Revised

An Integrated Platform for Automated Analysis of Protein NMR Structures

Yuanpeng Janet Huang[a], Hunter Moseley[a] , Michael C. Baran[a],

Cheryl Arrowsmith[b], Robert Powers[c], Roberto Tejero[a],

Thomas Szyperski[d], and Gaetano T. Montelione[a,e,*]

[a]Center for Advanced Biotechnology and Medicine,
Department of Molecular Biology and Biochemistry,
and Northeast Structural Genomics Consortium,
Rutgers University, Piscataway, NJ 08854

[b]Ontario Cancer Institute, Department of Medical Biophysics,
and Northeast Structural Genomics Consortium,
University of Toronto, Toronto, Ontario, Canada M5G 2M9.

[c]Department of Chemistry, University, Lincoln, NB 68588

[d]Department of Chemistry, and Northeast Structural Genomics Consortium,
University at Buffalo, The State University at New York, Buffalo, NY 14260

[e]Department of Biochemistry and Molecular Biology,
Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA

*Correspondence should be addressed to:
Prof. Gaetano T. Montelione
CABM-Rutgers University
679 Hoes Lane, Piscataway, NJ 08854
Ph: 732-235-5321; Fax: 732-235-5633
e-mail: guy@cabm.rutgers.edu

Running title: An Integrated Protein NMR Analysis Platform

Abstract

Recent developments provide automated analysis of NMR assignments and 3D structures of proteins. These approaches are generally applicable to proteins ranging from about 50 to 150 amino acids. In this chapter, we summarize progress by the Northeast Structure Genomics Consortium in standardizing the NMR data collection process for protein structure determination, and in building an integrated platform for automated protein NMR structure analysis. Our integrated platform includes the following principal steps: (i) standarized NMR data collection, (ii) standardized data processing (including spectral referencing and Fourier transformation), (iii) automated peak picking and peak list editing, (iv) automated analysis of resonance assignments, (v) automated analysis of NOESY data together with 3D structure determination, and (vi) methods for protein structure validation. In particular, the software AutoStructure for automated NOESY data analysis is described in this Chapter, together with a discussion of practical considerations for its use in a high throughput structure production effort. The critical area of data quality assessment has evolved significantly over the last few years, and involves evaluation of both intermediate and final peak lists, resonance assignments, and structural information derived from the NMR data. Methods for quality control of each of the major automated analysis steps in our platform are also discussed. Despite significant remaining challenges, when good quality data are available, automated analysis of protein NMR assignment and structures with this platform is both fast and reliable.

Introduction

With the advent of multidimensional and triple-resonance strategies for determining resonance assignments and 3D structures, it has become increasingly clear that protein NMR spectra have the quality and information content to allow largely automated and standardized analyses of assignments and structures for small proteins. This has been realized over the last few years in the development of automated methods for many of the steps in production NMR protein structure analysis. These advances are significant demonstrations of NMR as a powerful and accessible tool for biophysical chemistry, drug design, and functional genomics. In this article, we summarize our efforts in standardizing the NMR data collection process, building an integrated platform for automated NMR structure analysis, and demonstrating its impact for the NorthEast Structural Genomics consortium (NESG).

Overview of the Automated Protein Structure Analysis Process

The principal steps of automated NMR protein structure analysis are outlined in Figure 1. These include (i) Standardized Data Collection and Organization, (ii) Processing (including spectral referencing and Fourier transformation), (iii) Peak Picking and Peak List Editing, (iv) Resonance Assignment, and (v) Structure Determination (including analysis of conformation constraints, NOESY assignment, RDC data analysis and 3D structure generation). In building an automated data analysis platform, the input and output of each of these steps must be organized in a self-consistent way, ideally using a relational database (Baran, et al., 2002, Zolnai, et al., 2003). A key issue for automated analysis is validation of completeness, quality, and consistency of data generated in each of these principal steps. Recent efforts have focused on Peak List Validation, Resonance Assignment Validation, and Structure Validation. A critical issue for automation is data quality. These validation steps, and estimates in uncertainties in the derived information, are critical both for defining a robust and reliable automation process, and for interpreting the resulting resonance assignments and 3D structures.

