Automation for NMR-based Structural Proteomics

Hamid Eghbalnia, Arash Bahrami, Liya Wang, Marco Tonelli, Min S. Lee, Peter Lee, Klaas Hallenga,
Robert Tyler, Eldon L. Ulrich, W. Milo Westler, Zsolt Zolnai, and John L. Markley

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

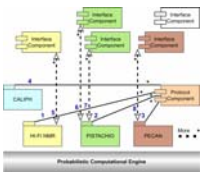
ABSTRACT

The overall goal of this collaboration between the National Magnetic Resonance Facility at Madison and the Center for Eukaryotic Structural Genomics is the seamless integration under the Sesame web-based software system of steps leading from NMR data collection to structure determination and data deposition. The platform will include (fast) data collection, data processing, automated assignment and secondary structure determination, structure determination and validation, and data deposition. We present novel developments: an interactive approach to fast data collection with peak identification ('HiFi-NMR'), a novel algorithm ('PISTACHIO') for the probabilistic assignment of NMR spectral data to backbone and side chain atoms, software ('PECAN') for determining protein secondary structure from chemical shifts and peptide sequence, and ('Caliph'), the Sesame module for organizing data depositions to the RCSB. PISTACHIO and PECAN are available for general use from the NMRFAM web site.

MANI MODEL

MANI is a model for High-throughput Structure determination from sparse data that is based on a rigorous mathematical substrate. The probabilistic computational substrate unifies the use of probabilities across various components of the system. The components interoperate using a protocol that is mainly based on NMR-STAR 3.0 data model. The overall model achieves important objectives important to any automation system that include:

- A rigorous approach with internal consistency checks
- A data model consistent with developing practices i.e. NMR-STAR 3.0
- Ability to add/remove/modify individual components to tailor/improve the system
- Ability to integrate with LIMS, for example, SESAME
- Ability to share standalone useful components



The figure illustrates the flexible integration model of MANI components.

The overall model for the MANI system is simple. A number of computational models (HI-FI NMR, PISTACHIO, etc) use the probabilistic computational engine to give probabilities for solution configurations. Each component communicates with others by passing data and/or commands. Format of the communicated data is based on the NMR-STAR 3.0. The protocol component is simply a dictionary that integrates various components. The CALIPH module is integrated using the same simple protocol. Any module can have any number of interface components. All modules can be used as standalone applications when needed. Examples of these standalone modules are MANI-LACS and MANI-ANSA and PEANUT.

The common use of the underlying probabilistic computational engine enables the enforcement of internal consistency rules. The consistency rules are enforced by iterating over various steps of the solution until results agree (within a probabilistic threshold). An example of this process is depicted at the top of the next column. In this example, assignments are checked for outliers and referencing problems and used to determine secondary structure. This information is used to refine probabilities in the computational model. This refinement is used to complete a subsequent assignment, reference check, and secondary structure identification. This process is repeated until no significant changes in probabilities are observed.

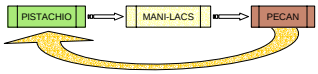


Figure illustrating the basic idea of internal consistency checks where results are iterated to achieve constant probability values across iterations. Additional components will be included as additional tools based on the probabilistic computational engine become available.

HI-FI NMR

Fast data collection and signal recognition

The objective for the High-resolution Iterative Frequency Identification for NMR (HiFi-NMR) effort is to develop methods and tools for faster collection and processing of multidimensional NMR data, while maintaining, or potentially improving, the resolution of data. Portions of this development have synergistic impact on approaches for automated assignments and structure determinations.

Our approach is based on the experimental approach of Kupce and Freeman. We collect data from two orthogonal planes of 3D data set, add an additional plane adaptively, detect the peaks and determine which next plane that would be the "most informative". Planes are added iteratively until the positions of all peaks have been distinguished. After the data are collected, an offline algorithm that uses statistical methods carry out an "optimal" reconstruction from the collected planes processes them. A modification of the HiFi-NMR approach can recover 4D data from a pair of 3D experiments, for example when 4D NOESY data are required.

