wwPDB NMR Structure Validation Summary Report (i)

Jun 22, 2024 – 10:52 AM EDT

PDB ID	:	5MHD
BMRB ID	:	34070
Title	:	Biosynthetic engineered A22S-B3K-B31R human insulin monomer structure in water/acetonitrile solutions.
Authors	:	Bocian, W.; Kozerski, L.; Bednarek, E.; Sitkowski, J.
Deposited on	:	2016-11-24

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:

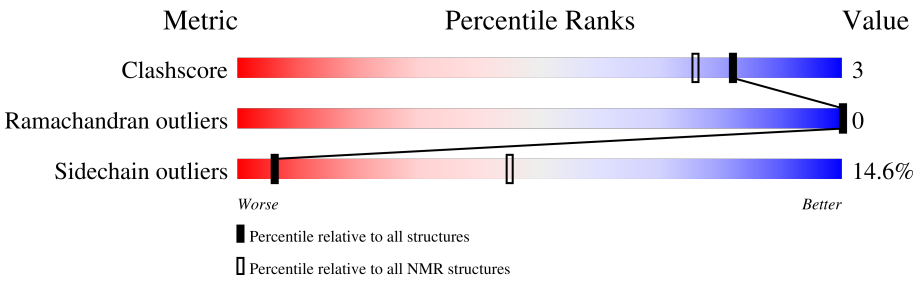
MolProbity	:	4.02b-467
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
wwPDB-RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
wwPDB-ShiftChecker	:	v1.2
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.37.1

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 52%.

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	22	68% 14% 5% 14%
2	B	31	52% 6% 42%

2 Ensemble composition and analysis (i)

This entry contains 20 models. Model 16 is the overall representative, medoid model (most similar to other models). The authors have identified model 1 as representative, based on the following criterion: *fewest violations*.

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:2-A:20, B:4-B:19, B:23-B:24 (37)	0.20	16

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

3 Entry composition (i)

There are 2 unique types of molecules in this entry. The entry contains 829 atoms, of which 406 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called Insulin.

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	S	
1	A	22	323	102	154	26	37	4	0

There is a discrepancy between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	22	SER	-	expression tag	UNP P01308

- Molecule 2 is a protein called Insulin.

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	S	
2	B	31	506	166	252	44	42	2	0

There is a discrepancy between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
B	3	LYS	ASN	conflict	UNP P01308

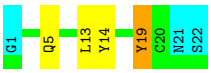
4 Residue-property plots (i)

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: Insulin

Chain A: 68% 14% 5% 14%



• Molecule 2: Insulin

Chain B: 52% 6% 42%

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

The representative model is number 16. Colouring as in section 4.1 above.

• Molecule 1: Insulin

Chain A: 64% 18% 5% 14%

• Molecule 2: Insulin

Chain B: 52% 42%

5 Refinement protocol and experimental data overview (i)

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

Of the 100 calculated structures, 20 were deposited, based on the following criterion: *20 structures for lowest energy.*

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

Software name	**Classification**	**Version**
DYANA	refinement	
Amber	structure calculation	14

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	351
Number of shifts mapped to atoms	351
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Assignment completeness (well-defined parts)	52%

6 Model quality (i)

6.1 Standard geometry (i)

The Z score for a bond length (or angle) is the number of standard deviations the observed value is removed from the expected value. A bond length (or angle) with $|Z| > 5$ is considered an outlier worth inspection. RMSZ is the (average) root-mean-square of all Z scores of the bond lengths (or angles).

Mol	Chain	Bond lengths		Bond angles	
		RMSZ	#Z>5	RMSZ	#Z>5
1	A	0.70±0.01	0±0/152 (0.0± 0.0%)	1.15±0.04	1±0/207 (0.5± 0.0%)
2	B	0.67±0.01	0±0/142 (0.0± 0.0%)	1.08±0.03	0±0/193 (0.0± 0.1%)
All	All	0.69	0/5880 (0.0%)	1.12	21/8000 (0.3%)

Chiral center outliers are detected by calculating the chiral volume of a chiral center and verifying if the center is modelled as a planar moiety or with the opposite hand. A planarity outlier is detected by checking planarity of atoms in a peptide group, atoms in a mainchain group or atoms of a sidechain that are expected to be planar.

Mol	Chain	Chirality	Planarity
1	A	0.0±0.0	1.0±0.0
All	All	0	20

There are no bond-length outliers.

All unique angle outliers are listed below. They are sorted according to the Z-score of the worst occurrence in the ensemble.

Mol	Chain	Res	Type	Atoms	Z	Observed(^{o})	Ideal(^{o})	Models	
								Worst	Total
1	A	19	TYR	CB-CG-CD1	-8.69	115.79	121.00	20	9
1	A	19	TYR	CB-CG-CD2	-8.41	115.95	121.00	3	11
2	B	6	LEU	CB-CG-CD1	5.07	119.62	111.00	9	1

There are no chirality outliers.

All unique planar outliers are listed below.

Mol	Chain	Res	Type	Group	Models (Total)
1	A	19	TYR	Sidechain	20

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	150	138	138	1±1
2	B	138	132	132	1±1
All	All	5760	5400	5400	28

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 3.

5 of 11 unique clashes are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
2:B:14:ALA:O	2:B:18:VAL:HG23	0.52	2.05	7	7
1:A:13:LEU:HG	2:B:18:VAL:HG22	0.49	1.85	7	1
1:A:16:LEU:HD11	2:B:11:LEU:CD2	0.48	2.38	13	7
1:A:13:LEU:HB3	2:B:18:VAL:HG22	0.47	1.86	8	1
1:A:16:LEU:HB2	2:B:18:VAL:HG21	0.46	1.86	8	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	19/22 (86%)	17±1 (92±4%)	2±1 (8±4%)	0±0 (0±0%)	100	100
2	B	18/31 (58%)	18±1 (98±3%)	0±1 (2±3%)	0±0 (0±0%)	100	100
All	All	740/1060 (70%)	702 (95%)	38 (5%)	0 (0%)	100	100

There are no Ramachandran outliers.

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	19/21 (90%)	17±1 (88±4%)	2±1 (12±4%)	8	51
2	B	15/27 (56%)	12±1 (82±8%)	3±1 (18±8%)	4	38
All	All	680/960 (71%)	581 (85%)	99 (15%)	6	45

5 of 10 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)
2	B	15	LEU	19
1	A	14	TYR	18
2	B	11	LEU	18
1	A	13	LEU	15
1	A	5	GLN	11

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 (i)

The completeness of assignment taking into account all chemical shift lists is 52% for the well-defined parts and 49% for the entire structure.

7.1 Chemical shift list 1

File name: working_cs.cif

Chemical shift list name: *SKRR_ ChemicalShift.txt*

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	351
Number of shifts mapped to atoms	351
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Number of shift outliers (ShiftChecker)	0

7.1.2 Chemical shift referencing (i)

No chemical shift referencing corrections were calculated (not enough data).

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 52%, i.e. 254 atoms were assigned a chemical shift out of a possible 485. 0 out of 9 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	**^{1}H**	**^{13}C**	**^{15}N**
Backbone	75/187 (40%)	75/76 (99%)	0/74 (0%)	0/37 (0%)
Sidechain	159/247 (64%)	159/164 (97%)	0/79 (0%)	0/4 (0%)
Aromatic	20/51 (39%)	20/25 (80%)	0/24 (0%)	0/2 (0%)
Overall	254/485 (52%)	254/265 (96%)	0/177 (0%)	0/43 (0%)

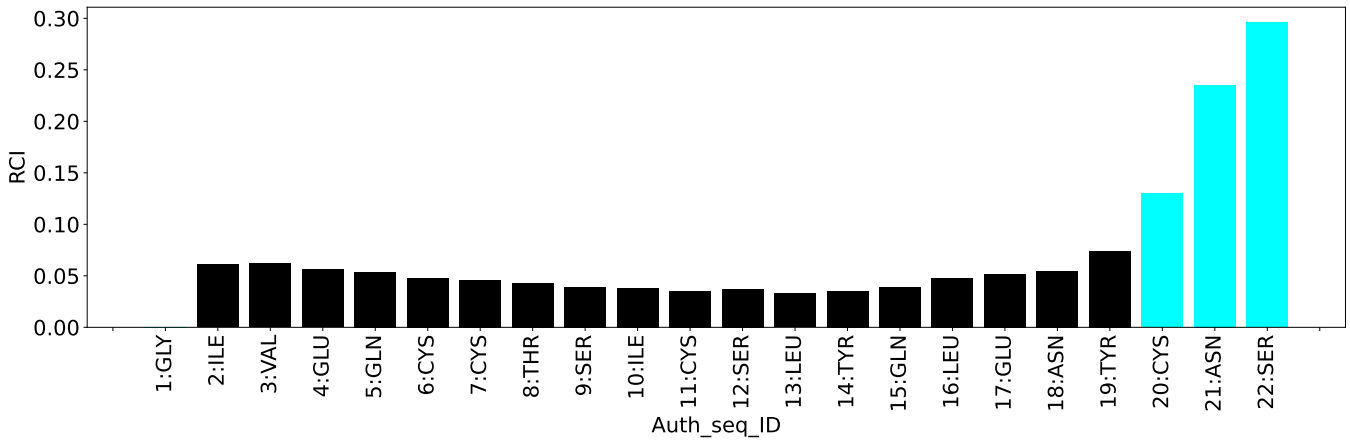
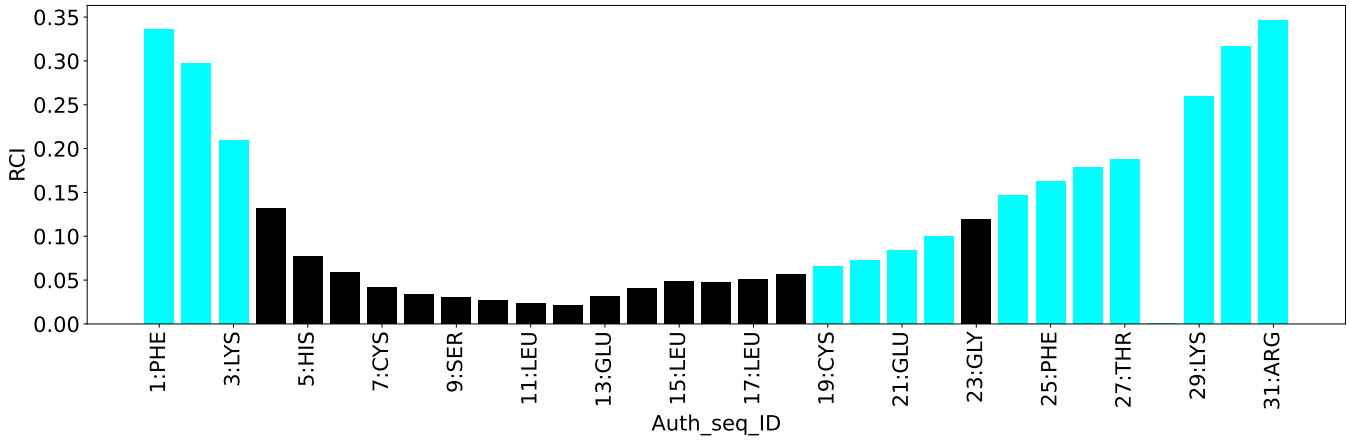
7.1.4 Statistically unusual chemical shifts (i)

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:



Random coil index (RCI) for chain B: