wwPDB NMR Structure Validation Summary Report ⓘ

Mar 6, 2022 – 12:10 PM EST

PDB ID	:	2KAE
Title	:	data-driven model of MED1:DNA complex
Authors	:	Lowry, J.A.; Gamsjaeger, R.; Thong, S.; Hung, W.; Kwan, A.H.; Broitman-Maduro, G.; Matthews, J.M.; Maduro, M.; Mackay, J.P.
Deposited on	:	2008-11-04

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

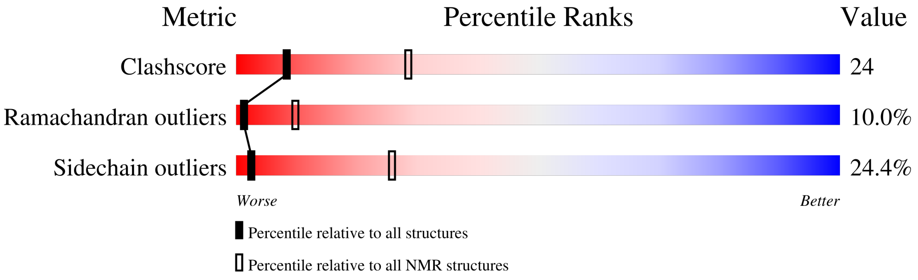
MolProbity	:	4.02b-467
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
ShiftChecker	:	2.27
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.27

1 Overall quality at a glance

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

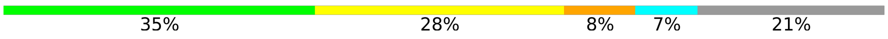
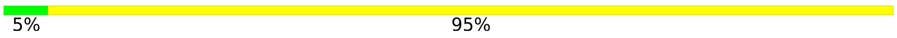
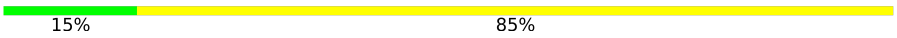
The overall completeness of chemical shifts assignment was not calculated.

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	71	
2	B	20	
2	C	20	

2 Ensemble composition and analysis ⓘ

This entry contains 10 models. Model 10 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:112-A:162 (51)	0.25	10

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 1 single-model cluster was found.

Cluster number	Models
1	3, 4, 5, 6, 8, 10
2	1, 2, 9
Single-model clusters	7

3 Entry composition

There are 3 unique types of molecules in this entry. The entry contains 2221 atoms, of which 937 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called GATA-type transcription factor.

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	S	
1	A	56	950	292	481	92	81	4	0

There are 15 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	104	GLY	-	expression tag	UNP Q9GSP3
A	105	SER	-	expression tag	UNP Q9GSP3
A	106	HIS	-	expression tag	UNP Q9GSP3
A	107	MET	-	expression tag	UNP Q9GSP3
A	108	ASN	-	expression tag	UNP Q9GSP3
A	109	LYS	-	expression tag	UNP Q9GSP3
A	110	LYS	-	expression tag	UNP Q9GSP3
A	167	GLU	-	expression tag	UNP Q9GSP3
A	168	THR	-	expression tag	UNP Q9GSP3
A	169	ASN	-	expression tag	UNP Q9GSP3
A	170	GLY	-	expression tag	UNP Q9GSP3
A	171	VAL	-	expression tag	UNP Q9GSP3
A	172	ASP	-	expression tag	UNP Q9GSP3
A	173	SER	-	expression tag	UNP Q9GSP3
A	174	PHE	-	expression tag	UNP Q9GSP3

- Molecule 2 is a DNA chain called 5'-D(*DCP*DGP*DGP*DAP*DAP*DAP*DAP*DGP*DTP*DAP*DTP*DAP*DCP*DTP*DTP*DTP*DTP*DCP*DCP*DG)-3'.

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	P	
2	C	20	635	196	228	74	118	19	0
			Total	C	H	N	O	P	
2	B	20	635	196	228	74	118	19	0

- Molecule 3 is ZINC ION (three-letter code: ZN) (formula: Zn).