Standardized Data Collection and Organization

The Organizational Challenge

The process of NMR-based protein structure analysis is challenged by requirements for properly executing, processing, and analyzing many separate NMR experiments. Unlike biomolecular crystallography, which generally involves a single type of data collection experiment, an NMR protein structure determination may require proper collection and analysis of 10 – 20 individual 2D, 3D, and 4D NMR spectra. These data must be highly self-consistent, as the input to the structure calculations is a composite generated from across these many data sets.

Standardized Data Collection

The challenges of organization for automated data analysis begin with data collection. As protein structure analysis relies on data from many different NMR experiments, it is critical that this data be self-consistent and fairly complete. Self-consistency can be particularly problematic when mixing data collected on different NMR spectrometers and/or using different samples of the protein under investigation. Efforts must be made to minimize spectrum-to-spectrum variability. In our laboratories, we generally collect all the data needed for a protein structure analysis back-to-back on the same sample and usually with the same NMR instrument. However, it is not always possible to collect in this manner, and even this strategy does not ensure consistency across spectra since sample heating effects can depend on decoupler duty cycles, which are different across NMR experiments. Fortunately, the latest generation NMR probes, and particularly cryogenic probes, exhibit less sample heating from decoupling than previous generation probes.

Another critical organizational issue for automated data analysis is the use of a standardized set of NMR pulse sequences for data collection. Each implementation of a sophisticated NMR experiment involves data collection and processing parameters that are unique to that implementation. It is very difficult to construct an analysis platform that is completely flexible with respect to all possible permutations. Well-defined sets of NMR data collection strategies creates the basis for a robust analysis platform, providing consistent types of input data and guiding users to a better understanding of which NMR experiments are essential, optional (but useful), or superfluous. In general, different protein classes (e.g. small ^{15}N, ^{13}C-enriched proteins vs. larger perdeuterated ^{15}N,^{13}C-proteins) require different data collection strategies; but a standardized set of experiments for each of these general classes can be defined. Within our "standard data collection sets", some experiments are defined as "required" while others are labeled "optional." Typically, "optional" experiments are only carried out when the quality evaluation of the "required" set deems it necessary.

It is also valuable to define the adjustable (sample dependent) and fixed parameters of data collection and processing for each NMR experiment in each "standard set." For example, in generating triple-resonance spectra for automated analysis of resonance assignments, we constrain the digital resolution in "matching dimensions of complementary spectra" (e.g. the ^{13}C dimensions of HNCA and HNcoCA spectra) to be identical, in order to maximize accuracy in matching intraresidue and sequential crosspeaks between these spectra. In the activities of our structural genomics project (www.nesg.org), one of the most critical innovations providing high-efficiency NMR structure generation has been the establishment of standardized data collection strategies and carefully-considered default data collection and processing parameters.

The development of a package for employing reduced-dimensionality (RD) NMR spectroscopy (Szyperski, et al., 1993) for complete protein resonance assignment (Szyperski, et al., 2002) exemplifies this point. The 'RDpack' (Y. Xia, D. K. Sukumaran, C. Arrowsmith, T. Szyperski, in preparation) comprises pulse sequences, parameter sets, scripts and macros for efficient *de novo* implementation of RD NMR experiments as well as rapid adjustment of parameter sets when using VARIAN INOVA spectrometers (Fig. 2). The RDpack is freely available for academic users and contains 11 experiments. 3D $\underline{\mathrm{H}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha\beta}$coNH, 3D $\underline{\mathrm{HACA}}$coNH and 3D $\underline{\mathrm{HC}}$ccoNH-TOCSY sequentially correlate proton and carbon shifts of residue *i*-1 with the amide proton and nitrogen shifts of residue *i*, 3D $\underline{\mathrm{H}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha\beta}$NH, 3D HN$\underline{\mathrm{CAHA}}$ and 3D $\underline{\mathrm{H}}^{\alpha\beta}\underline{\mathrm{C}}^{\alpha\beta}$coHA provide complementary intraresidue connectivities, and 3D HN<$\underline{\mathrm{CO}}$,$\underline{\mathrm{CA}}$>

affords both sequential and intraresidue connectivities. Aliphatic side chains are assigned by use of 3D HCCH COSY and TOCSY, while aromatic spin system identification relies on 2D HBCBcgcdHD and 2D ^{1}H-TOCSY HCH COSY. All parameter sets offer flags to conveniently select (i) central peak acquisition from ^{13}C steady state magnetization (Szyperski, et al., 1996), (ii) transverse relaxation optimized spectroscopy (TROSY) (Pervushin, et al., 1997) type data acquisition, and (iii) ^{2}H-decoupling.