Protein name	CalC	CalC
Number of residues	101	101
Experiment	CBCA(CO)NH	HNCO
Number of planes required	4	2
Percent peak recovery	99%	100%+

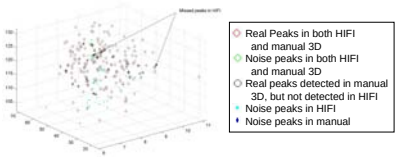
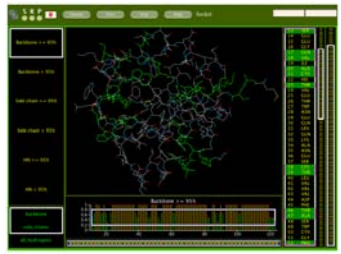


Figure showing results of HIFI NMR approach for CBCA(CO)NH experiment using 4 adaptively selected planes.

	HIFI	3-D (Manually peak picked)
Number of reported peaks:	215	223
Number of peaks assigned in final structure (Real peaks):	176	178
Number of peaks not assigned in final structure (Noise peaks):	39	45
Total data collection time	2 hrs	22 hrs

PISTACHIO

PISTACHIO is a probabilistic approach to assignment of NMR backbone and sidechain resonance data. Input in the form of XEasy file format for any combination of the following experiments is supported : HSQC, HNCO, CBCA(CO)NH, HNCACB, HN(CO)CACB, HNCA, HN(CO)CA, HN(CA)CO, HBHA(CO)NH, C(CO)NH, HN(CO)(CA)CB, HN(CA)CB, H(CCO)NH, N15-TOCSY, and HCCH-TOCSY. Output in two formats is supported. The XEasy Prot file format is supported
for those users that prefer the standalone mode. NMR-STAR 3.0 is supported as the native output format. The standalone interface is available on the web at http://bija.nmrfam.wisc.edu/PISTACHIO. The LIMS integrated input through SESAME will be available through the same webpage soon. Visualization and inspection output modes such as depicted in the screenshot at the top of the next column will be available soon. Some assignment results using PISTACHIO are shown in the following table, with CESG proteins highlighted in red.



Assignment visualization module is designed to enable visualization of the assignment at any stage of the process. In the figure above, the assignments with probability less than 95% are highlighted in green. Controls allow examination of assignments from a number of viewpoints and user-selected subsets.

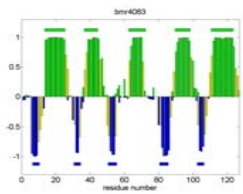
75 protein designator	# of residues correct/ # assignable residues	work-station time (hour)	P > .95 (Backbone)	percent Correct #	workstation time-sidechain (hour)	percent correct (side chain)
At2g24940	106/106 (3*)	0.2	100%	100%	1	95%
At1g77540	96/97 (6)	0.2	100%	99%	0.1	95%
At2g23090	82/84(2)	0.1	98%	97%	##	##
CalC	92/94(4)	0.1	97%	97%	0.1	85%
At9g22580	99/108 (3)	4	95%	92%	1	86%
CESO73	108/118 (2)	4	95%	92%	##	##
At3g17210	96/107 (5)	5	95%	90%	1	90%
At3g51030	108/120 (4)	4	95%	90%	##	##
At5g01610	95/115(5)	6	90%	83%	1	80%
At3g16450*	232/291 (8)	5	85%*	80%	2	75%
At1g23750	122/152 (5)	6	85%	80%	##	##
Acyl Carrier protein	80/82 (0)	0.7	98%	98%	##	##
BMRB 5106	61/68 (2)	1	95%	90%	##	##
YgpX	71/89 (2)	4	80%	80%	1	75%
P-Gamma	62/77 (10)	4	85%	80%	##	##

Correct assignment in comparison with manual assignment
No sidechain data available for that protein
(*) Number of Prolines in the sequence subtracted from total count.

PECAN

'PECAN' (Protein Energetic Conformational Analysis from NMR chemical shifts) optimizes a combination of information sources including residue-specific statistical energy function to yield energetic descriptions most favorable to predicting secondary structure. PECAN can achieve 90% accuracy in well-structured regions in a database of over 50,000 residues. Using the energy model, PECAN constructs a probabilistic description wherein each residue is assigned a probability of belonging to a designated state (e.g. Helix, sheet, etc.). The model has the advantage of identifying "intermediate regions" where a strict geometric assignment of state may depend on threshold.

The probabilistic secondary structure output is shown below. The yellow bars indicate regions we have called "intermediate" above.



PECAN is available at http://bija.nmrfam.wisc.edu/PECAN

INTEGRATION WITH SESAME MODULES: CALIPH AND CAMEL

Caliph Module for data validation and deposition.

Caliph is being created to collect, in pre-validated 'deposition-ready' form, the detailed specific data needed for PDB and BMRB depositions. Most of the required information will have been accumulated in the Sesame database. Caliph will prompt collection of any remaining information required for depositions. NMR data will be used to construct an NMR-STAR file that will be submitted through the new ADIT-NMR joint interface to PDB and BMRB (to be released by the middle of 2005); these depositions will include peak lists, restraints, and raw (time-domain) data sets. This tool will validate results prior to deposition to ensure that they meet CESG and community standards. For NMR structures, Caliph will launch evaluations by the WHAT-IF, PROCHEK-NMR, and the PDB and BMRB validation software suites. The validation approaches for X-ray structures described in will also be implemented under Sesame.

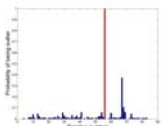
Camel will be expanded to support fast data collection methods, primarily the HiFi NMR approach and relevant post-processing part of HiFi NMR. The front end to the PISTACHIO and PECAN software tools will be a View within the Camel Module. Any outside users will be able to upload their chemical shift data (peak lists) and obtain the calculated assignments and secondary shift predictions, along with measures of their probable accuracy in standard NMR-STAR files or XEasy formatted files.

TOOLS: MANI-LACS, MANI-ANSA, PEANUT

MANI-LACS: Linear Analysis of Chemical Shifts for reference correction and outlier detection. MANI-LACS can detect potential outliers using linear analysis of chemical shifts. An outlier may be the result of miss assignment of chemical shifts or it may be a sign of relevant information that should be examined further. MANI-LACS uses the probabilistic computational engine and reports probabilities for the presence of outliers.
MANI-LACS is available at http://www.bija.nmrfam.wisc.edu/MANI-LACS/ .

MANI-ANSA: Automated NMR Spectral Alignment for registering multiple spectra. MANI-ANSA is used to align reported peaks from multiple spectra. Proper registration of peaks is an important step of any automatic assignment procedure. It is an easy task for humans to align spectra but difficult for computers when different data sets may miss different data points. MANI-ANSA accepts multiple peak-lists as input and reports a reference peak list consistent with all data.

PEANUT: Peak Enhancement using Adaptive Non-Uniform Tempering. PEANUT is used for automatic peak recovery using a robust algorithm that can capture both weak and strong signals. The fast peak extraction tool (PEANUT, Peak Enhancement using Adaptive Non-Uniform Tempering) developed by CESG-NMRFAM, (PEANUT) builds on ideas used in a new class of image-processing algorithm developed recently.



Bar plot showing the probability of a residue being an outlier. The horizontal axis is the amino acid sequence and the vertical axis is probability. The red bar is the outlier residue HIS56 in the BMRB entry 4998.

FUTURE PLANS

Expansion of the capabilities of the MANI system in a number of directions is currently planned.

To support the production capabilities of the system, the existing modules are being integrated into a web available environment that will be open to the community. Close collaboration with production environments will be used to fine-tune the system.

New extensions will be available in the near future that provide new, more efficient, methods for determining initial structure models based on the probabilistic models. Important data collection time savings will also be added – For example, a 4D to 3D NOESY method that would need little more time than 3D data collection.

Longer term, the probabilistic model and model-based constructions will be extended to cover the entire NMR pipeline to give on-demand updates of structure based on current available data.

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative P50 GM 64598 and P41RR02301