVV9
A	67	SER	CYS	ENGINEERED MUTATION	UNP Q9NVV9
A	83	GLU	-	EXPRESSION TAG	UNP Q9NVV9
A	84	LEU	-	EXPRESSION TAG	UNP Q9NVV9
A	85	VAL	-	EXPRESSION TAG	UNP Q9NVV9
A	86	PRO	-	EXPRESSION TAG	UNP Q9NVV9
A	87	ARG	-	EXPRESSION TAG	UNP Q9NVV9

- Molecule 2 is a DNA chain called DNA (5'-D(*GP*CP*TP*TP*GP*TP*GP*TP*GP*GP*GP*CP*AP*GP*CP*G)-3').

Mol	Chain	Residues	Atoms						Trace
2	B	16	Total	C	H	N	O	P	0
			512	157	181	62	97	15	

- Molecule 3 is a DNA chain called DNA (5'-D(P*CP*GP*CP*TP*GP*CP*CP*CP*AP*CP*AP*CP*AP*AP*GP*C)-3').

Mol	Chain	Residues	Atoms						Trace
3	C	16	Total	C	H	N	O	P	0
			500	152	178	61	93	16	

- Molecule 4 is ZINC ION (three-letter code: ZN) (formula: Zn).

Mol	Chain	Residues	Atoms	
4	A	1	Total	Zn
			1	1

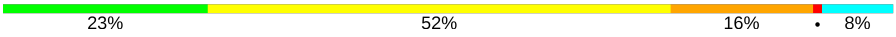
4 Residue-property plots

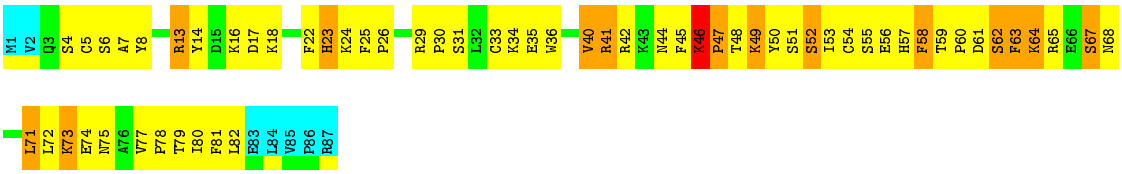
4.1 Average score per residue in the NMR ensemble

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: THAP domain-containing protein 1

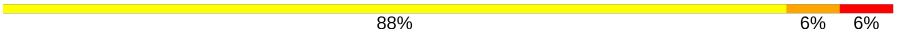
Chain A:

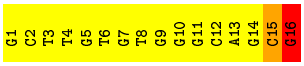




• Molecule 2: DNA (5'-D(*GP*CP*TP*TP*GP*TP*GP*TP*GP*GP*GP*CP*AP*GP*CP*G)-3')

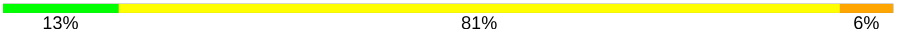
Chain B:





• Molecule 3: DNA (5'-D(P*CP*GP*CP*TP*GP*CP*CP*CP*AP*CP*AP*CP*AP*AP*GP*C)-3')

Chain C:

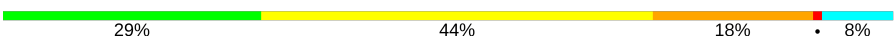


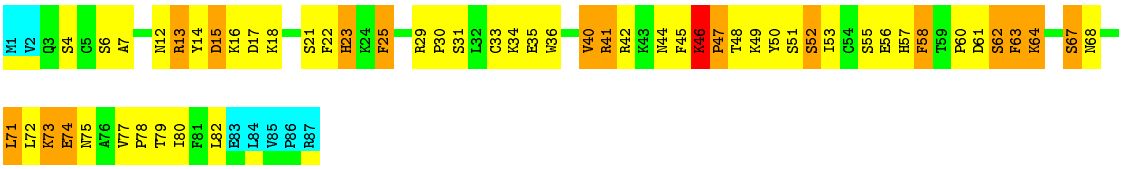
4.2 Residue scores for the representative (medoid) model from the NMR ensemble

The representative model is number 11. Colouring as in section 4.1 above.

• Molecule 1: THAP domain-containing protein 1

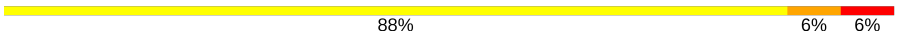
Chain A:

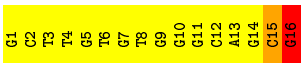




• Molecule 2: DNA (5'-D(*GP*CP*TP*TP*GP*TP*GP*TP*GP*GP*GP*CP*AP*GP*CP*G)-3')

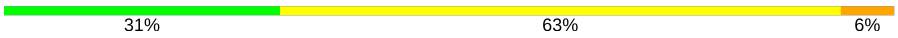
Chain B:





• Molecule 3: DNA (5'-D(P*CP*GP*CP*TP*GP*CP*CP*CP*AP*CP*AP*CP*AP*AP*GP*C)-3')

Chain C:



5 Refinement protocol and experimental data overview

The models were refined using the following method: *Rigid body docking, Semi flexible simulated annealing, Water refinement.*

Of the 200 calculated structures, 17 were deposited, based on the following criterion: *structures with the least restraint violations.*

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

Software name	Classification	Version
CNS	refinement	1.21

No chemical shift data was provided. No validations of the models with respect to experimental NMR restraints is performed at this time.

COVALENT-GEOMETRY INFOmissingINFO

5.1 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	657	653	652	40±5
2	B	331	181	182	47±2
3	C	322	178	178	25±3
All	All	22287	17204	17204	1781

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 45.

5 of 243 unique clashes are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
2:B:15:DC:H1'	2:B:16:DG:N7	1.12	1.60	7	17
2:B:7:DG:C2'	2:B:8:DT:H71	0.94	1.91	5	10
2:B:14:DG:H2"	2:B:15:DC:C5	0.91	1.99	16	17
2:B:7:DG:H2'	2:B:8:DT:H72	0.91	1.41	6	7
1:A:25:PHE:HD2	1:A:33:CYS:HG	0.89	1.11	7	17

5.2 Torsion angles (i)

5.2.1 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	80/87 (92%)	57±2 (71±2%)	14±2 (18±2%)	9±1 (11±1%)	1 8
All	All	1360/1479 (92%)	970 (71%)	239 (18%)	151 (11%)	1 8

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	46	LYS	17
1	A	64	LYS	17
1	A	67	SER	17
1	A	41	ARG	17
1	A	47	PRO	17

5.2.2 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	75/82 (91%)	44±2 (58±3%)	31±2 (42±3%)	0 4
All	All	1275/1394 (91%)	744 (58%)	531 (42%)	0 4

5 of 53 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	34	LYS	17
1	A	49	LYS	17
1	A	71	LEU	17
1	A	56	GLU	17
1	A	77	VAL	17

5.2.3 RNA

There are no RNA molecules in this entry.

5.3 Non-standard residues in protein, DNA, RNA chains

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

5.4 Carbohydrates

There are no carbohydrates in this entry.

5.5 Ligand geometry

Of 1 ligands modelled in this entry, 1 is monoatomic - leaving 0 for Mogul analysis.

5.6 Other polymers

There are no such molecules in this entry.

5.7 Polymer linkage issues

There are no chain breaks in this entry.

6 Chemical shift validation

No chemical shift data were provided