Mol	Chain	Residues	Atoms	
			Total	Zn
3	A	1	1	1

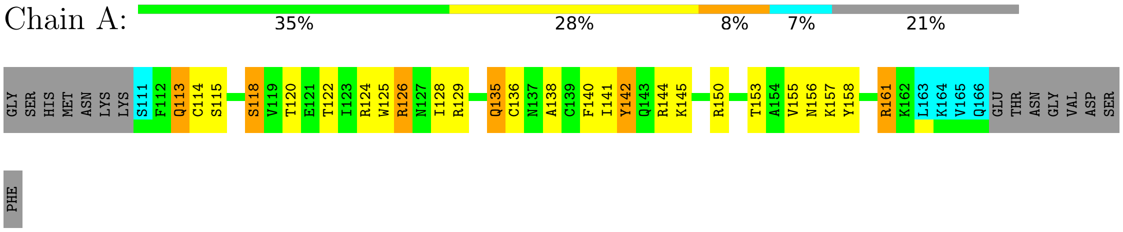
4 Residue-property plots ⓘ

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: GATA-type transcription factor

Chain A:



- Molecule 2: 5'-D(*DCP*DGP*DGP*DAP*DAP*DAP*DAP*DGP*DTP*DAP*DTP*DAP*DCP*DTP*DTP*DTP*DTP*DCP*DCP*DG)-3'

Chain C:

- Molecule 2: 5'-D(*DCP*DGP*DGP*DAP*DAP*DAP*DAP*DGP*DTP*DAP*DTP*DAP*DCP*DTP*DTP*DTP*DTP*DCP*DCP*DG)-3'

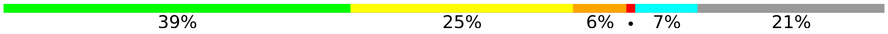
Chain B:

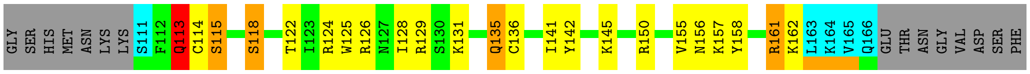
4.2 Residue scores for the representative (medoid) model from the NMR ensemble

The representative model is number 10. Colouring as in section 4.1 above.

- Molecule 1: GATA-type transcription factor

Chain A:





- Molecule 2: 5’-D(*DCP*DGP*DGP*DAP*DAP*DAP*DAP*DGP*DTP*DAP*DTP*DAP*DCP*DTP*DTP*DTP*DTP*DCP*DCP*DG)-3’

Chain C:

- Molecule 2: 5’-D(*DCP*DGP*DGP*DAP*DAP*DAP*DAP*DGP*DTP*DAP*DTP*DAP*DCP*DTP*DTP*DTP*DTP*DCP*DCP*DG)-3’

Chain B:

5 Refinement protocol and experimental data overview ⓘ

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

Of the 10 calculated structures, 10 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**
CNS	refinement	

No chemical shift data was provided.

6 Model quality ⓘ

6.1 Standard geometry ⓘ

Bond lengths and bond angles in the following residue types are not validated in this section: ZN

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	430	435	435	19±3
2	B	407	228	228	21±4
2	C	407	228	228	22±4
All	All	12450	8910	8910	519

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 24.

5 of 155 unique clashes are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
1:A:161:ARG:HB2	2:B:14:DT:H73	0.68	1.66	8	8
1:A:141:ILE:HG22	2:C:29:DT:H71	0.67	1.65	6	2
1:A:119:VAL:HG23	1:A:135:GLN:HE22	0.66	1.50	2	1
1:A:161:ARG:HH22	2:B:16:DT:H71	0.64	1.52	2	3
2:B:14:DT:C2’	2:B:15:DT:H71	0.64	2.23	4	3

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	51/71 (72%)	34±1 (66±2%)	12±1 (24±2%)	5±1 (10±2%)	1	10
All	All	510/710 (72%)	338 (66%)	121 (24%)	51 (10%)	1	10

5 of 8 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	128	ILE	10
1	A	129	ARG	10
1	A	118	SER	9
1	A	113	GLN	8
1	A	126	ARG	5

6.3.2 Protein sidechains

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	48/66 (73%)	36±1 (76±2%)	12±1 (24±2%)	2	26
All	All	480/660 (73%)	363 (76%)	117 (24%)	2	26

5 of 27 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	135	GLN	10
1	A	142	TYR	10
1	A	156	ASN	10
1	A	122	THR	9
1	A	145	LYS	9

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)

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

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 ⓘ

No chemical shift data were provided