Center for Eukaryotic Structural Genomics

Protein Structure Initiative

Pine and HIFI-C: Robust Assignment and Fast Data Collection in NMR

Arash Bahrami, Gabriel Cornilescu, Marco Tonelli, John L. Markley, and Hamid R. Eghbalnia

University of Wisconsin-Madison, Department of Biochemistry, 433 Babcock Drive, Madison, Wisconsin, USA 53706-1549, http://www.uwstructuralgenomics.org

Abstract

"Probabilistic Interaction Network of Evidence" (PINE) demonstrates a significant advance toward our goal of building an automated framework for NMR structure determination. PINE uses a probabilistic model to combine experimental data with derived information from databases and uses probabilistic rules to reach self-consistent results. PINE integrates modules, which can be used individually as standalone tools, into a probabilistic paradigm to provide adaptive and self-correcting interactions in the steps of structure determination. PINE outperforms the isolated modules (in terms of completeness and correctness) and supersedes them by offering new capabilities and streamlined operation. PINE, which accepts multiple standard formats for the peak lists used as input data, provides a complete and robust probabilistic model of the protein up to the level of complete chemical shift assignment and secondary structure. PINE supports several output formats. The automated computation, which seldom takes more than one hour, provides results comparable to those achievable by an expert over a period of days or weeks. A beta-release version of PINE, available online since July, 2006, has been received positively by users. In addition to input from traditional data collection methods, PINE can accept peak lists derived directly from our automated data collection framework (HIFI-NMR[1]), without the need for data processing and peak picking. PINE provides the proof of concept for our goal of incorporating all steps in protein structure determinations into a probabilistic framework in which results from NMR data analysis can drive the data collection strategy.

"High-resolution Iterative Frequency Identification of Couplings"[5] (HIFI-C) is a novel approach for the rapid collection and processing of multi-dimensional NMR coupling data. By extending the adaptive and intelligent data collection approach introduced earlier in HIFI NMR[1], HIFI-C collects one or more optimal two-dimensional (2D) planes, identifies peaks, and determines couplings with high resolution and precision with the following features: (1) Adaptive tilted plane data collection provides an intelligent trade off between data collection time and accuracy; (2) Data from independent planes can provide a statistical measure of reliability for each measured coupling; (3) Fast data collection enables measurements in cases where sample stability is a limiting factor; (4) For samples that are stable, fast data collection enables more accurate determinations of couplings, particularly for larger biomolecules.

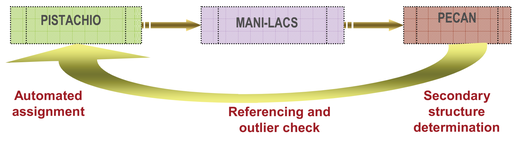
PINE

PINE (Probabilistic Interaction Network of Evidence) represents the realization of a novel algorithm for complete and fully automated backbone and sidechain assignment, secondary structure determination, detection and correction of possible referencing error, and outlier detection. PINE improves, integrates, and outperforms the individual tools on which it is based: PISTACHIO[2], PECAN[3] and LACS[4]. The input to PINE can be a any subset of following NMR experiments:

- HSQC
- HNCO
- CBCA(CO)NH
- HNCAB
- HN(CO)CACB
- HN(CO)CA
- HN(CA)CO
- HNCA
- HN(CO)(CA)CB
- HN(CA)CB
- C(CO)NH
- HBHA(CO)NH
- H(CCO)NH
- HCCH-TOCSY

Furthermore, the probabilistic framework developed for PINE enables the incorporation of other available prior information into the process, such as prior assignments, information from selective labeling, or previously-assembled spin systems.

The following picture shows the iterative steps carried out by the current version of PINE:



In order to establish the fully probabilistic network used by PINE, it was necessary to redefine every variable within each individual step in random (probabilistic) terms. PINE establishes the whole network by considering the conditional dependence of the random variables, and carries out an iterative updating probabilities until they converge to the final solution.

[1] Hamid R. Eghbalnia, Arash Bahrami, Marco Tonelli, Klaus Hallenga, and John L. Markley (2005) *J. Am. Chem. Soc.*, 127(36) 12528 – 12536.

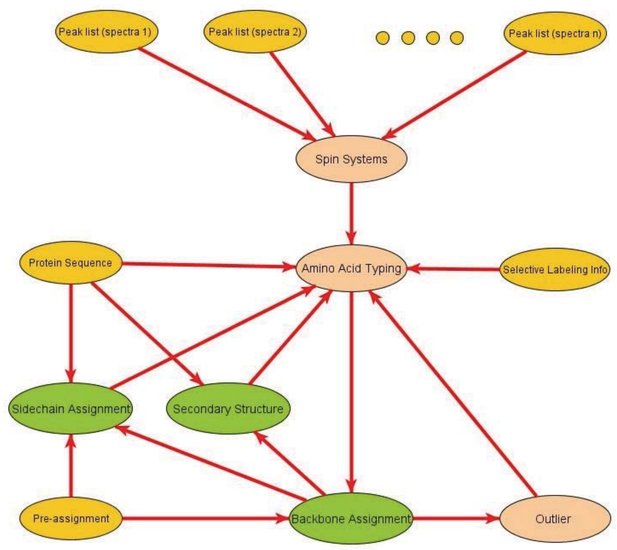
[2] Hamid R. Eghbalnia, Arash Bahrami, Liya Wang, Amir Assadi, and John L. Markley (2005) *J. Biomol. NMR*, 32(3):219-233.

[3] Hamid R. Eghbalnia, Liya Wang, Arash Bahrami, Amir Assadi, and John L. Markley (2005) *J. Biomol. NMR*, 32(1):71-81.

[4] Liya Wang, Hamid R. Eghbalnia, Arash Bahrami, and John L. Markley (2005) *J. Biomol. NMR*, 32(1):13-22.

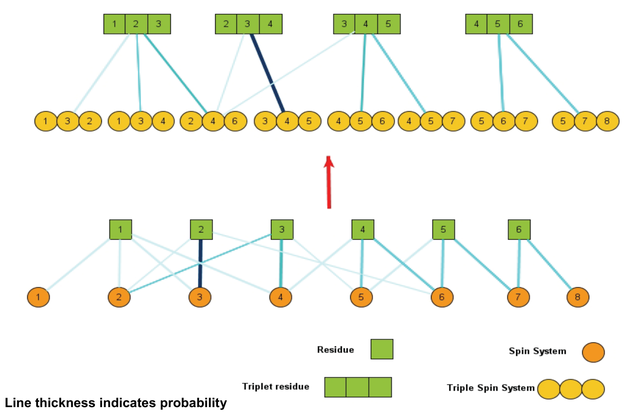
[5] Gabriel Cornilescu, Arash Bahrami, Marco Tonelli, John L. Markley, Hamid R. Eghbalnia. *Submitted.*

General Overview of Probabilistic Network Defined by PINE

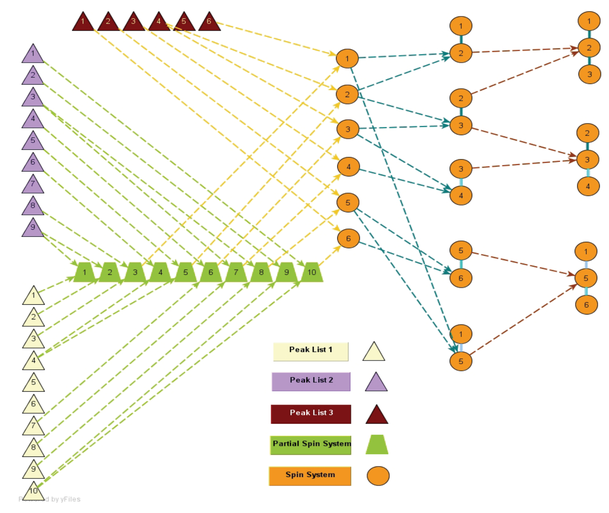


The actual probabilistic network developed by PINE is more complicated than indicated above, because80
every variable presented in the figure above has its own probabilistic network. Examples of such networks (Amino Acid Typing Network and Spin System Generation Network) are shown below.

Amino Acid Typing Network



Spin System Generation Network

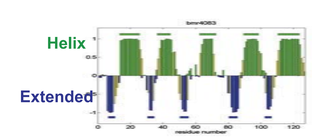


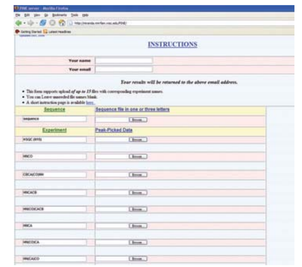
PINE Web Server

PINE is available for public use through a web server at http://miranda.nmrfam.wisc.edu/PINE/ .

The PINE server is exceptionally convenient to use. No installation or parameter setting is required, because all the parameters are automatically analyzed according to the quality and consistency of the uploaded peak lists. The program also can be run as a single executable file for the purpose of integration to another software tool. The PINE server supports different input and output formats including Sparky, XEasy, BMRB (NMR-STAR). PINE also accepts prior information, including prior assignments, peak identifications derived from selective labeling, and assembled spin systems.

Probabilistic Secondary Structure





Native Probabilistic PISTACHIO Output

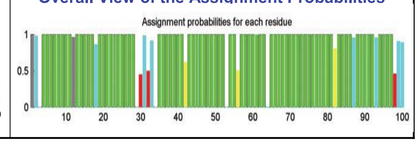
Residue_Name	P(H,N)	H	N	CO	CA	CB	P(H,N) P(no_assignment)	H	N	P(H,N)	H	N	P(H,N)	H	N
1 MET	0.000	0.000	0.00	0.00	55.29	34.51	0.000	0.000	0.00	0.000	0.00	0.000	0.000	0.00	0.000
2 ASN	0.691	9.899	125.16	0.00	52.03	40.68	0.214	9.432	116.54	0.000	0.000	0.00	0.000	0.000	0.00	0.095
3 THR	1.000	9.121	116.72	0.00	59.37	63.99	0.000	0.000	0.00	0.000	0.000	0.00	0.000	0.000	0.00	0.000
4 VAL	1.000	7.977	127.97	0.00	61.66	36.07	0.000	0.000	0.00	0.000	0.000	0.00	0.000	0.000	0.00	0.000
5 CYS	1.000	8.310	126.57	0.00	59.14	31.70	0.000	0.000	0.00	0.000	0.000	0.00	0.000	0.000	0.00	0.000

...

NMR-Star Format

1	1	MET	CA	C	55.291	1.000	0
2	1	MET	CB	C	34.509	1.000	0
3	2	ASN	N	N	125.160	1.000	0
4	2	ASN	H	H	9.899	1.000	0
5	2	ASN	CA	C	52.031	1.000	0
6	2	ASN	CB	C	40.684	1.000	0
7	3	THR	N	N	116.723	1.000	0

Overall View of the Assignment Probabilities

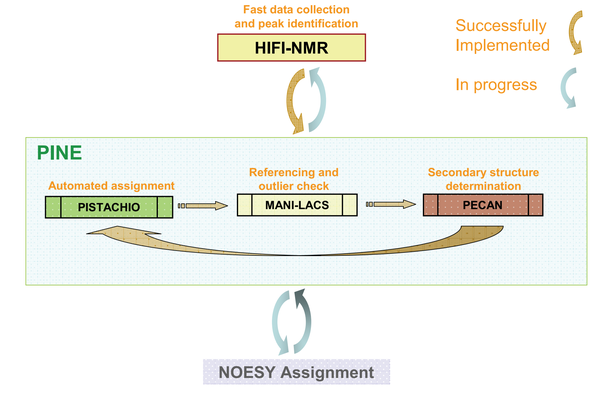


Statistics of PINE Users Location

Total number of jobs submitted Since July 2006: 319 jobs

PINE and HIFI Integration

The overall goal of the project is to provide an interactive probabilistic framework for the whole process of NMR structure determination of proteins. The following picture presents the integration of fast data collection and resonance assignment (Backbone, sidechain, and NOESY assignment) without any manual intervention. We have tested the implemented part (brown arrows) for three proteins of various sizes.

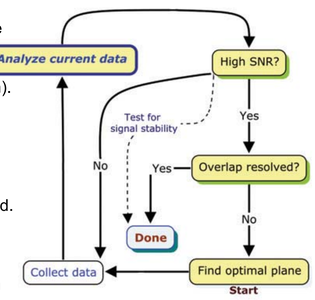


	Data Collection Time (HIFI)	Assignment Time (PINE)	Accuracy of the Assignment
Brazzein 53 aa	12 Hours	5 Min	98%
Ubiquitin 76 aa	14 Hours	5 Min	98%
Flavodoxin 176 aa	48 Hours	1 Hour	80%

HIFI–C: A Fast and Robust Method for Determining NMR Couplings from Adaptive 3D to 2D Projections

The accurate determination of small couplings is frequently of importance in biomolecular NMR spectroscopy. Applications include the collection of dihedral angle constraints, the identification and quantification of trans hydrogen bond couplings, and the measurement of residual dipolar couplings (RDCs). By building on the methods proposed in HIFI-NMR (adaptive, tilted-plane data collection), we have developed a novel method (HIFI-C[5]) for the fast collection and analysis of coupling data. The underlying principle behind HIFI-C is to combine the high digital resolution provided by 2D spectra with the ability of tilted-plane data collection to separate overlapped peaks. HIFI-C works very well with the quantitative J method that encodes the coupling in the peak amplitudes, thus avoiding the additional overlap present in spectra with chemical shift encoded couplings. HIFI-C uses the HIFI-NMR approach to adaptively acquire one or more tilted 2D planes with the highest likelihood of obtaining the needed coupling information by minimizing peak overlap. Because the data for each plane are collected and processed independently, results from multiple planes can be used to ascertain the reliability and robustness of individual RDC values. The time savings result from automated termination of data collection at the point where additional acquisitions would not improve the analysis.

The adaptive decision process for optimal data collection starts with the selection of the optimal plane with minimal peak overlap (this may be known from prior HIFI data collection). Data for this plane are collected and analyzed for satisfactory signal-to-noise and for the presence if signal degradation between early and later scans. It the sample is shown to be unstable, data collection is terminated. If the sample is robust, the choice between collecting additional scans for the current plane or collecting a plane at a different angle is made on the basis of the degree of peak overlap. The process continues until a complete measure of accuracy and precision for each data point has been determined.

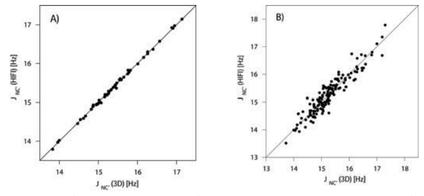


To illustrate the applicability of the HIFI-C approach, we demonstrate its use in determining RDC values for the N-C′ vector. This is one of the smallest dipolar couplings commonly measured, and its determination requires the highest measurement precision.

To enable benchmarking and comparison of results from standard 3D and HIFI-C experiments, data for each protein were collected in 3D and HIFI mode in both isotropic and aligned states.

We selected three proteins for testing the HIFI-C method. GB3 is a benchmark protein of 55 amino acids with available high-resolution crystal and high quality NMR structures; this made it possible to compare how RDC values determined by HIFI-C and 3D agreed with the structures. The other two were examples of medium-sized proteins: PRP24-12 (149 residues) and At5g22580.1 (111 residues, a 25 kDa homodimer).

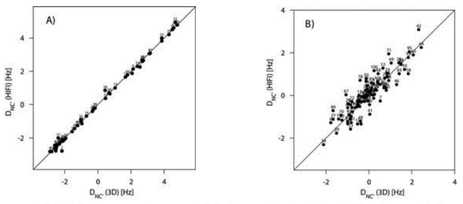
Comparison in Isotropic Conditions



Correlation and RMSD comparison of couplings collected by HIFI-C and 3D. Agreement between the two was within experimental error.

(A) GB3 protein (R = 99.8%, rmsd = 0.2 Hz). The total data collection times were 1.7 h for HIFI-C and 7.9 h for 3D.
(B) PRP24-12 protein (R = 94.0%, rmsd = 0.2 Hz). The total data collection times were 14.6 h for HIFI-C and 44.1 h for 3D .

Comparison in Anisotropic (Aligned) State



Correlation and RMSD comparison of RDCs collected by HIFI-C and 3D. Agreement between the two was within experimental error.

(A) GB3 protein (R = 0.998, rmsd = 0.15 Hz). The total data collection times were 3.1 h for HIFI-C and 17.5 h for 3D.
(B) At5g22580 25 kDa homodimeric protein (R = 0.940, rmsd = 0.6 Hz). The total data collection times were 48.6 h for HIFI-C and 63.1 h for 3D.

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative grant number U54 GM074901.
HRE is supported by NLM grant 5K22LM8992.