Full wwPDB NMR Structure Validation Report (i)

Apr 27, 2016 – 04:30 AM BST

PDB ID	:	2N8G
Title	:	NMR Structure of the homeodomain transcription factor Gbx1[E23R,R58E] from Homo sapiens
Authors	:	Proudfoot, A.K.; Serrano, P.; Geralt, M.; Wuthrich, K.; Joint Center for Structural Genomics (JCSG); Partnership for Stem Cell Biology (STEMCELL)
Deposited on	:	2015-10-15

This is a Full wwPDB NMR Structure Validation 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:

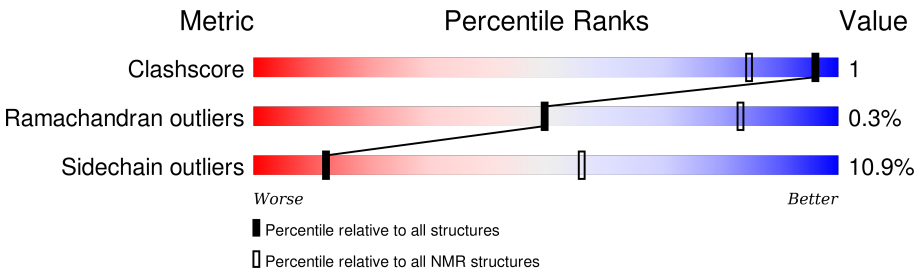
Cyrange	:	Kirchner and Güntert (2011)
NmrClust	:	Kelley et al. (1996)
MolProbity	:	4.02b-467
Mogul	:	unknown
Percentile statistics	:	20151230.v01 (using entries in the PDB archive December 30th 2015)
RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
ShiftChecker	:	rb-20027457
Ideal 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 87%.

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	71	66% 6% 28%

2 Ensemble composition and analysis (i)

This entry contains 20 models. Model 1 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:13-A:63 (51)	0.36	1

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 6 single-model clusters were found.

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

3 Entry composition (i)

There is only 1 type of molecule in this entry. The entry contains 1198 atoms, of which 613 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called Homeobox protein GBX-1.

Mol	Chain	Residues	Atoms						Trace
			Total	C	H	N	O	S	
1	A	71	1198	369	613	116	99	1	0

There are 3 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
A	1	GLY	-	EXPRESSION TAG	UNP Q14549
A	23	ARG	GLU	ENGINEERED MUTATION	UNP Q14549
A	58	GLU	ARG	ENGINEERED MUTATION	UNP Q14549

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: Homeobox protein GBX-1

Chain A:

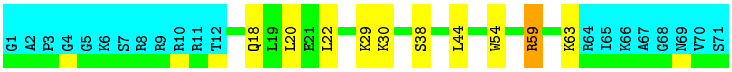
4.2 Scores per residue for each member of the ensemble

Colouring as in section 4.1 above.

4.2.1 Score per residue for model 1 (medoid)

- Molecule 1: Homeobox protein GBX-1

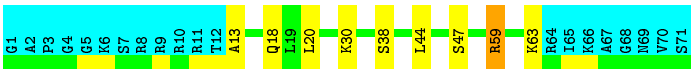
Chain A:



4.2.2 Score per residue for model 2

- Molecule 1: Homeobox protein GBX-1

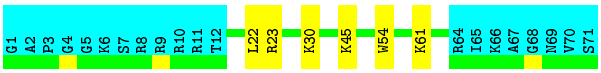
Chain A:



