Full wwPDB NMR Structure Validation Report (i)

Nov 9, 2020 – 02:10 PM GMT

PDB ID	:	6HPJ
Title	:	Structure of human SRSF1 RRM1 bound to AACAAA RNA
Authors	:	Allain, F.T.H.; Clery, A.
Deposited on	:	2018-09-21

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
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:

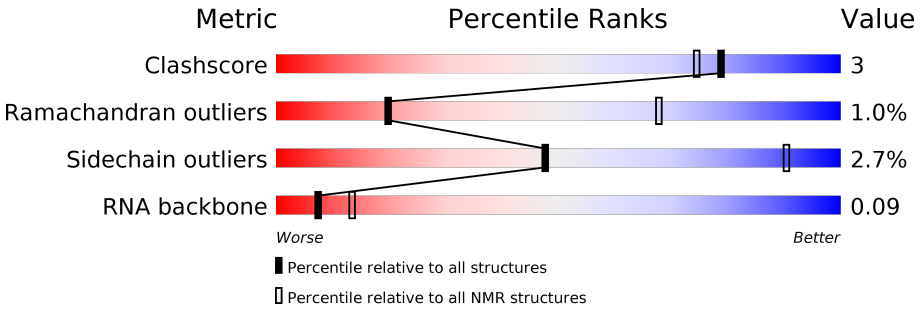
Cyrange	:	Kirchner and Güntert (2011)
NmrClust	:	Kelley et al. (1996)
MolProbity	:	4.02b-467
Percentile statistics	:	20191225.v01 (using entries in the PDB archive December 25th 2019)
RCI	:	v_1n_11_5_13_A (Berjanski et al., 2005)
PANAV	:	Wang et al. (2010)
ShiftChecker	:	2.14.6
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et al. (1996)
Validation Pipeline (wwPDB-VP)	:	2.14.6

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 is 72%.

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
RNA backbone	4643	676

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	B	161	37% • • 13% 48%
2	A	6	17% 67% 17%

2 Ensemble composition and analysis

This entry contains 20 models. Model 17 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	B:16-B:23, B:30-B:49, B:55-B:89 (63)	0.53	17

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 1 single-model cluster was found.

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

3 Entry composition

There are 2 unique types of molecules in this entry. The entry contains 1506 atoms, of which 705 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1.

Mol	Chain	Residues	Atoms						Trace
1	B	84	Total	C	H	N	O	S	0
			1311	422	637	121	130	1	

There are 12 discrepancies between the modelled and reference sequences:

Chain	Residue	Modelled	Actual	Comment	Reference
B	-63	MET	-	initiating methionine	UNP P19909
B	-62	GLN	-	expression tag	UNP P19909
B	-7	GLY	-	linker	UNP P19909
B	-6	SER	-	linker	UNP P19909
B	-5	HIS	-	linker	UNP P19909
B	-4	HIS	-	linker	UNP P19909
B	-3	HIS	-	linker	UNP P19909
B	-2	HIS	-	linker	UNP P19909
B	-1	HIS	-	linker	UNP P19909
B	0	HIS	-	linker	UNP P19909
B	37	SER	TYR	conflict	UNP Q07955
B	72	SER	TYR	conflict	UNP Q07955

- Molecule 2 is a RNA chain called RNA (5'-R(*AP*AP*CP*AP*AP*A)-3').

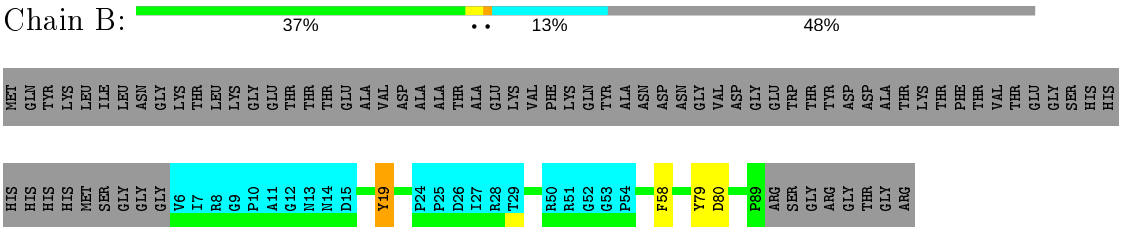
Mol	Chain	Residues	Atoms						Trace
2	A	6	Total	C	H	N	O	P	0
			195	59	68	28	35	5	

4 Residue-property plots

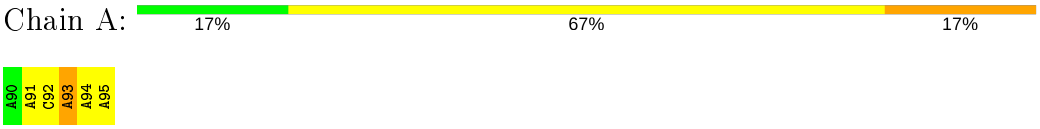
4.1 Average score per residue in the NMR ensemble

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: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1



• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

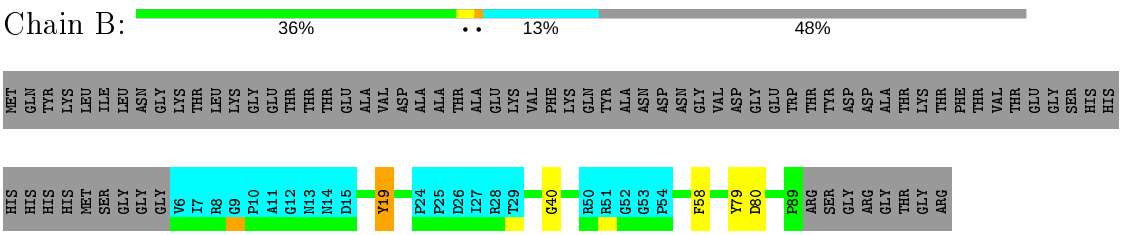


4.2 Scores per residue for each member of the ensemble

Colouring as in section 4.1 above.

4.2.1 Score per residue for model 1

• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1



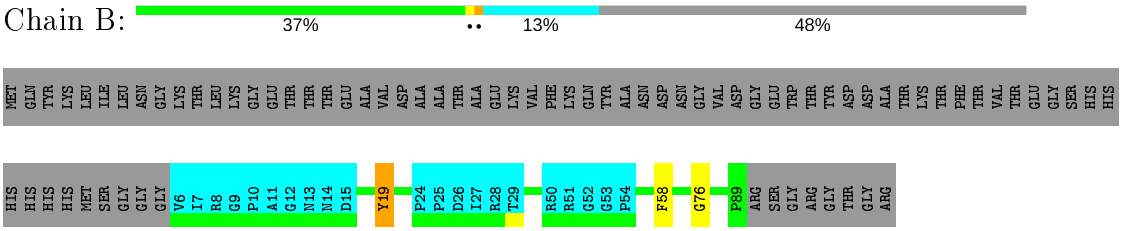
• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

Chain A:

4.2.2 Score per residue for model 2

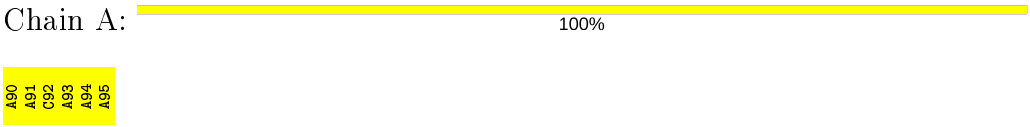
- Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



- Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

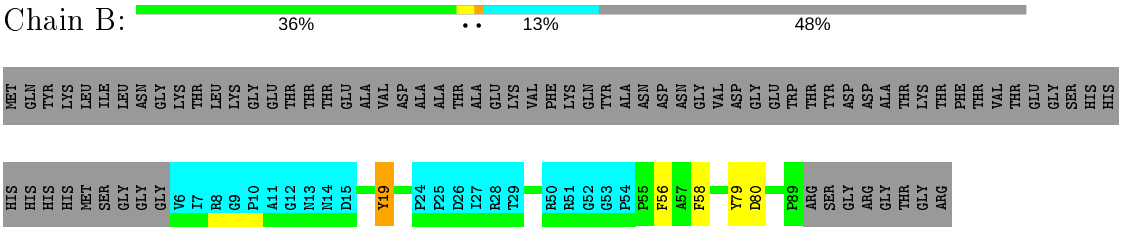
Chain A:



4.2.3 Score per residue for model 3

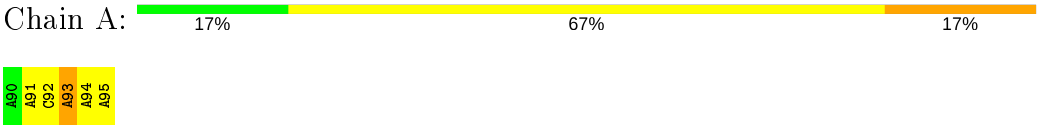
- Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



- Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

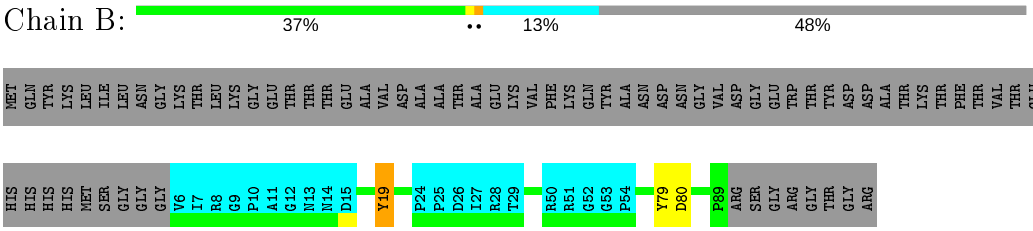
Chain A:



4.2.4 Score per residue for model 4

- Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



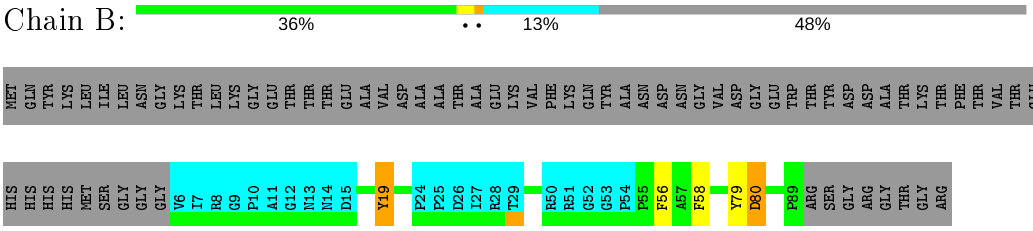
• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

Chain A:

4.2.5 Score per residue for model 5

• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



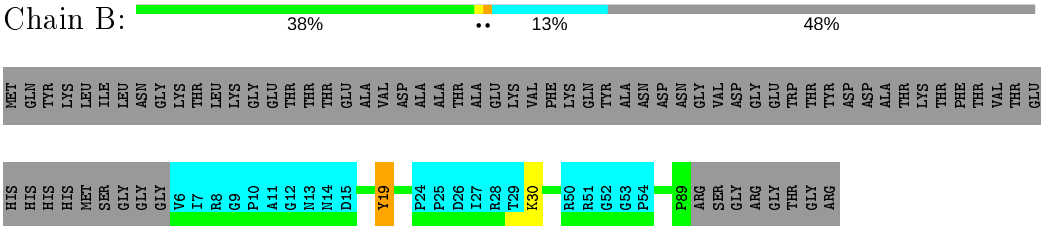
• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

Chain A:

4.2.6 Score per residue for model 6

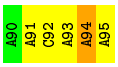
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

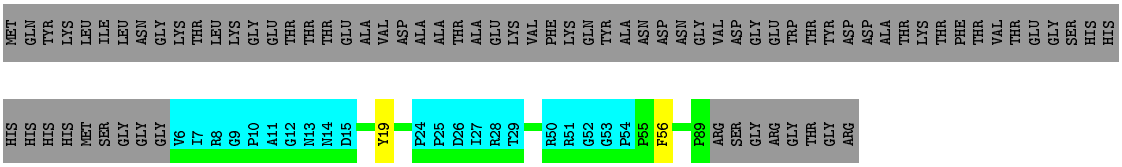
Chain A:



4.2.7 Score per residue for model 7

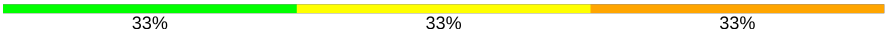
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

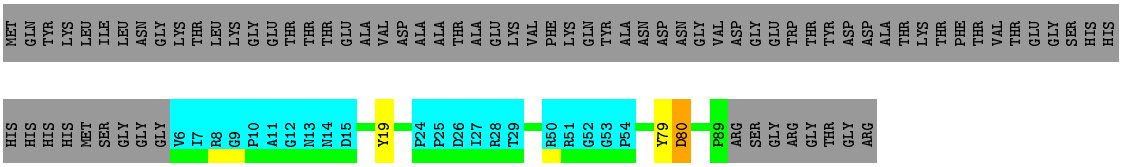
Chain A:



