Structural Bioinformatics

PINE-SPARKY.2 for automated NMR-based protein structure research

Woonghee Lee[*], John L. Markley[*]

National Magnetics Resonance Facility at Madison and Biochemistry Department, University of Wisconsin-Madison, Madison, WI 53706, USA

SUPPLEMENTARY INFORMATION

Supplementary text

Implementation details. The GUI of *PINE-SPARKY.2* is written in Python 2.7 and Tkinter. It uses *NDP-PLOT* for probability visualization. The different tools are integrated mainly through the execution of Python scripts that carry out auto-conversion and auto-launching. For example, the *PINE* web server provides the probabilistic outputs of *PINE*, *PECAN*, and *LACS* in STAR format (Hall, 1991), and *PINE-SPARKY.2* uses these as input for conversion scripts, which generate 2D *NDP-PLOT* graphs that display the probabilities. Another script utilizes *PINE* output to carry out a grid search of the *PACSY* database to detect hydrophobic cores and transfer the result to *NDP-PLOT*. The unique *Key* generated upon PINE submission is formatted as YYMMDD_HHMMSS_RANDOM by date, time and random numbers.

The default sampling size for *CS-Rosetta* runs is 10,000 structures; this value can be changed on the BMRB *CS-Rosetta* web page. We recommend using *CS-Rosetta* only after most of the backbone signals have been assigned and with the pre-assignment option enabled. The *PINE* web server manages the queue for submitted *CS-Rosetta* jobs and also notifies the user when the job is finished. The calculated 3D structures are automatically imported into the same URL that the user received with assignment results from *PINE*. The output from *CS-Rosetta* is a best model and a bundle of the 10 best models in PDB format (Berman *et al.*, 2007).

For structure visualizations, the user can load their 3D structure and type *@pine_pymol, @core_pymol or @rci_pymol* in the *PyMOL* command-line. The color scheme in *PyMO*L for *PINE* assignment probabilities is the same as used in the 2D probability bar graph (Fig. S2B) (green: ≥0.99; cyan: ≥0.85; yellow: ≥0.5; red: <0.5; gray: no assignment). The default color scheme for relative solvent accessible surface (Lee and Richards, 1971) is: red for buried (<0.1), purple for medium accessibility (between 0.1 and 0.3), and blue for exposed to water (>0.3). The color scheme for RCI-S^2 (random coil index order parameter) values are gradient colors between green and red, with green most ordered and red least ordered (Fig. S2C; Fig. S4C).

Details concerning the CS-Rosetta structure of AeSCP-2. Peaks from the 2D ^{1}H,^{15}N-HSQC, 3D CBCA(CO)NH, and HNCACB spectra were picked by *APES* and verified manually by using *Strip Plot* (Shin *et al.*, 2008). Because the protein exhibits a set of peaks from a minor conformer, only the dominant signals were picked (Fig. S3A). The *PINE* assignment job took 3 minutes to complete. By clicking *Check* and *Okay* to a series of interactive questions, *NDP-PLOT* visualizations (Fig. S2A-D) were generated on the basis of answers to the series of interactive questions. Four helices and five strands were detected by *PECAN* (Fig. S2A), and spin system assignments were complete except for the first residue and prolines (Fig. S2B). *LACS* linear analysis of the backbone chemical shifts determined that there was no error in ^{13}C chemical shift referencing (Fig. S2C). Potential chemical shift assignment errors were flagged for I34 and E103; these turned out to be false-positives in the *Strip Plot* tool. F8, F32 and V45 were the three most probable core residues detected (Fig. S2D). 651 of 726 peaks from spectra were automatically assigned, and the resonance assignment completeness for backbone atoms was 98.38 % after generating *PINE* labels and accepting assignments with >0.5 probability with *PINE-SPARKY.2* (Fig. 1G; Fig. S3B). The *CS-Rosetta* 3D structure calculation was finished three days later. The notification email from the *PINE* web server contained the *Key* identifier; by entering this *Key* and clicking *Check*, the *PINE-SPARKY* importer automatically downloaded csrosetta_url.html, csrosetta_best.pdb, csrosetta_best_10.pdb, csrosetta_url.html and csroset-ta_best_10_pdb.zip files into the *PINE* sub-directory. We used *@pine_pymol*, *@core_pymol* and *@rci_pymol* to color *PINE* assignment probabilities (Fig. S4A), *PACSY* hydrophobicities (Fig. S4B) and RCI S^2 values (Fig. S4C) on the 3D structure. The *CS-Rosetta* structure closely matched the structure from manually assigned data based on NOE restraints and deposited in the PDB (Entry Number 2KSH). We superimposed the best *CS-Rosetta* structure (csrosetta_best.pdb) on the 2KSH structure: the pairwise backbone r.m.s.d was 1.21 Å, and the all-heavy-atom r.m.s.d was 2.14 Å for structured regions.

Supplementary figures

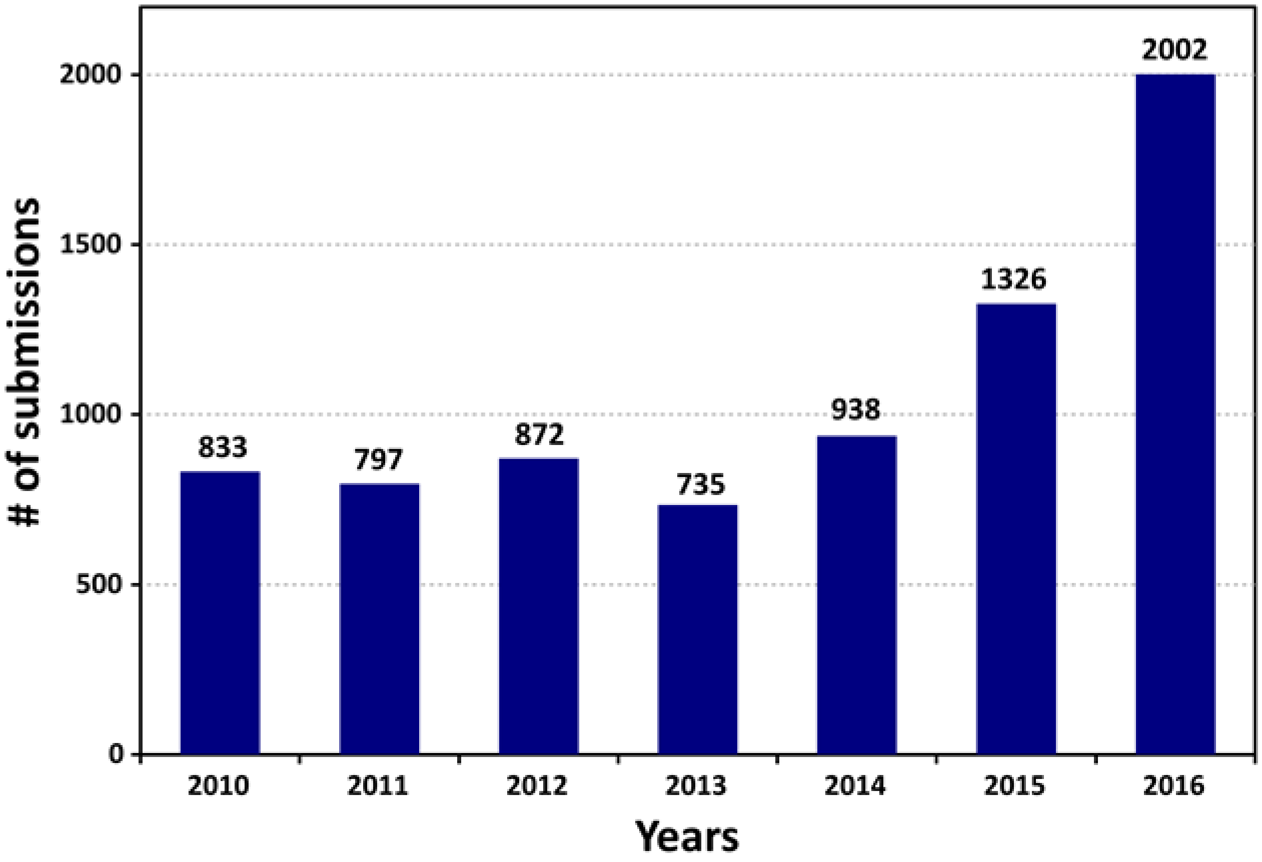

Fig. S1. Growth in the utilization of the *PINE* web server over time. In 2016, *PINE* served 2,002 jobs for the biomolecular NMR community. According to statistics from BMRB entries, *PINE* is one of the most popular computer algorithms for chemical shift assignment.

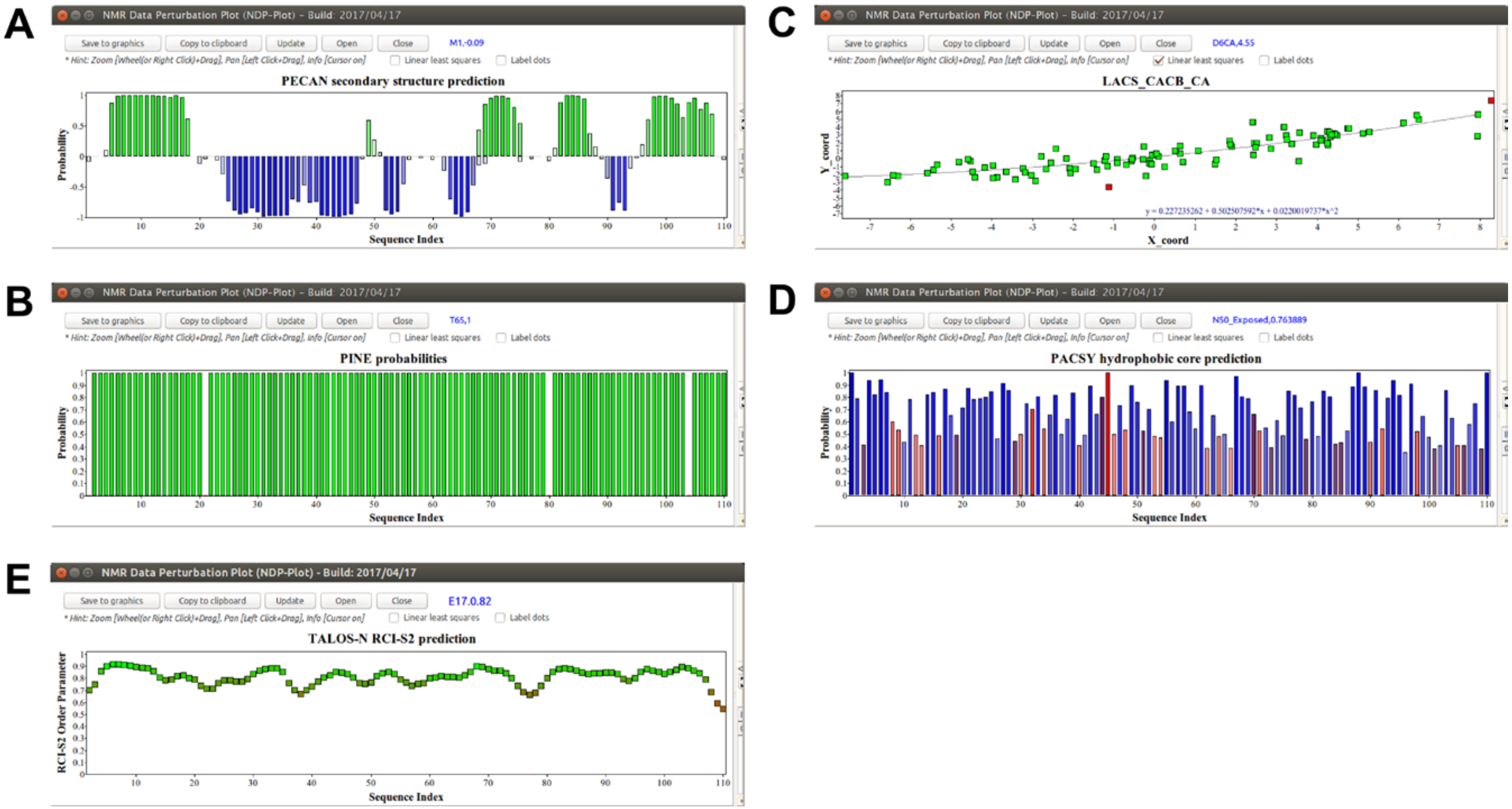
A
NMR Data Perturbation Plot (NDP-Plot) - Build: 2017/04/17
Save to graphics
Copy to clipboard
Update
Open
Close
M1,-0.09
* Hint: Zoom [Wheel(or Right Click)+Drag], Pan [Left Click+Drag], Info [Cursor on]
Linear least squares
Label dots
PECAN secondary structure prediction
Probability
Sequence Index
B
NMR Data Perturbation Plot (NDP-Plot) - Build: 2017/04/17
T65,1
PINE probabilities
Probability
Sequence Index
C
NMR Data Perturbation Plot (NDP-Plot) - Build: 2017/04/17
D6CA,4.55
LACS_CACB_CA
Y_coord
X_coord
y = 0.227235262 + 0.502507592*x + 0.0220019737*x^2
D
NMR Data Perturbation Plot (NDP-Plot) - Build: 2017/04/17
N50_Exposed,0.763889
PACSY hydrophobic core prediction
Probability
Sequence Index
E
NMR Data Perturbation Plot (NDP-Plot) - Build: 2017/04/17
E17,0.82
TALOS-N RCI-S2 prediction
RCI-S2 Order Parameter
Sequence Index

Fig. S2. *NDP-PLOT* visualizations launched automatically by the *PINE-SPARKY.2* plug-in to *NMRFAM-SPARKY*. **A**. Secondary structure per residue from the *PECAN* algorithm (green, helix; blue, strand). **B**. *LACS* (Linear Analysis of Chemical Shifts) indicating 0 ppm referencing error and potential errors for I34 and E103 in red. **C**. Spin system assignment probabilities from the *PINE* algorithm (green, ≥ 0.99; cyan, ≥ 0.80; yellow, ≥ 0.50; red, < 0.50; gray, no available assignment). Gaps indicate proline residues **D**. Hydrophobic core residue detected by *PACSY* DB search (red, buried; purple, medium; blue, exposed). F8, F32 and V45 are found to be the three most probable core residues. **E.** RCI S^2 (random coil index order parameter) per residue from TALOS-N.

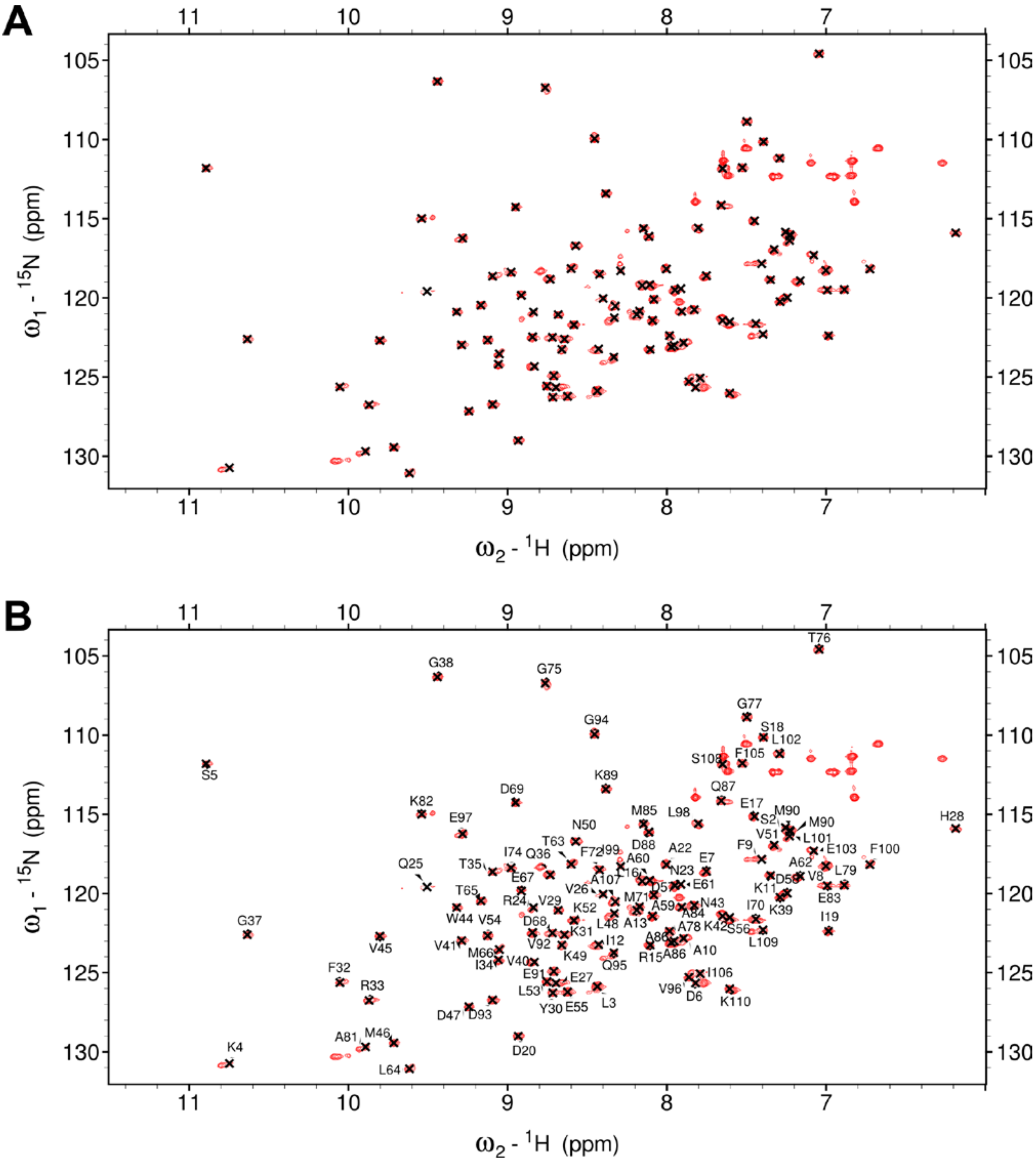

Fig. S3. ^{1}H,^{15}N-HSQC spectra from AeSCP-2 before and after running *PINE-SPARKY.2*. **A**. Peaks were identified by *APES* (two-letter-code *ae*) and manually verified by the *Integrative NMR* approach using *Strip Plot* (two-letter-code *sp*). Side chain peaks and weaker peaks from minor conformers were not picked. **B**. *PINE-SPARKY.2* carried out automated assignment of all the selected peaks in the spectrum in three minutes.

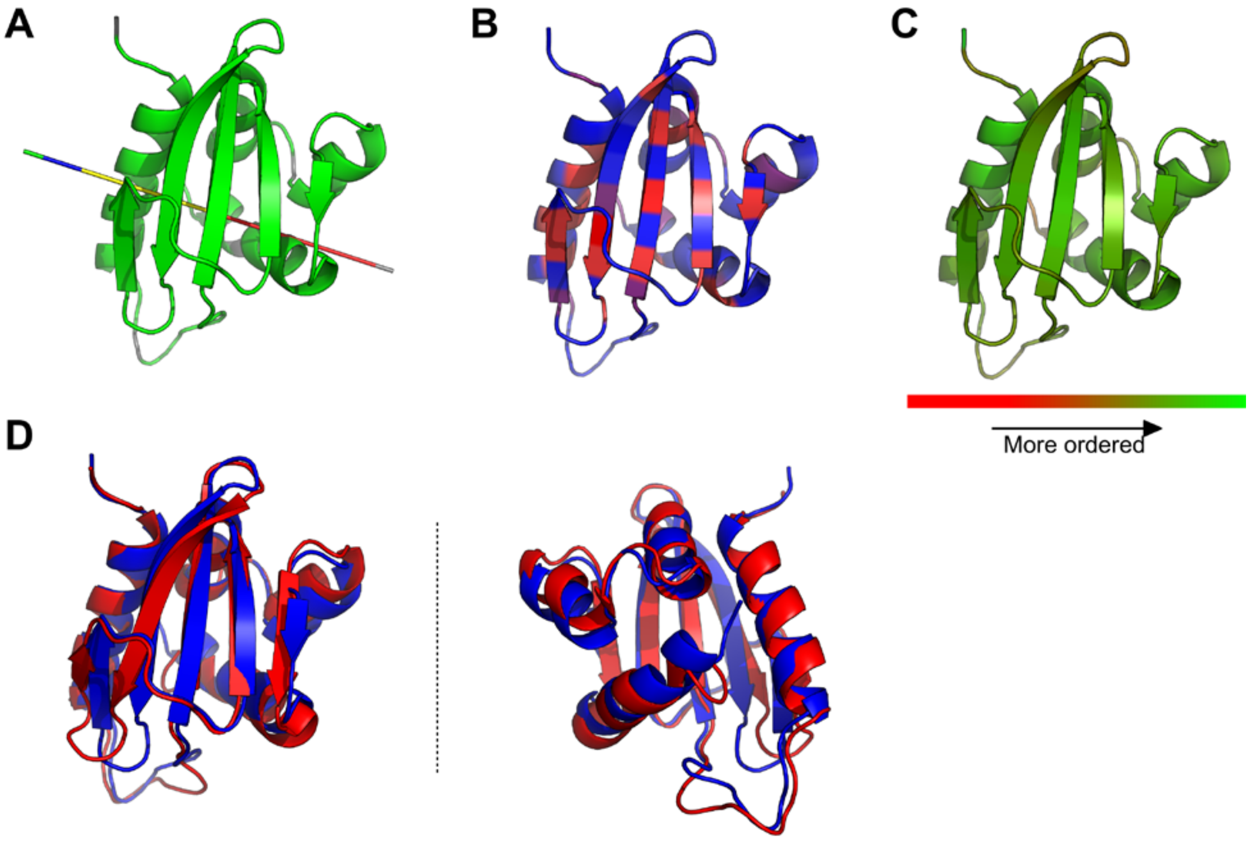

Fig. S4. 3-D structure of AeSCP-2 generated by *CS-Rosetta* on the basis of backbone chemical shifts as launched by *PINE-SPARKY.2* and visualized by *PyMOL*. **A**. Structure colored according to *PINE* assignment probabilities by using the *@pine_pymol* command provided by PINE-SPARKY.2. The green color indicates assignment probabilities ≥0.99. **B**. Hydrophobic residues detected by *PACSY* DB visualized by the *@core_pymol* command. Color codes are the same as in Fig. S2D. **C**. RCI-S^2 representing protein flexibility is predicted by TALOS-N visualized by the *@rci_pymol* command. **D**. Lowest energy model of AeSCP-2 from *CS-Rosetta* (red) superimposed on the medoid conformer (blue) of the deposited structure determined from NOE data (PDB 2KSH). Backbone and all heavy atom pairwise r.m.s.d.s were 1.21 Å and 2.14 Å, respectively.

Supplemental References

Berman, H.*, et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**(Database issue):D301-303.

Hall, S.R. (1991) The STAR File: A New Format for Electronic Data Transfer and Archiving. *Journal of Chemical Information and Computing Sciences* **31**:326-333.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

Shin, J., Lee, W. and Lee, W. (2008) Structural proteomics by NMR spectroscopy. *Expert Rev Proteomics* **5**(4):589-601.