4.2.3 Score per residue for model 3

- Molecule 1: Homeobox protein GBX-1

Chain A: 63% 8% 28%



4.2.4 Score per residue for model 4

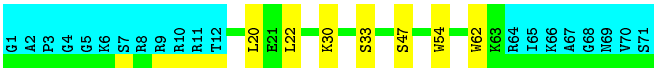
- Molecule 1: Homeobox protein GBX-1

Chain A: 63% 8% 28%

4.2.5 Score per residue for model 5

- Molecule 1: Homeobox protein GBX-1

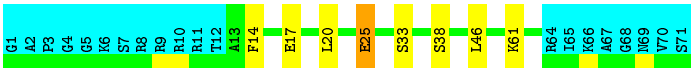
Chain A: 62% 10% 28%



4.2.6 Score per residue for model 6

- Molecule 1: Homeobox protein GBX-1

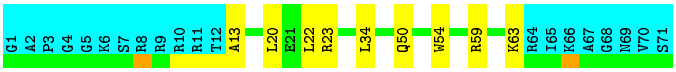
Chain A: 61% 10% • 28%



4.2.7 Score per residue for model 7

- Molecule 1: Homeobox protein GBX-1

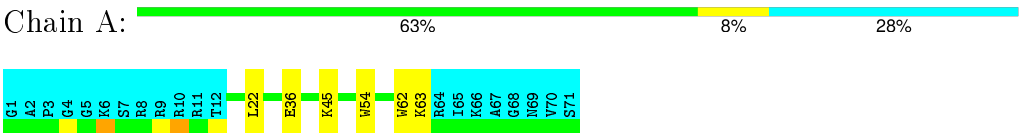
Chain A: 59% 13% 28%



4.2.8 Score per residue for model 8

- Molecule 1: Homeobox protein GBX-1

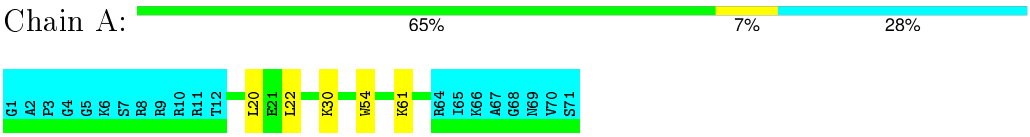
Chain A:



4.2.9 Score per residue for model 9

- Molecule 1: Homeobox protein GBX-1

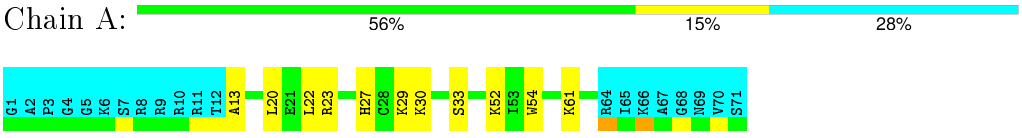
Chain A:



4.2.10 Score per residue for model 10

- Molecule 1: Homeobox protein GBX-1

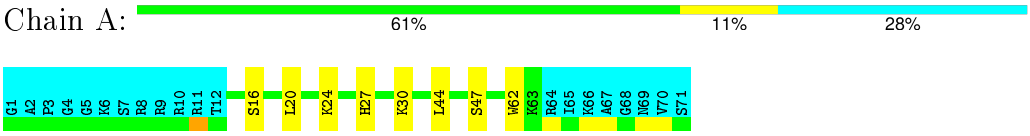
Chain A:



4.2.11 Score per residue for model 11

- Molecule 1: Homeobox protein GBX-1

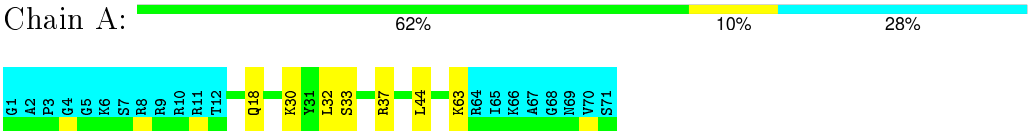
Chain A:



4.2.12 Score per residue for model 12

- Molecule 1: Homeobox protein GBX-1

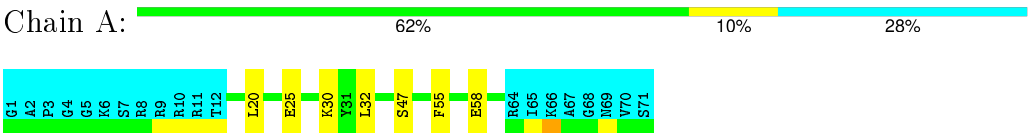
Chain A:



4.2.13 Score per residue for model 13

- Molecule 1: Homeobox protein GBX-1

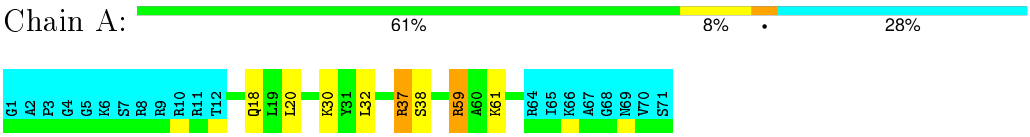
Chain A:



4.2.14 Score per residue for model 14

- Molecule 1: Homeobox protein GBX-1

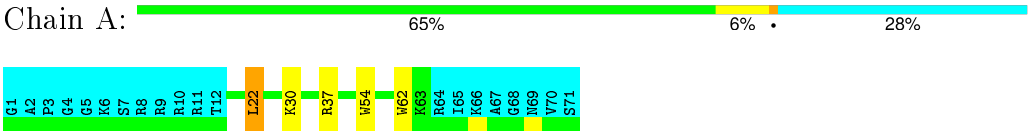
Chain A:



4.2.15 Score per residue for model 15

- Molecule 1: Homeobox protein GBX-1

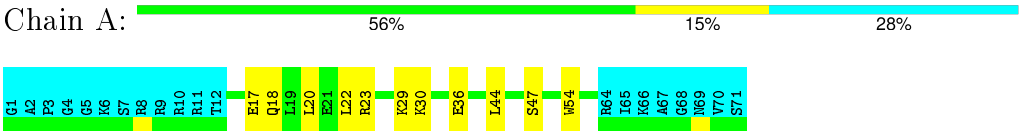
Chain A:



4.2.16 Score per residue for model 16

- Molecule 1: Homeobox protein GBX-1

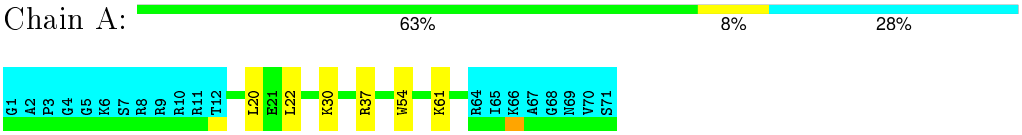
Chain A:



4.2.17 Score per residue for model 17

- Molecule 1: Homeobox protein GBX-1

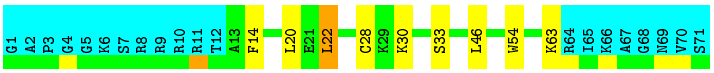
Chain A:



4.2.18 Score per residue for model 18

- Molecule 1: Homeobox protein GBX-1

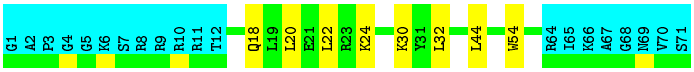
Chain A:



4.2.19 Score per residue for model 19

- Molecule 1: Homeobox protein GBX-1

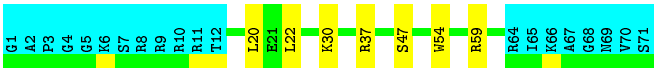
Chain A:



4.2.20 Score per residue for model 20

- Molecule 1: Homeobox protein GBX-1

Chain A:



5 Refinement protocol and experimental data overview (i)

The models were refined using the following method: *molecular dynamics*.

Of the 80 calculated structures, 20 were deposited, based on the following criterion: *target function*.

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

Software name	Classification	Version
DYANA	refinement	
OPAL	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)	2n8g_cs.str
Number of chemical shift lists	1
Total number of shifts	836
Number of shifts mapped to atoms	836
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Assignment completeness (well-defined parts)	87%

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)

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	0.60±0.01	0±0/447 (0.0±0.0%)	0.99±0.04	0±1/599 (0.1±0.1%)
All	All	0.60	0/8940 (0.0%)	0.99	9/11980 (0.1%)

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	0.2±0.4
All	All	0	5

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	37	ARG	NE-CZ-NH2	-7.82	116.39	120.30	14	2
1	A	59	ARG	NE-CZ-NH2	-5.96	117.32	120.30	14	2
1	A	23	ARG	NE-CZ-NH2	-5.43	117.59	120.30	10	3
1	A	25	GLU	CA-CB-CG	5.32	125.11	113.40	6	1
1	A	37	ARG	NE-CZ-NH1	5.20	122.90	120.30	14	1

There are no chirality outliers.

All unique planar outliers are listed below. They are sorted by the frequency of occurrence in the ensemble.

Mol	Chain	Res	Type	Group	Models (Total)
1	A	37	ARG	Sidechain	3
1	A	59	ARG	Sidechain	1
1	A	23	ARG	Sidechain	1

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	437	450	450	1±1
All	All	8740	9000	9000	22

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 1.

All unique clashes are listed below, sorted by their clash magnitude.

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
1:A:22:LEU:HD23	1:A:54:TRP:CD2	0.59	2.32	18	12
1:A:18:GLN:HA	1:A:44:LEU:HD13	0.50	1.82	19	1
1:A:14:PHE:CE2	1:A:22:LEU:HD21	0.49	2.41	18	1
1:A:18:GLN:HA	1:A:44:LEU:HD22	0.48	1.85	16	4
1:A:14:PHE:CE2	1:A:46:LEU:HD13	0.47	2.44	18	2
1:A:32:LEU:HD13	1:A:55:PHE:CE2	0.44	2.47	13	1
1:A:22:LEU:HD23	1:A:54:TRP:CG	0.41	2.50	7	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	51/71 (72%)	49±1 (95±3%)	2±1 (5±2%)	0±0 (0±1%)	50	83
All	All	1020/1420 (72%)	971 (95%)	46 (5%)	3 (0%)	50	83

All 1 unique Ramachandran outliers are listed below.

Mol	Chain	Res	Type	Models (Total)
1	A	13	ALA	3

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	47/61 (77%)	42±2 (89±3%)	5±2 (11±3%)	12 56
All	All	940/1220 (77%)	838 (89%)	102 (11%)	12 56

All 28 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	20	LEU	16
1	A	30	LYS	16
1	A	47	SER	7
1	A	61	LYS	6
1	A	63	LYS	6
1	A	59	ARG	5
1	A	33	SER	5
1	A	38	SER	5
1	A	29	LYS	4
1	A	62	TRP	4
1	A	32	LEU	3
1	A	17	GLU	3
1	A	36	GLU	2
1	A	45	LYS	2
1	A	22	LEU	2
1	A	25	GLU	2
1	A	24	LYS	2
1	A	27	HIS	2
1	A	21	GLU	1
1	A	18	GLN	1
1	A	16	SER	1
1	A	58	GLU	1
1	A	28	CYS	1
1	A	50	GLN	1
1	A	52	LYS	1
1	A	44	LEU	1
1	A	37	ARG	1
1	A	34	LEU	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 87% for the well-defined parts and 78% for the entire structure.

7.1 Chemical shift list 1

File name: 2n8g_cs.str

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	836
Number of shifts mapped to atoms	836
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Number of shift outliers (ShiftChecker)	5

7.1.2 Chemical shift referencing (i)

The following table shows the suggested chemical shift referencing corrections.

Nucleus	# values	Correction ± precision, *ppm*	Suggested action
$^{13}C_{\alpha}$	69	-0.13 ± 0.11	None needed (< 0.5 ppm)
$^{13}C_{\beta}$	66	0.14 ± 0.08	None needed (< 0.5 ppm)
$^{13}C'$	63	-0.19 ± 0.17	None needed (< 0.5 ppm)
^{15}N	68	0.38 ± 0.37	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 87%, i.e. 620 atoms were assigned a chemical shift out of a possible 710. 9 out of 9 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	^{1}H	^{13}C	^{15}N
Backbone	249/255 (98%)	101/102 (99%)	97/102 (95%)	51/51 (100%)
Sidechain	314/382 (82%)	197/226 (87%)	112/135 (83%)	5/21 (24%)

Continued on next page...

Continued from previous page...

	Total	^{1}H	^{13}C	^{15}N
Aromatic	57/73 (78%)	29/39 (74%)	26/30 (87%)	2/4 (50%)
Overall	620/710 (87%)	327/367 (89%)	235/267 (88%)	58/76 (76%)

The following table shows the completeness of the chemical shift assignments for the full structure. The overall completeness is 78%, i.e. 760 atoms were assigned a chemical shift out of a possible 969. 10 out of 10 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	^{1}H	^{13}C	^{15}N
Backbone	335/353 (95%)	135/141 (96%)	132/142 (93%)	68/70 (97%)
Sidechain	368/543 (68%)	225/323 (70%)	137/181 (76%)	6/39 (15%)
Aromatic	57/73 (78%)	29/39 (74%)	26/30 (87%)	2/4 (50%)
Overall	760/969 (78%)	389/503 (77%)	295/353 (84%)	76/113 (67%)

7.1.4 Statistically unusual chemical shifts (i)

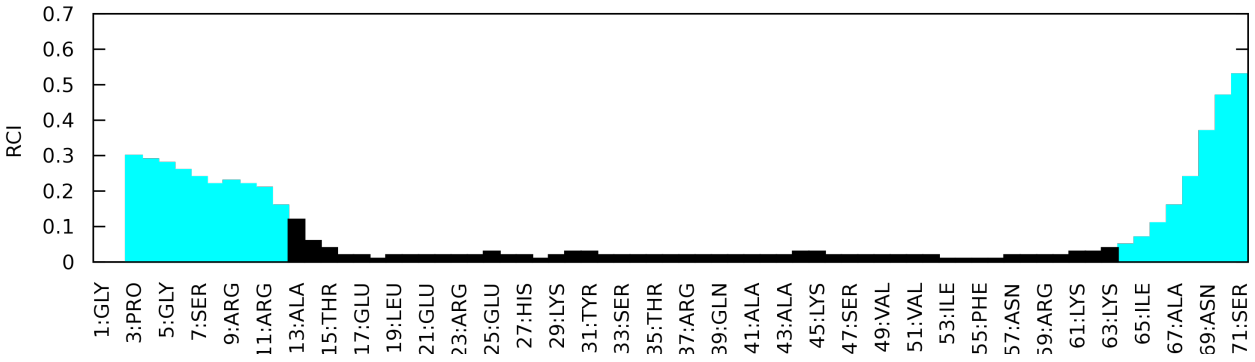
The following table lists the statistically unusual chemical shifts. These are statistical measures, and large deviations from the mean do not necessarily imply incorrect assignments. Molecules containing paramagnetic centres or hemes are expected to give rise to anomalous chemical shifts.

Mol	Chain	Res	Type	Atom	Shift, *ppm*	Expected range, *ppm*	Z-score
1	A	58	GLU	HB2	-0.03	3.08 – 0.98	-9.8
1	A	58	GLU	HG3	0.35	3.31 – 1.21	-9.1
1	A	58	GLU	HG2	0.39	3.33 – 1.23	-9.0
1	A	58	GLU	HB3	0.32	3.10 – 0.90	-7.7
1	A	22	LEU	HB2	-0.81	3.32 – -0.08	-7.1

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:PRO
5:GLY
7:SER
9:ARG
11:ARG
13:ALA
15:THR
17:GLU
19:LEU
21:GLU
23:ARG
25:GLU
27:HIS
29:LYS
31:TYR
33:SER
35:THR
37:ARG
39:GLN
41:ALA
43:ALA
45:LYS
47:SER
49:VAL
51:VAL
53:ILE
55:PHE
57:ASN
59:ARG
61:LYS
63:LYS
65:ILE
67:ALA
69:ASN
71:SER