Local Data Organization and Archiving

Biomolecular NMR research groups require efficient and simple access to archival NMR data, both for routine storage purposes and for the development and testing of novel computational methods for data analysis. Common methods of archiving raw NMR data [usually in the form of time domain free-induction decay (FID) data] in use in most biomolecular NMR laboratories are often inefficient, out-dated, and error-prone, leading to frequent loss of valuable data that are both hard and expensive to obtain. The growing demands on data organization and formatting in submitting NMR data and structures to public databases like the BMRB (Seavey, et al., 1991) and the PDB (Berman, et al., 2000) also require simple methods of harvesting NMR data and moving this information from the NMR laboratory into appropriate archival formats. This is particularly challenging for the several pilot projects in structural proteomics (Chance, et al., 2002, Gong, et al., 2003, Heinemann, et al., 2000, Kennedy, et al., 2002, Terwilliger, 2000, Yokoyama, et al., 2000) which are being encouraged to submit into the public domain many more data items than have been traditionally expected from a conventional structural biology project. The goal of a standardized archive is to not only increase laboratory productivity through organization, but also to support future NMR methods development by organizing laboratory data into a format which can easily be retrieved, reproduced and shared across the community. If properly organized and archived, these data will be invaluable to the NMR community in efforts to develop new data collection and analysis technologies.

Examples of recently described NMR Laboratory Information Management System (LIMS) solutions are the Sesame (Zolnai, et al., 2003) and SPINS (Baran, et al., 2002) databases. SPINS (Standardized ProteIn NMR Storage) (Baran, et al., 2002) is an object-oriented relational database and data model that provides facilities for high-volume NMR data archival, data organization, and dissemination of raw NMR FID data to the public domain by automatic preparation of the header files needed for simple submission to the BMRB (Seavey, et al., 1991).

NMR Spectral Processing

Several NMR spectral processing issues need to be carefully considered for successful automated data analysis. Particularly important are accurate and precise chemical shift referencing in the direct and indirect dimensions using IUPAC-defined referencing methods (Wishart, et al., 1995) with dimethylsilapentane-5-sulfonic acid (DSS) as the reference compound. Accurate ^{13}C, ^{15}N, and ^{1}H referencing is essential for ensuring the development of an accurate database of chemical shift values (Zhang, et al., 2003). Proper chemical shift referencing for aliphatic ^{13}C and ^{1}H resonances is also critical for accurate amino acid typing

(Grzesiek and Bax, 1993, Moseley, et al., 2001, Zimmerman, et al., 1997) and secondary structure analysis (Wishart and Sykes, 1994), generating information that is used in most automated assignment and structure programs. In our laboratories, we externally calibrate the synthesizer offsets on each NMR spectrometer with a sample of 1 mM DSS in ^2H_2O at neutral pH and at multiple temperatures, and then use these calibrations to define the corresponding chemical shift value of the carrier offset in each dimension of each NMR spectrum (Monleon, et al., 2002).

As with NMR data collection, similar amounts of zero-filling and/or linear prediction, and similar window functions should be applied to matching dimensions across spectra to provide comparable final digital resolutions (Montelione, et al., 1999, Moseley, et al., 2001). This allows the use of the tightest possible “match tolerances” in later steps of automated analysis. We typically zero-fill the direct H^N dimension to 1024 complex points and 2-fold linear predict and zero-fill each indirect dimension to 256 or 512 complex points. Even though this copious increase in digital resolution goes beyond usual theoretical recommendations, such processing can aid peak picking software that does not interpolate peak centers well. The use of linear prediction also suppresses severe Fourier truncation artifacts (e.g., sinc wiggles) and reduces line broadening effects of window functions (Koehl, 1999). This can have a significant impact in crowded regions of a spectrum. Linear prediction generally produces cleaner spectra and better shaped peaks, thus improving the performance of the peak picking algorithms, providing higher quality peak lists, and ultimately improving the performance of later automated analysis steps (Moseley, et al., 2001). It is also critical to apply ridge-suppression and baseline correction in each spectral dimension to improve their quality, which can be very important for later restrictive peak picking steps (Monleon, et al., 2002).