4.2.8 Score per residue for model 8

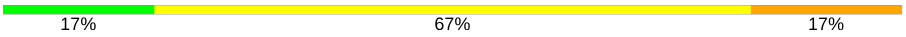
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

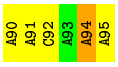
Chain B:



• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

Chain A:

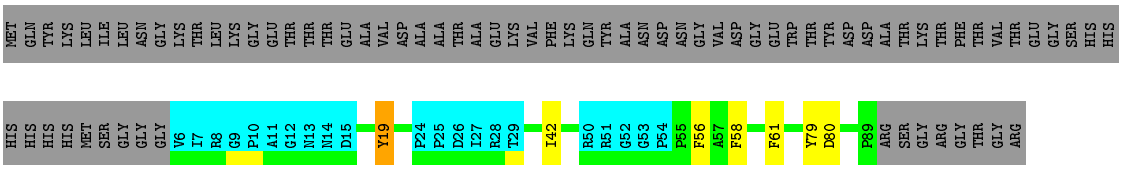




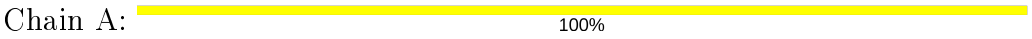
4.2.9 Score per residue for model 9

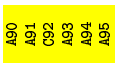
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



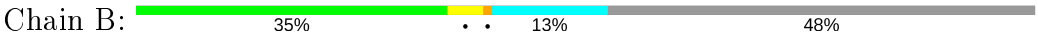
• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

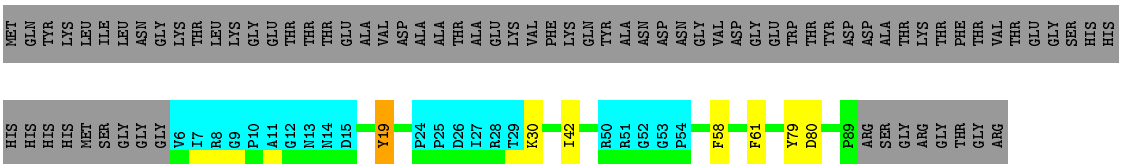




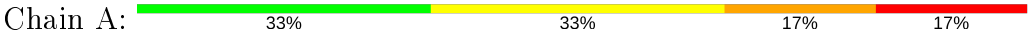
4.2.10 Score per residue for model 10

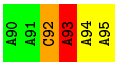
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1





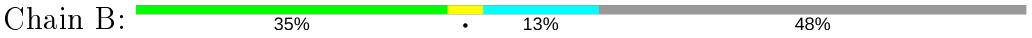
• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

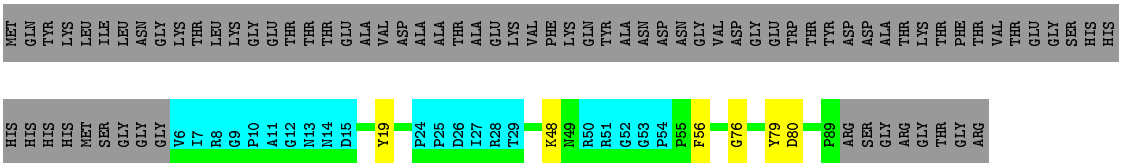




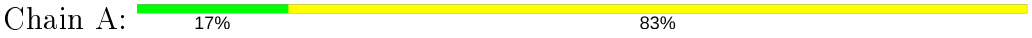
4.2.11 Score per residue for model 11

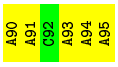
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1





• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

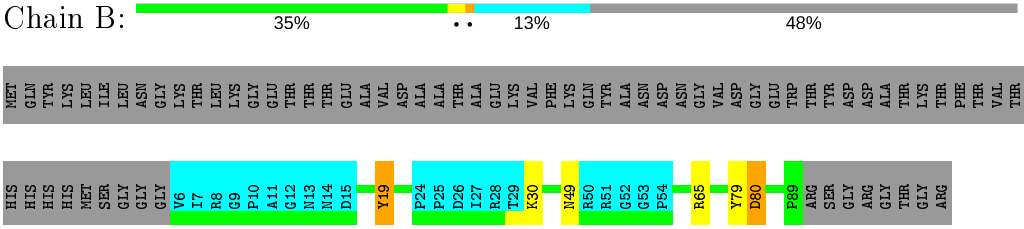




4.2.12 Score per residue for model 12

• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



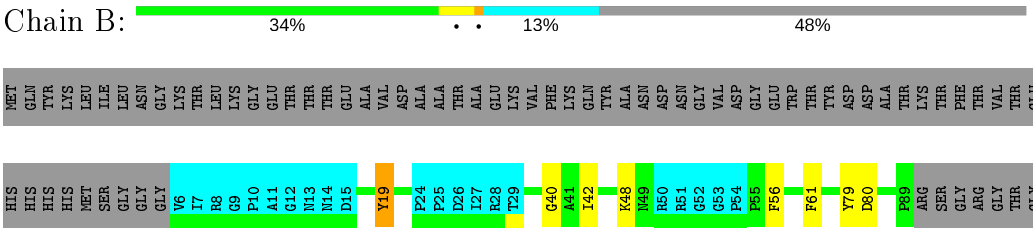
• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

Chain A:

4.2.13 Score per residue for model 13

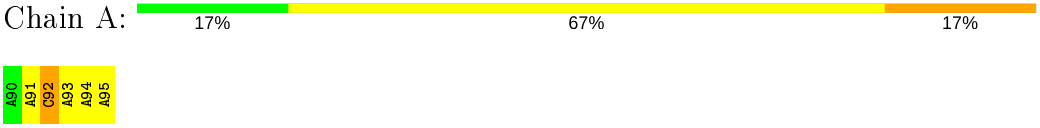
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

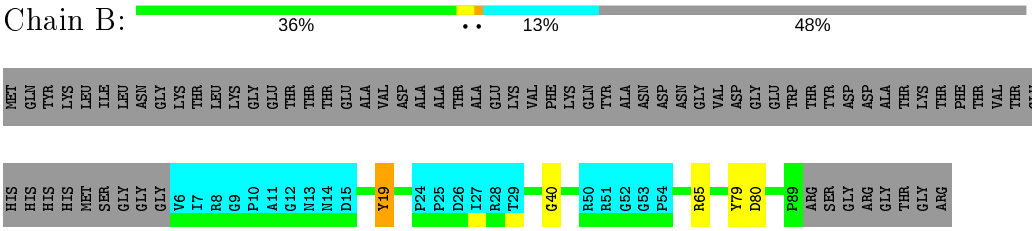
Chain A:



4.2.14 Score per residue for model 14

• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



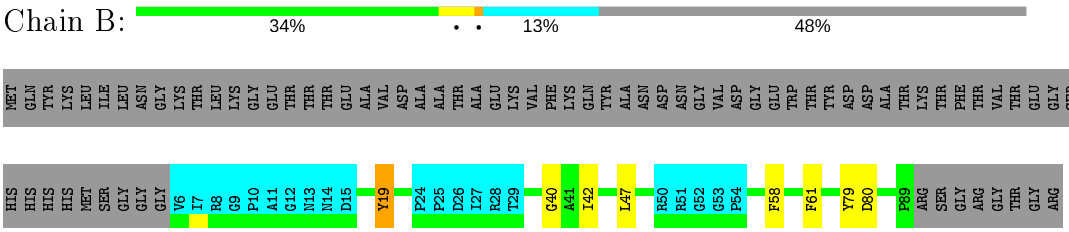
• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

Chain A:

4.2.15 Score per residue for model 15

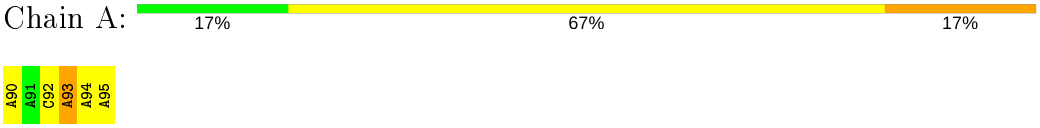
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

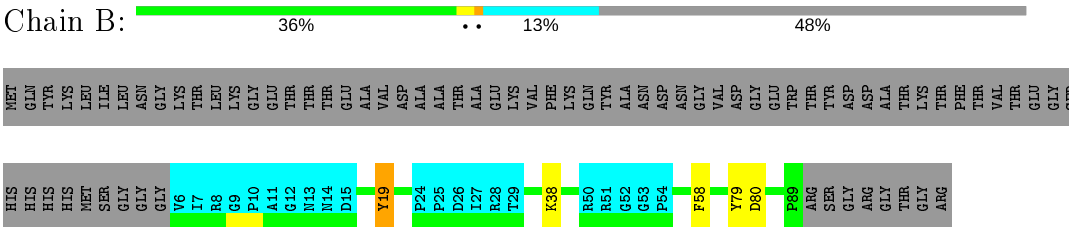
Chain A:



4.2.16 Score per residue for model 16

• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



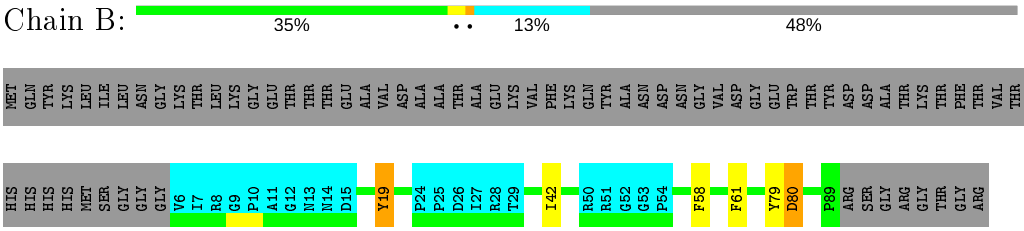
• Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

Chain A:

4.2.17 Score per residue for model 17 (medoid)

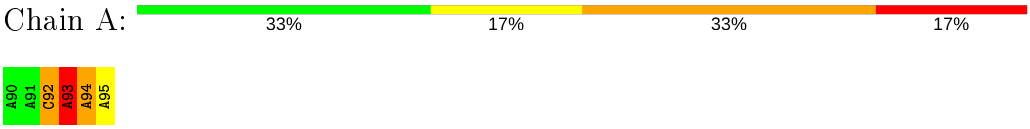
• Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



- Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

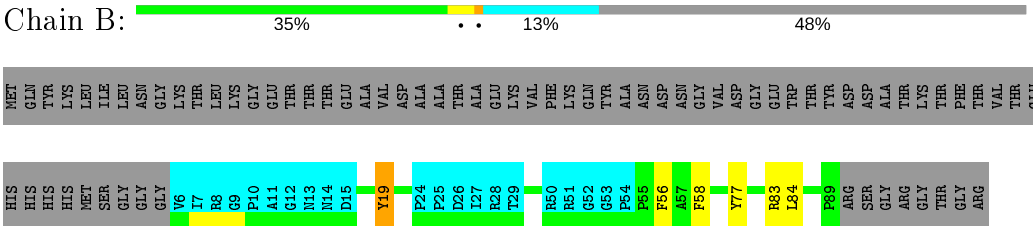
Chain A:



4.2.18 Score per residue for model 18

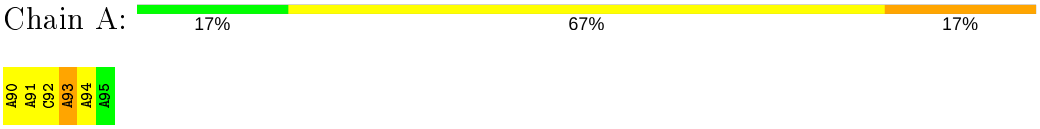
- Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



- Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

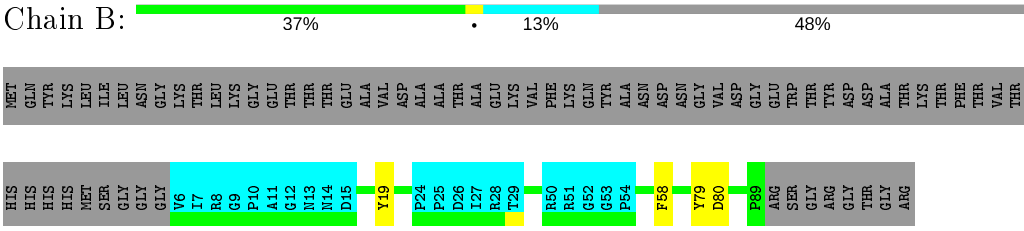
Chain A:



4.2.19 Score per residue for model 19

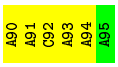
- Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



- Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

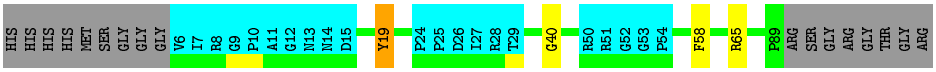
Chain A:



4.2.20 Score per residue for model 20

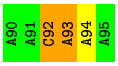
- Molecule 1: Immunoglobulin G-binding protein G,Serine/arginine-rich splicing factor 1

Chain B:



- Molecule 2: RNA (5'-R(*AP*AP*CP*AP*AP*A)-3')

Chain A:



5 Refinement protocol and experimental data overview

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

Of the 50 calculated structures, 20 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
Amber	refinement	
CYANA	structure calculation	

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	844
Number of shifts mapped to atoms	844
Number of unparsed shifts	0
Number of shifts with mapping errors	0
Number of shifts with mapping warnings	0
Assignment completeness (well-defined parts)	72%

No validations of the models with respect to experimental NMR restraints is performed at this time.

6 Model quality

6.1 Standard geometry

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	B	0.59±0.00	0±0/531 (0.0± 0.0%)	0.72±0.01	0±0/716 (0.0± 0.0%)
2	A	1.05±0.02	0±0/143 (0.0± 0.0%)	1.34±0.05	1±1/221 (0.4± 0.3%)
All	All	0.72	0/13480 (0.0%)	0.91	16/18740 (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
2	A	0.0±0.0	0.1±0.3
All	All	0	2

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
2	A	91	A	O4'-C1'-N9	6.27	113.22	108.20	1	3
2	A	90	A	O4'-C1'-N9	5.96	112.97	108.20	8	4
2	A	94	A	O4'-C1'-N9	5.70	112.76	108.20	2	5
2	A	93	A	O4'-C1'-N9	5.62	112.70	108.20	5	3
2	A	95	A	O4'-C1'-N9	5.01	112.21	108.20	7	1

There are no chirality outliers.

All unique planar outliers are listed below.

Mol	Chain	Res	Type	Group	Models (Total)
2	A	93	A	Sidechain	2

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	B	519	482	482	3±2
2	A	127	68	68	2±1
All	All	12920	11000	11000	66

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.

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

Atom-1	Atom-2	Clash(Å)	Distance(Å)	Models	
				Worst	Total
1:B:42:ILE:HD11	1:B:61:PHE:CE2	0.56	2.36	15	3
1:B:84:LEU:HD22	1:B:84:LEU:N	0.54	2.17	18	1
1:B:58:PHE:CE1	2:A:93:A:C4	0.53	2.95	5	4
1:B:58:PHE:CE2	2:A:93:A:C4	0.53	2.97	16	5
1:B:79:TYR:CG	1:B:80:ASP:N	0.52	2.78	17	14
1:B:19:TYR:CE1	2:A:92:C:C6	0.48	3.02	9	10
1:B:42:ILE:HD11	1:B:61:PHE:CZ	0.46	2.46	17	2
1:B:58:PHE:CE2	2:A:93:A:C6	0.46	3.03	17	1
1:B:84:LEU:CD2	1:B:84:LEU:N	0.45	2.79	18	1
1:B:79:TYR:CD2	1:B:80:ASP:N	0.45	2.84	14	2
1:B:56:PHE:CE2	2:A:92:C:H4'	0.45	2.47	5	3
1:B:42:ILE:HD12	1:B:42:ILE:N	0.44	2.27	17	1
1:B:19:TYR:CE2	2:A:92:C:C6	0.44	3.05	2	6
1:B:48:LYS:HE2	1:B:56:PHE:CZ	0.43	2.48	11	2
1:B:58:PHE:CD2	2:A:93:A:C2	0.43	3.07	19	3
1:B:58:PHE:CZ	2:A:93:A:C5	0.42	3.06	5	1
1:B:79:TYR:CD1	1:B:80:ASP:N	0.42	2.88	3	3
2:A:93:A:H5"	2:A:93:A:C8	0.40	2.51	7	1
1:B:42:ILE:N	1:B:42:ILE:HD12	0.40	2.32	9	1
1:B:47:LEU:C	1:B:47:LEU:HD13	0.40	2.37	15	1
1:B:56:PHE:CE1	2:A:92:C:H4'	0.40	2.51	3	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	B	62/161 (39%)	54±2 (87±3%)	7±2 (12±4%)	1±1 (1±1%)	20 68
All	All	1240/3220 (39%)	1078 (87%)	149 (12%)	13 (1%)	20 68

All 5 unique Ramachandran outliers are listed below. They are sorted by the frequency of occurrence in the ensemble.

Mol	Chain	Res	Type	Models (Total)
1	B	40	GLY	5
1	B	80	ASP	4
1	B	76	GLY	2
1	B	77	TYR	1
1	B	49	ASN	1

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	B	54/130 (42%)	53±1 (97±1%)	1±1 (3±1%)	48 90
All	All	1080/2600 (42%)	1051 (97%)	29 (3%)	48 90

All 6 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	B	19	TYR	20
1	B	65	ARG	3
1	B	30	LYS	3
1	B	83	ARG	1
1	B	56	PHE	1

Continued on next page...

Continued from previous page...

Mol	Chain	Res	Type	Models (Total)
1	B	38	LYS	1

6.3.3 RNA

Mol	Chain	Analysed	Backbone Outliers	Pucker Outliers	Suiteness
2	A	5/6 (83%)	3±1 (63±18%)	1±1 (20±15%)	0.09±0.07
All	All	104/120 (87%)	63 (61%)	20 (19%)	0.09

The overall RNA backbone suiteness is 0.09.

All unique RNA backbone outliers are listed below:

Mol	Chain	Res	Type	Models (Total)
2	A	94	A	16
2	A	93	A	14
2	A	95	A	14
2	A	91	A	11
2	A	92	C	8

All unique RNA pucker outliers are listed below:

Mol	Chain	Res	Type	Models (Total)
2	A	94	A	8
2	A	91	A	6
2	A	90	A	4
2	A	93	A	2

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

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

6.5 Carbohydrates

There are no monosaccharides in this entry.

6.6 Ligand geometry

There are no ligands in this entry.

6.7 Other polymers

There are no such molecules in this entry.

6.8 Polymer linkage issues

There are no chain breaks in this entry.

7 Chemical shift validation

The completeness of assignment taking into account all chemical shift lists is 72% for the well-defined parts and 65% for the entire structure.

7.1 Chemical shift list 1

File name: working_cs.cif

Chemical shift list name: *starch_ output*

7.1.1 Bookkeeping

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	844
Number of shifts mapped to atoms	844
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

The following table shows the suggested chemical shift referencing corrections.

Nucleus	# values	Correction ± precision, *ppm*	Suggested action
$^{13}C_{\alpha}$	56	-0.33 ± 0.37	None needed (< 0.5 ppm)
$^{13}C_{\beta}$	64	0.13 ± 0.12	None needed (< 0.5 ppm)
$^{13}C'$	0	—	None (insufficient data)
^{15}N	70	0.34 ± 0.64	None needed (< 0.5 ppm)

7.1.3 Completeness of resonance assignments

The following table shows the completeness of the chemical shift assignments for the well-defined regions of the structure. The overall completeness is 72%, i.e. 655 atoms were assigned a chemical shift out of a possible 911. 8 out of 8 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	^{1}H	^{13}C	^{15}N
Backbone	231/309 (75%)	121/123 (98%)	50/126 (40%)	60/60 (100%)
Sidechain	320/403 (79%)	204/236 (86%)	114/144 (79%)	2/23 (9%)

Continued on next page...

Continued from previous page...

	Total	^{1}H	^{13}C	^{15}N
Aromatic	60/85 (71%)	41/45 (91%)	19/40 (48%)	0/0 (—%)
Overall	655/911 (72%)	410/470 (87%)	183/352 (52%)	62/89 (70%)

The following table shows the completeness of the chemical shift assignments for the full structure. The overall completeness is 65%, i.e. 765 atoms were assigned a chemical shift out of a possible 1168. 8 out of 9 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

	Total	^{1}H	^{13}C	^{15}N
Backbone	268/406 (66%)	142/161 (88%)	56/168 (33%)	70/77 (91%)
Sidechain	393/563 (70%)	254/332 (77%)	135/194 (70%)	4/37 (11%)
Aromatic	60/85 (71%)	41/45 (91%)	19/40 (48%)	0/0 (—%)
Overall	765/1168 (65%)	481/604 (80%)	210/444 (47%)	74/120 (62%)

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.

Random coil index (RCI) for chain B: