# wwPDB NMR Structure Validation Summary Report (i)

Apr 26, 2016 – 07:28 PM BST

| | | |
|---|---|---|
| PDB ID | : | 2CTO |
| Title | : | Solution structure of the HMG box like domain from human hypothetical protein FLJ14904 |
| Authors | : | Tomizawa, T.; Kigawa, T.; Sato, M.; Koshiba, S.; Inoue, M.; Kamatari, Y.O.; Yokoyama, S.; RIKEN Structural Genomics/Proteomics Initiative (RSGI) |
| Deposited on | : | 2005-05-24 |

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
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:

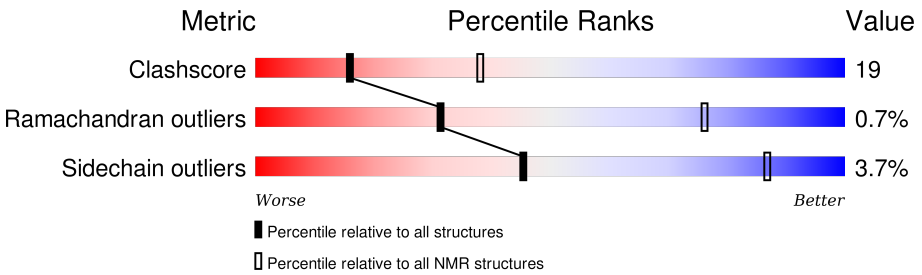
| | | |
|---|---|---|
| Cyrange | : | Kirchner and Güntert (2011) |
| NmrClust | : | Kelley et al. (1996) |
| MolProbity | : | 4.02b-467 |
| Mogul | : | unknown |
| Percentile statistics | : | 20151230.v01 (using entries in the PDB archive December 30th 2015) |
| RCI | : | v_1n_11_5_13_A (Berjanski et 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 90%.

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 | 93 | 43% 18% 39% |

# 2 Ensemble composition and analysis (i)

This entry contains 20 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: *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:16-A:72 (57) | 0.17 | 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 4 clusters and 6 single-model clusters were found.

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

# 3 Entry composition (i)

There is only 1 type of molecule in this entry. The entry contains 1438 atoms, of which 710 are hydrogens and 0 are deuteriums.

- Molecule 1 is a protein called novel protein.

| Mol | Chain | Residues | Atoms | | | | | | Trace |
|---|---|---|---|---|---|---|---|---|---|
| | | | Total | C | H | N | O | S | |
| 1 | A | 93 | 1438 | 453 | 710 | 134 | 138 | 3 | 0 |

There are 13 discrepancies between the modelled and reference sequences:

| Chain | Residue | Modelled | Actual | Comment | Reference |
|---|---|---|---|---|---|
| A | 1 | GLY | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 2 | SER | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 3 | SER | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 4 | GLY | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 5 | SER | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 6 | SER | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 7 | GLY | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 88 | SER | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 89 | GLY | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 90 | PRO | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 91 | SER | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 92 | SER | - | CLONING ARTIFACT | UNP Q96JY0 |
| A | 93 | GLY | - | CLONING ARTIFACT | UNP Q96JY0 |

# 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: novel protein

Chain A: 43% 18% 39%

G1 S2 S3 G4 S5 S6 G7 M8 P9 N10 R11 K12 A13 S14 R15 N16 A17 V22 I26 P27 V36 V39 A40 D41 A42 I43 P44 Y45 C46 S47 S48 D49 W50 L53 E57 A62 G73 K74 D75 P76 G77 P78 S79 E80 K81 Q82 K83 P84 V85 F86 T87 S88

G89 P90 S91 S92 G93

## 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: novel protein

Chain A: 38% 22% • 39%

G1 S2 S3 G4 S5 S6 G7 M8 P9 N10 R11 K12 A13 S14 R15 N16 A17 Y18 Y19 K25 I26 P27 E28 L29 R30 V36 D41 A42 I43 P44 Y45 C46 S47 S48 D49 W50 L53 R54 E55 E56 E57 A62 G73 K74 D75 P76 G77 P78 S79 E80 K81 Q82 K83 P84

V85 F86 T87 S88 G89 P90 S91 S92 G93

# 5 Refinement protocol and experimental data overview (i)

The models were refined using the following method: *torsion angle dynamics, restrainted molecular dynamics.*

Of the 100 calculated structures, 20 were deposited, based on the following criterion: *target function, structures with the lowest energy, structures with the least restraint violations.*

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

| Software name | Classification | Version |
|---|---|---|
| CYANA | structure solution | 2.0.17 |
| CYANA | refinement | 2.0.17 |

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) | BMRB entry 11132 |
|---|---|
| Number of chemical shift lists | 1 |
| Total number of shifts | 1055 |
| Number of shifts mapped to atoms | 1055 |
| Number of unparsed shifts | 0 |
| Number of shifts with mapping errors | 0 |
| Number of shifts with mapping warnings | 0 |
| Assignment completeness (well-defined parts) | 90% |

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 | 481 | 468 | 468 | 18±2 |
| All | All | 9620 | 9360 | 9360 | 357 |

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 19.

5 of 67 unique clashes are listed below, sorted by their clash magnitude.

| Atom-1 | Atom-2 | Clash(Å) | Distance(Å) | Models | |
|---|---|---|---|---|---|
| | | | | Worst | Total |
| 1:A:22:VAL:CG1 | 1:A:39:VAL:HG13 | 0.75 | 2.12 | 8 | 9 |
| 1:A:50:TRP:CE3 | 1:A:53:LEU:HD12 | 0.73 | 2.19 | 19 | 18 |
| 1:A:22:VAL:HG11 | 1:A:39:VAL:HG13 | 0.69 | 1.65 | 6 | 7 |
| 1:A:62:ALA:HB1 | 1:A:66:ARG:NH2 | 0.67 | 2.03 | 14 | 1 |
| 1:A:26:ILE:HD11 | 1:A:39:VAL:N | 0.66 | 2.05 | 20 | 8 |

## 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 | 57/93 (61%) | 52±1 (91±2%) | 5±1 (9±2%) | 0±1 (1±1%) | 31 76 |
| All | All | 1140/1860 (61%) | 1035 (91%) | 97 (9%) | 8 (1%) | 31 76 |

All 2 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 | 47 | SER | 7 |
| 1 | A | 17 | ALA | 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 | A | 47/75 (63%) | 45±1 (96±2%) | 2±1 (4±2%) | 45 87 |
| All | All | 940/1500 (63%) | 905 (96%) | 35 (4%) | 45 87 |

5 of 12 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 | 36 | VAL | 15 |
| 1 | A | 23 | GLN | 4 |
| 1 | A | 49 | ASP | 4 |
| 1 | A | 41 | ASP | 3 |
| 1 | A | 69 | ARG | 2 |

### 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 90% for the well-defined parts and 84% for the entire structure.

## 7.1 Chemical shift list 1

File name: BMRB entry 11132

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 | 1055 |
| Number of shifts mapped to atoms | 1055 |
| Number of unparsed shifts | 0 |
| Number of shifts with mapping errors | 0 |
| Number of shifts with mapping warnings | 0 |
| Number of shift outliers (ShiftChecker) | 6 |

### 7.1.2 Chemical shift referencing (i)

The following table shows the suggested chemical shift referencing corrections.

| Nucleus | # values | Correction ± precision, *ppm* | Suggested action |
|---|---|---|---|
| $^{13}C_{\alpha}$ | 86 | 2.24 ± 0.25 | Should be applied |
| $^{13}C_{\beta}$ | 79 | 2.47 ± 0.11 | Should be applied |
| $^{13}C'$ | 81 | 2.32 ± 0.19 | Should be applied |
| $^{15}N$ | 73 | 0.56 ± 0.26 | Should be applied |

### 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 90%, i.e. 687 atoms were assigned a chemical shift out of a possible 767. 0 out of 7 assigned methyl groups (LEU and VAL) were assigned stereospecifically.

| | Total | $^{1}H$ | $^{13}C$ | $^{15}N$ |
|---|---|---|---|---|
| Backbone | 278/279 (100%) | 111/111 (100%) | 114/114 (100%) | 53/54 (98%) |
| Sidechain | 335/414 (81%) | 211/246 (86%) | 121/141 (86%) | 3/27 (11%) |

*Continued on next page...*

*Continued from previous page...*

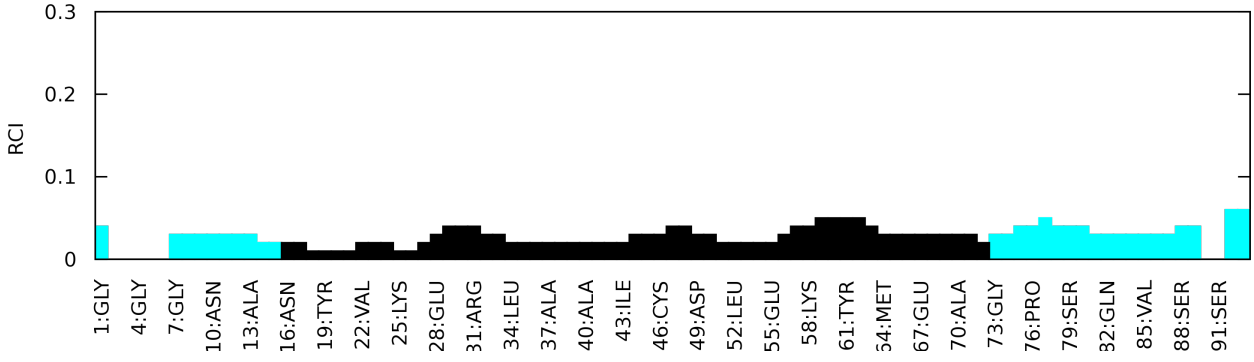
| | Total | $^{1}$H | $^{13}$C | $^{15}$N |
|---|---|---|---|---|
| Aromatic | 74/74 (100%) | 38/38 (100%) | 34/34 (100%) | 2/2 (100%) |
| Overall | 687/767 (90%) | 360/395 (91%) | 269/289 (93%) | 58/83 (70%) |

### 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 | 15 | ARG | HD3 | 0.90 | 4.36 – 1.86 | -8.8 |
| 1 | A | 15 | ARG | HD2 | 1.29 | 4.27 – 1.97 | -8.0 |
| 1 | A | 58 | LYS | HE3 | 1.63 | 3.86 – 1.96 | -6.8 |
| 1 | A | 58 | LYS | HE2 | 1.77 | 3.87 – 1.97 | -6.0 |
| 1 | A | 58 | LYS | HG3 | -0.31 | 2.76 – -0.04 | -6.0 |
| 1 | A | 15 | ARG | HG2 | -0.04 | 2.92 – 0.22 | -6.0 |

### 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: