BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 12 2011, pages 1727–1728
doi:10.1093/bioinformatics/btr200

Data and text mining

Advance Access publication April 21, 2011

PONDEROSA, an automated 3D-NOESY peak picking program, enables automated protein structure determination

Woonghee Lee[1,2,*], Jin Hae Kim[3], William M. Westler[1] and John L. Markley[1,2,3,*]

[1]National Magnetic Resonance Facility at Madison, [2]Biochemistry Department and [3]Graduate Program in Biophysics, University of Wisconsin-Madison, Madison, WI 53706, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: PONDEROSA (Peak-picking Of Noe Data Enabled by Restriction of Shift Assignments) accepts input information consisting of a protein sequence, backbone and sidechain NMR resonance assignments, and 3D-NOESY (^{13}C-edited and/or ^{15}N-edited) spectra, and returns assignments of NOESY crosspeaks, distance and angle constraints, and a reliable NMR structure represented by a family of conformers. PONDEROSA incorporates and integrates external software packages (TALOS+, STRIDE and CYANA) to carry out different steps in the structure determination. PONDEROSA implements internal functions that identify and validate NOESY peak assignments and assess the quality of the calculated three-dimensional structure of the protein. The robustness of the analysis results from PONDEROSA's hierarchical processing steps that involve iterative interaction among the internal and external modules. PONDEROSA supports a variety of input formats: SPARKY assignment table (.shifts) and spectrum file formats (.ucsf), XEASY proton file format (.prot), and NMR-STAR format (.star). To demonstrate the utility of PONDEROSA, we used the package to determine 3D structures of two proteins: human ubiquitin and *Escherichia coli* iron-sulfur scaffold protein variant IscU(D39A). The automatically generated structural constraints and ensembles of conformers were as good as or better than those determined previously by much less automated means.

Availability: The program, in the form of binary code along with tutorials and reference manuals, is available at http://ponderosa.nmrfam.wisc.edu/.

Contact: whlee@nmrfam.wisc.edu; markley@nmrfam.wisc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 24, 2011; revised on April 6, 2011; accepted on April 7, 2011

1 INTRODUCTION

A major challenge of structural biology is to close the gap between known sequences of proteins [$>1\times10^8$ in GenBank (Benson *et al.*, 2008)] and their 3D structures ($\sim1\times10^5$ in PDB; Berman *et al.*, 2000). Automation now plays a key role in speeding up the determination of protein structures by X-ray crystallography. However, the determination of protein structures by NMR spectroscopy includes a larger number of steps that present greater challenges for automation. The steps basically are sequential; however, some of them may need to be iterated in order to yield a satisfactory protein structure. Software packages have been developed to automate individual steps, and in some cases to pipeline several steps (Bahrami *et al.*, 2009; Lopez-Mendez and Güntert, 2006). One of the challenges has been to automate the final steps beyond backbone and sidechain peak assignment, including the determination of torsion angle constraints, the assignment of NOESY cross peaks and the determination of distance constraints, the analysis of secondary structure, and the calculation of a validated 3D protein structure. The PONDEROSA (Peak-picking Of Noe Data Enabled by Restriction of Shift Assignments) software package described here bridges this gap and is meant to be used with an automated resonance assignment package such as PINE-NMR introduced earlier by our group (Bahrami *et al.*, 2009).

*To whom correspondence should be addressed.

2 IMPLEMENTATION

PONDEROSA (Supplementary Fig. S1A) accepts resonance assignments in popular file formats (SPARKY; T.D. Goddard and D. G. Kneller, SPARKY 3; University of California, San Francisco, XEASY; Bartels *et al.*, 1995 or NMR-STAR; http://www.bmrb.wisc.edu/dictionary/), an amino acid sequence file in either one- or three-letter code, and ^{13}C-NOESY and/or ^{15}N-NOESY datasets in SPARKY (.ucsf) format. By integrating internal functions and external programs, PONDEROSA provides as output NOE peak lists, NOE assignments, structural constraints and a family of conformers representing the 3D structure.

Internal functions: The major internal functions of PONDEROSA simulate and validate NOESY peaks and manage interactions among the internal and external software routines. PONDEROSA uses available resonance assignments to simulate all possible short, medium- and long-range peaks (Supplementary Figs S2 and S3). Members of the set of simulated peaks are validated by comparing them to peaks detected in the experimental NOESY datasets under different threshold levels. The sets of validated peak lists are provided to the external programs that determine torsion angle restraints, assign NOESY peaks, calculate structures and analyze secondary structure. The results from these programs are recycled to PONDEROSA for the next iteration (Supplementary Fig. S4).

PONDEROSA examines the effect of the threshold level on a structural quality score that incorporates the root mean standard deviation (RMSD) of backbone atoms in structured regions as determined by STRIDE, the number of constraint and van der Waals violations, and number of residues in favored and disallowed Ramachandran regions. If both ^{13}C- and ^{15}N-edited NOESY

data are present, PONDEROSA interactively determines optimal thresholds for each.

External programs: PONDEROSA interacts with TALOS+ (Shen *et al.*, 2009) for identifying structured regions and for determining torsion angle restraints from assigned chemical shifts, STRIDE (Frishman *et al.*, 1995) for analyzing secondary structure, and CYANA (Güntert, 2004) for assigning NOESY cross peaks and calculating 3D structures.

Graphical User Interface: An intuitive graphical user interface (Supplementary Fig. S1B) enables specification of the number of CPU nodes, steps and cycles to be used in CYANA iterations, the limit on the number of NOESY peaks to be searched for on the basis of local peak maxima, and the weighting factors for RMSD distance violations and torsion angle dispersions.

3 RESULTS AND CONCLUSION

We selected two proteins to illustrate the use of PONDEROSA for NOESY peak picking and automated structure determination: human ubiquitin (76 residues) and *Escherichia coli* iron–sulfur scaffold protein variant IscU(D39A) (128 residues). We chose human ubiquitin because it is a well-known test sample for protein NMR technology development with 3D structures deposited in the Protein Data Bank (PDB), e.g. 1D3Z (Cornilescu *et al.*, 1998). We chose IscU(D39A) (Kim *et al*., 2009) because it is a larger protein with a recently deposited non-automatically derived NMR structure (PDB 2KQK) that exhibited variation in the position of secondary structural elements within the family of 20 conformers. In determining the structures of both proteins, we used ^{1}H-^{15}N HSQC, ^{1}H-^{13}C HSQC, CBCA(CO)NH, HNCACB and HBHA(CO)NH datasets for backbone assignments, and (H)CC(CO)NH, H(CC)(CO)NH and HCCH-TOCSY datasets for sidechain assignments. We used NMRpipe (Delaglio *et al.*, 1995) to process all spectra and then converted the spectra to SPARKY (.ucsf) files. We used PINE-NMR and PINE-SPARKY (Lee *et al.*, 2009) to assign the spectra of human ubiquitin, but assigned IscU(D39A) by a manual assignment strategy. We processed 3D ^{13}C-NOESY and ^{15}N-NOESY datasets with NMRPipe and converted the spectra to .ucsf files for input to PONDEROSA. The total times required for the structure determinations with 24 CPUs were 9 h for human ubiquitin and 15 h for IscU(D39A).

The 20 best conformers of human ubiquitin determined by PONDEROSA (Supplementary Fig. S1C) had a RMSD of 0.09 Å for backbone atoms and 0.48 Å for all heavy atoms in structured regions. The 20 best conformers of IscU(D39A) determined by PONDEROSA had an RMSD of 0.20 Å for backbone atoms and 0.61 Å for all heavy atoms in structured regions. The structures determined by PONDEROSA were very similar to those determined earlier by more manual approaches: 1.15 Å RMSD for the backbone atoms of human ubiquitin (PONDEROSA versus 1D3Z) and 1.30 Å for structured backbone atoms IscU(D39A) (PONDEROSA versus 2KQK) (Supplementary Fig. S1D). Analysis by two standard validation suites, PSVS (Bhattacharya *et al*., 2007) and iCing (http://nmr.cmbi.ru.nl/icing/#welcome), revealed that the PONDEROSA-derived structures were of equivalent quality to the structures of the same proteins in the Protein Data Bank (1D3Z and 2KQK) determined by less automated means.

Funding: National Institutes of Health (grant number P41 RR02301).

Conflict of Interest: none declared

REFERENCES

Bahrami,A. *et al.* (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput. Biol.*, **5**, 1–12.

Bartels,C. *et al.* (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR*, **6**, 1–10.

Benson,D. *et al.* (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.

Berman,H. *et al.* (2000) The Protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhattacharya,A. *et al*. (2007) Evaluating protein structures determined by structural genomics consortia. *Proteins*, **66**, 4, 778–795.

Cornilescu,G. *et al.* (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.*, **120**, 6836–6837.

Delaglio,F. *et al.* (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.

Frishman,D and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Genetics*, **23**, 566–579.

Güntert, P. (2004) Automated NMR structure calculation with CYANA. *Methods Mol. Biol.*, **278**, 353–378.

Kim,J.H. *et al.* (2009) Structure and dynamics of the iron-sulfur cluster assembly scaffold protein IscU and its interaction with the cochaperone HscB. *Biochemistry*, **48**, 6062–6071.

Lee,W. *et al.* (2009) PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy. *Bioinformatics*, **25**, 2085–2087.

Lopez-Mendez,B. and Güntert,P. (2006) Automated protein structure determination from NMR spectra, *J. Am. Chem. Soc.*, **128**, 13112–13122.

Shen,Y. *et al.* (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, **44**, 4, 213–223.