wwPDB NMR Structure Validation Summary Report (i)

May 28, 2020 – 10:44 pm BST

PDB ID	:	2KHM
Title	:	Structure of the C-terminal non-repetitive domain of the spider dragline silk protein ADF-3
Authors	:	Hagn, F.X.; Eisoldt, L.; Hardy, J.G.; Vendrely, C.; Coles, M.; Scheibel, T.; Kessler, H.
Deposited on	:	2009-04-09

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:

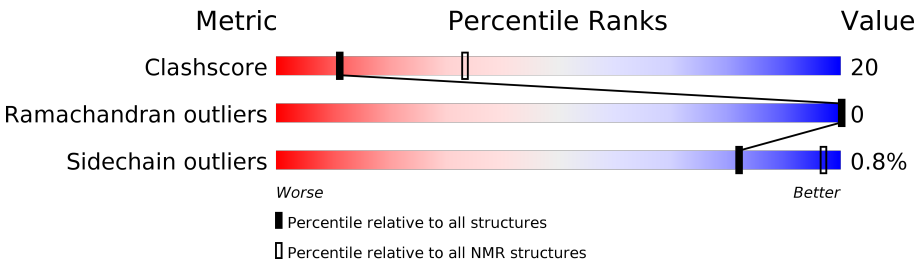
Cyrange	:	Kirchner and Güntert (2011)
NmrClust	:	Kelley et al. (1996)
MolProbity	:	4.02b-467
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
ShiftChecker	:	2.11
Ideal geometry (proteins)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

1 Overall quality at a glance (i)

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

The overall completeness of chemical shifts assignment is 97%.

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	140	51% 23% 26%
1	B	140	50% 24% 26%

2 Ensemble composition and analysis (i)

This entry contains 21 models. Model 19 is the overall representative, medoid model (most similar to other models). The authors have identified model 1 as representative, based on the following criterion: *minimized average*.

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:36-A:138, B:36-B:138 (206)	0.12	19

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 3 clusters and 3 single-model clusters were found.

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

3 Entry composition (i)

There is only 1 type of molecule in this entry. The entry contains 3712 atoms, of which 1856 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called Fibroin-3.

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	S	
1	A	140	1856	558	928	164	200	6	0
1	B	140	1856	558	928	164	200	6	0

There are 40 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	1	MET	-	expression tag	UNP Q16987
A	2	ALA	-	expression tag	UNP Q16987
A	3	SER	-	expression tag	UNP Q16987
A	4	MET	-	expression tag	UNP Q16987
A	5	THR	-	expression tag	UNP Q16987
A	6	GLY	-	expression tag	UNP Q16987
A	7	GLY	-	expression tag	UNP Q16987
A	8	GLN	-	expression tag	UNP Q16987
A	9	GLN	-	expression tag	UNP Q16987
A	10	MET	-	expression tag	UNP Q16987
A	11	GLY	-	expression tag	UNP Q16987
A	12	ARG	-	expression tag	UNP Q16987
A	13	GLY	-	expression tag	UNP Q16987
A	14	SER	-	expression tag	UNP Q16987
A	15	MET	-	expression tag	UNP Q16987
A	34	ALA	VAL	SEE REMARK 999	UNP Q16987
A	40	ALA	VAL	SEE REMARK 999	UNP Q16987
A	67	ASN	LYS	SEE REMARK 999	UNP Q16987
A	68	GLN	HIS	SEE REMARK 999	UNP Q16987
A	140	GLY	-	SEE REMARK 999	UNP Q16987
B	1	MET	-	expression tag	UNP Q16987
B	2	ALA	-	expression tag	UNP Q16987
B	3	SER	-	expression tag	UNP Q16987
B	4	MET	-	expression tag	UNP Q16987
B	5	THR	-	expression tag	UNP Q16987
B	6	GLY	-	expression tag	UNP Q16987
B	7	GLY	-	expression tag	UNP Q16987
B	8	GLN	-	expression tag	UNP Q16987
B	9	GLN	-	expression tag	UNP Q16987

Continued on next page...

Continued from previous page...

Chain	Residue	Modelled	Actual	Comment	Reference
B	10	MET	-	expression tag	UNP Q16987
B	11	GLY	-	expression tag	UNP Q16987
B	12	ARG	-	expression tag	UNP Q16987
B	13	GLY	-	expression tag	UNP Q16987
B	14	SER	-	expression tag	UNP Q16987
B	15	MET	-	expression tag	UNP Q16987
B	34	ALA	VAL	SEE REMARK 999	UNP Q16987
B	40	ALA	VAL	SEE REMARK 999	UNP Q16987
B	67	ASN	LYS	SEE REMARK 999	UNP Q16987
B	68	GLN	HIS	SEE REMARK 999	UNP Q16987
B	140	GLY	-	SEE REMARK 999	UNP Q16987

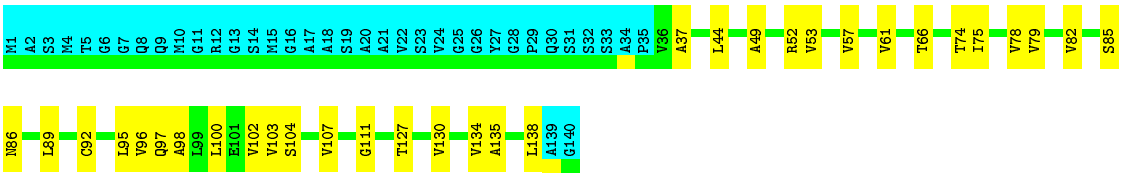
4 Residue-property plots (i)

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.

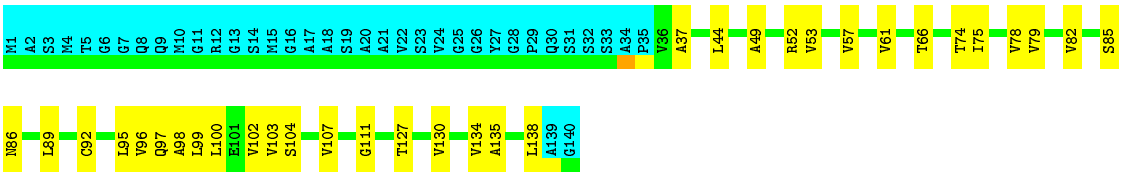
- Molecule 1: Fibroin-3

Chain A: 51% 23% 26%



- Molecule 1: Fibroin-3

Chain B: 50% 24% 26%

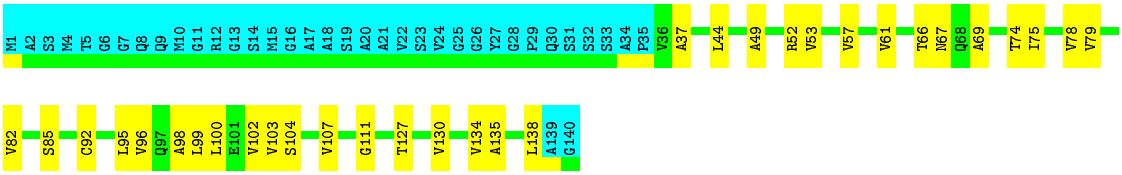


4.2 Residue scores for the representative (medoid) model from the NMR ensemble

The representative model is number 19. Colouring as in section 4.1 above.

- Molecule 1: Fibroin-3

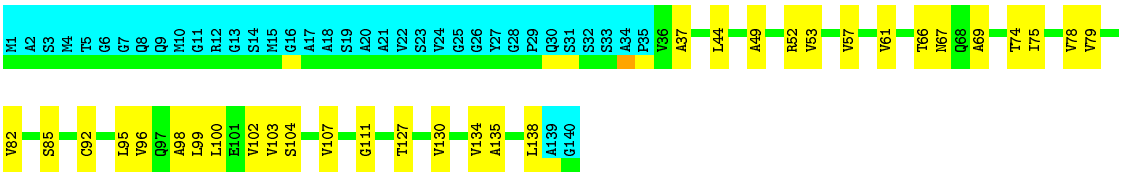
Chain A: 51% 23% 26%



- Molecule 1: Fibroin-3

Chain B:

51%
23%
26%


M1 A2 S3 M4 T5 G6 G7 Q8 Q9 M10 G11 R12 G13 S14 M15 G16 A17 A18 S19 A20 A21 V22 S23 V24 G25 G26 Y27 G28 P29 Q30 S31 S32 S33 A34 P35 V36 A37 L44 A49 R52 V53 V57 V61 T66 N67 Q68 A69 T74 I75 V78 V79
V82 S85 C92 L95 V96 Q97 A98 L99 L100 E101 V102 V103 S104 V107 G111 T127 V130 V134 A135 L138 A139 G140

5 Refinement protocol and experimental data overview (i)

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

Of the 50 calculated structures, 21 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
X-PLOR NIH	structure solution	
X-PLOR NIH	refinement	

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)	input_cs.cif
Number of chemical shift lists	1
Total number of shifts	3050
Number of shifts mapped to atoms	3050
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Assignment completeness (well-defined parts)	97%

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	697	708	708	40±4
1	B	697	708	708	40±4
All	All	29274	29736	29736	1192

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 20.

5 of 113 unique clashes are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
1:A:138:LEU:HD22	1:B:37:ALA:HB3	0.84	1.49	1	16
1:A:37:ALA:HB3	1:B:138:LEU:HD22	0.82	1.49	1	16
1:A:111:GLY:CA	1:B:66:THR:HG21	0.76	2.11	12	17
1:A:66:THR:HG21	1:B:111:GLY:CA	0.75	2.10	8	17
1:A:98:ALA:O	1:A:102:VAL:HG23	0.71	1.85	12	21

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	103/140 (74%)	100±1 (97±1%)	3±1 (3±1%)	0±0 (0±0%)	100	100
1	B	103/140 (74%)	100±1 (97±1%)	3±1 (3±1%)	0±0 (0±0%)	100	100
All	All	4326/5880 (74%)	4207 (97%)	119 (3%)	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	80/101 (79%)	79±1 (99±1%)	1±1 (1±1%)	82	97
1	B	80/101 (79%)	79±1 (99±1%)	1±1 (1±1%)	84	97
All	All	3360/4242 (79%)	3334 (99%)	26 (1%)	82	97

All 4 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	85	SER	13
1	B	85	SER	11
1	A	83	SER	1
1	B	83	SER	1

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 97% for the well-defined parts and 94% for the entire structure.

7.1 Chemical shift list 1

File name: input_cs.cif

Chemical shift list name: *assigned_chem_shift_list_1*

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	3050
Number of shifts mapped to atoms	3050
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Number of shift outliers (ShiftChecker)	2

7.1.2 Chemical shift referencing (i)

The following table shows the suggested chemical shift referencing corrections.

Nucleus	# values	Correction ± precision, *ppm*	Suggested action
$^{13}C_{\alpha}$	276	-0.60 ± 0.07	Should be applied
$^{13}C_{\beta}$	244	0.21 ± 0.05	None needed (< 0.5 ppm)
$^{13}C'$	240	-0.50 ± 0.05	Should be applied
^{15}N	266	0.07 ± 0.19	None needed (< 0.5 ppm)

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 97%, i.e. 2044 atoms were assigned a chemical shift out of a possible 2112. 46 out of 48 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	^{1}H	^{13}C	^{15}N
Backbone	1000/1018 (98%)	406/406 (100%)	394/412 (96%)	200/200 (100%)
Sidechain	1012/1062 (95%)	590/614 (96%)	394/412 (96%)	28/36 (78%)

Continued on next page...

Continued from previous page...

	Total	^{1}H	^{13}C	^{15}N
Aromatic	32/32 (100%)	16/16 (100%)	16/16 (100%)	0/0 (—%)
Overall	2044/2112 (97%)	1012/1036 (98%)	804/840 (96%)	228/236 (97%)

7.1.4 Statistically unusual chemical shifts (i)

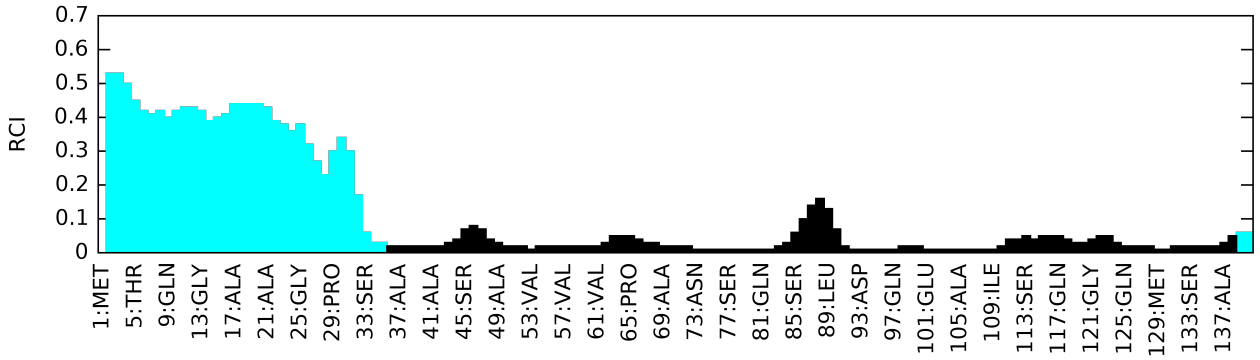
The following table lists the statistically unusual chemical shifts. These are statistical measures, and large deviations from the mean do not necessarily imply incorrect assignments. Molecules containing paramagnetic centres or hemes are expected to give rise to anomalous chemical shifts.

Mol	Chain	Res	Type	Atom	Shift, *ppm*	Expected range, *ppm*	Z-score
1	B	43	ARG	NE	67.00	92.63 – 76.73	-11.1
1	A	43	ARG	NE	67.00	92.63 – 76.73	-11.1

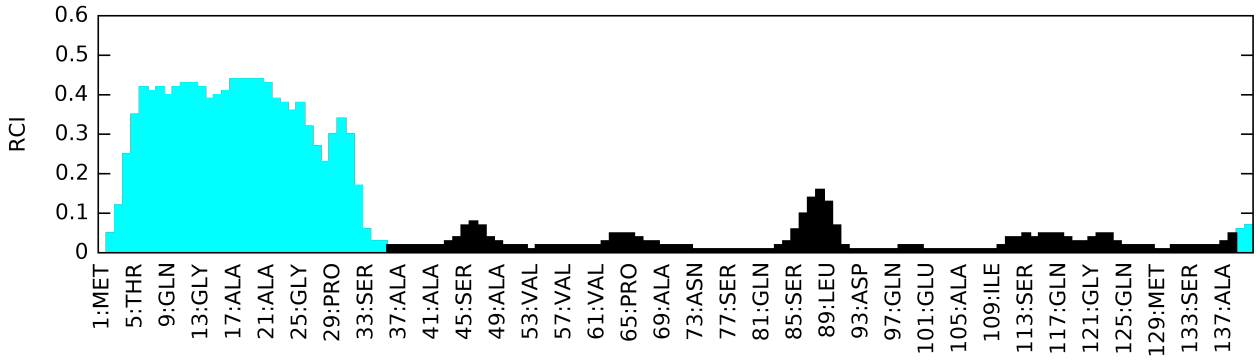
7.1.5 Random Coil Index (RCI) plots (i)

The images below report *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:



Random coil index (RCI) for chain B:


RCI
0.6
0.5
0.4
0.3
0.2
0.1
0
1:MET
5:THR
9:GLN
13:GLY
17:ALA
21:ALA
25:GLY
29:PRO
33:SER
37:ALA
41:ALA
45:SER
49:ALA
53:VAL
57:VAL
61:VAL
65:PRO
69:ALA
73:ASN
77:SER
81:GLN
85:SER
89:LEU
93:ASP
97:GLN
101:GLU
105:ALA
109:ILE
113:SER
117:GLN
121:GLY
125:GLN
129:MET
133:SER
137:ALA