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 overall umbrella of a software system of steps leading from NMR data collection to structure determination and data deposition. The platform will include (fast) data collection, 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 and Camel) the Sesame web-based software system of steps leading from deposition to deposition. The figure illustrates the basic idea of internal consistency checks where results are iterated to achieve constant probability values across iterations. Additional components will be included as 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 be extracted and accumulated in the Sesame database. Caliph will prompt collection of any remaining information required for deposition. Deposition will be used to construct an NMR-STAR that will be submitted through the new ADP-STAR portal to PDB and BMRB (to be released by the middle of 2005); these deposits will include peak lists, restraints, and new time-domain data. This tool will validate results prior to deposition to ensure that they meet CESG and community standards. For NMR data the Sesame software suite will be enhanced with software modules to handle all new steps in the data processing and structure determination continuum as described in this paper. Sesame 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 PEANUT 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.

MANI MODEL

MANI is a model for High-Throughput Spectra Data Analysis. The model is based on NMR-STAR 3.0 data model. The overall model for the MANI system is simple. A number of components (HiFi-NMR, PISTACHIO, etc) use the probabilistic computational engine to give probabilities for solution configuration. Each component communicates with others for 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 approach 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 overall model for the MANI system is simple. A number of components (HiFi-NMR, PISTACHIO, etc) use the probabilistic computational engine to give probabilities for solution configuration. Each component communicates with others for 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 approach 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 repeated until no significant changes in probabilities are observed.

HI-FI NMR

Field 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 ‘on-the-fly’ analysis of results using NMR data, while maintaining, or potentially improving, the resolution. Portions of this development have synergistic impact on approaches for automated assignments and structure determination.

Our approach is based on the experimental approach of Kopecky and Freeman. We collect data from two orthogonal planes of 3D data set, and add an additional plane adaptively to gain information and detect which planes that would be the “most informative”. Frames are added iteratively until the position where all planes have been discharged. All 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 approach will recover data from a pair of 3D experiments, for example when 4D NOESY data are acquired.

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 peaks. It is a way to tune for humans to align spectra but difficult for computers when different data sets may miss different data points. MANI-LACS uses the probabilistic computational engine and reports probabilities for the presence of outliers.

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 for which a strict geometric assignment of state may depend on threshold.

FUTURE PLANS

Expansion of the capabilities of the MANI system in a number of directions is currently planned. 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 secondary structure models on the probabilistic models. Important data collection time savings will also be added. For example, a 4D to 3D NOESY can be done in less than half the time of the 4D.

Long-term, the probabilistic model and model-based constructions will be extended to cover the entire NMR pipeline to give near real-time updates of structures based on current available data.