Several high quality NMR processing programs have been developed over the last several years, including Felix (Molecular Simulations, Inc., San Diego, USA), NMRPipe (Delaglio, et al., 1995), PROSA (Guntert, et al., 1992), VNMR (Varian, Inc., Palo Alto, CA, U.S.A), and XWinNMR (Bruker Analytik GmbH, Karlsruhe, Germany). NMR data processing requires expert knowledge of many technical concepts and terms, presenting barriers to scientists not familiar with the deeper details of NMR spectroscopy. However, many of the parameters associated with the referencing and processing of NMR data, though specific to the pulse sequence program and particular spectrometer used to record the data, are relatively sample independent. Given the constraints of the data collection process as defined by the NMR pulse sequence, only a few adjustable parameters need to be considered by a user, and most of these can be set to usable default values based on general laboratory experience. Accordingly, there are several steps in the analysis of NMR data that may be viewed as routine tasks, but often demand non trivial amounts of time, knowledge of NMR theory, and familiarity with technical features of the specific data collection methods and/or processing software.

To address these data organization issues, we have developed AutoProc (Monleon, et al., 2002), a data dictionary together with a set of software tools designed to allow a non-expert in NMR spectroscopy to accurately reference multidimensional NMR spectra, generate and run appropriate conversion scripts, and process NMR data using the software package NMRPipe (Delaglio, et al., 1995). AutoProc takes as input FID files along with libraries of spectrometer and pulse-sequence specific description (table) files. It converts the data into a processing format, references the data in the direct and indirect dimensions using spectrometer-specific calibrations, and creates processing scripts suitable for running NMRPipe. It is straightforward to modify AutoProc to work with other script-based processing software like Felix (Molecular Simulations, Inc., San Diego, U.S.A.) or PROSA (Guntert, et al., 1992).

Peak Picking

Peak picking represents one of the crucial steps of NMR data analysis that has resisted successful automation for the purpose of automated resonance assignment and structure determination. This is due largely to cross peak overlap and artifacts associated with large peaks, especially solvent diagonal peaks. Multidimensional NMR spectra often exhibit artifacts of baseline distortions, intense solvent lines, ridges, and/or sinc wiggles. These problems are sometimes exacerbated by different processing methods that can dramatically affect line shape, intensity, and resolution of peaks as well as the severity of spectral artifacts.

Most automated peak pickers (Eccles, et al., 1991, Garrett, et al., 1991, Goddard and Kneller, 2000, Herrmann, et al., 2002, Koradi, et al., 1998, Orekhov, et al., 2001) rely on properties of an individual peak along with a model of the noise generated in the spectrum to determine whether a peak is valid or not; though, one approach has looked at comparative properties of doublets (Andrec and Prestegard, 1998). Many programs perform restricted peak picking or filtered peak picking which is a form of peak list editing where one peak list is filtered against another in comparable dimensions (Goddard and Kneller, 2000, Monleon, et al., 2002, Zimmerman, et al., 1997). ATNOS (Herrmann, et al., 2002) is software for automated NOESY peak picking. It uses NOESY symmetry relationships along with restrictive peak picking against an assigned resonance list to guide the automated peak picking while using a ridge detection method to minimize peak picking along ridges. ANTOS has been used together with NOESY assignment and the structure determination software CANDID(Herrmann, et al., 2002) and DYANA (Guntert, et al., 1997) to iteratively identify and assign NOESY cross peaks.

In our laboratories, peak picking is usually done using the restrictive peak picking and peak editing facilities in the program Sparky (Goddard and Kneller, 2000) or XEasy. Additional software, AutoPeak (Monleon, et al., 2002), uses peak lists generated from manually peak picked 2D ^{15}N-^{1}H HSQC, ^{13}C-^{1}H HSQC spectra as frequency-filters across raw peak lists from 3D spectra. For the peaks which pass these filters, Sparky reports linewidth, root mean square fits to Lorenzian line shape, and peak intensity data can be used to further filter artifactual entries in the initial peak list table. Despite the sophistication of these automatic peak picking and editing methods, it is generally necessary to follow up with further editing (inclusion and exclusion) of peak lists by manual inspection of the spectra.

This manual editing is guided by a data completeness quality report generated from initial analysis of data [i.e., the examine_spin_systems.pl (ESS) report from the AutoPeak software]. For an experienced spectroscopist, peak list editing for a typical set of NMR spectra used for backbone resonance assignments is completed in about 1 day, and can be streamlined by doing some of the peak list editing while some data collection is still in progress (Moseley, et al., 2001).

Interspectral Registration and Quality Assessment of Peak Lists

Quality assessment of input peak lists for further steps in the automated NMR analysis is crucial for the success of automation. We use several quality assessments of peak lists when judging if the peak lists are good enough for the later steps of automation. These include (i) peak list registration, (ii) the examine_expected_peaks.pl (EEP) report, and (iii) the examine_spin_systems.pl (ESS) reports of the AutoPeak software suite (Moseley, et al., 2001). The first quality assessment is the ability to register peak lists to each other in their comparable dimensions. Registration is an often overlooked step that is absolutely required for good performance in automated resonance assignment and NOESY assignment steps. In our current platform, a distance matrix approach [calculate_registration (Monleon, et al., 2002)] is used to register peak lists from different spectra using resonance frequencies common to pairs of spectra. This approach has the added benefit of providing standard deviations of matching frequencies that can be used to derive appropriate tolerances for later steps in the automated NMR analysis. These standard deviations, along with a count of the peaks that contributed to their calculation, provide scores that can used to assess the quality of the peak lists they come from. Interspectral registration data, and other (EEP and ESS) spin system quality reports provided by the AutoPeak software suite, are used to determine if a set of peak lists is of good enough quality for automated NMR analysis, and to identify problematic or incomplete peak lists.

AutoAssign - Automated Analysis of Backbone Resonance Assignments

Significant progress has been made recently in automated analysis of resonance assignments, particularly using triple-resonance NMR data. Several laboratories are developing programs that automate either backbone or complete resonance assignments (reviewed in refs (Baran, et al., 2004, Moseley and Montelione, 1999, Zimmerman and Montelione, 1995)). Most automated programs use the same general analysis scheme which originates from the classical strategy developed by Wüthrich and co-workers (Billeter, et al., 1982, Wagner and Wuthrich, 1982, Wuthrich, 1986).

Most commonly used algorithms for automated analysis of resonance assignments include the following steps (Moseley and Montelione, 1999): i) register peak lists in comparable dimensions (registering/aligning); ii) group resonances into spin systems (grouping); iii) identify amino acid type of spin systems (typing); iv) find and link sequential spin systems into segments (linking); and v) map spin system segments onto the primary sequence (mapping). Different automation programs implement each step with varying degrees of success; however, overall robustness is often dictated by the performance of the weakest step. The different automated resonance assignment programs are typically categorized by the methods they use in the mapping step. These methods include simulated annealing/Monte Carlo algorithms (Buchler, et al., 1997, Leutner, et al., 1998, Lukin, et al., 1997), genetic algorithms (Bartels, et al., 1996, Bartels, et al., 1997), exhaustive search algorithms (Andrec and Levy, 2002, Atreya, et al., 2000, Coggins and Zhou, 2003, Guntert, et al., 2000), heuristic comparison to predicted chemical shifts derived from homologous proteins (Gronwald, et al., 1998), and heuristic best-first algorithms (Hyberts and Wagner, 2003, Li and Sanctuary, 1997, Zimmerman, et al., 1994, Zimmerman, et al., 1997).

We develop and use the automated backbone resonance assignment program AutoAssign (Moseley, et al., 2001, Zimmerman, et al., 1997). AutoAssign is a constraint-based expert system (heuristic best first mapping algorithm) designed to determine backbone H^N, H^α, $^{13}C'$, $^{13}C^\alpha$, ^{15}N, and $^{13}C^\beta$ resonance assignments from peak lists derived from a set of triple resonance spectra with common H^N-^{15}N resonance correlations. The original implementation of AutoAssign was written in LISP with a Tcl/Tk-based graphical user interface (GUI) (Zimmerman, et al., 1997). The current version of AutoAssign is written in C++ with a Java-based GUI (Moseley, et al., 2001). The program can handle data obtained on uniformly ^{15}N-^{13}C doubly-labeled; uniformly or partially-deuterated, ^2H-^{15}N-^{13}C triply-labeled; and selectively methyl-protonated, uniformly or partially-deuterated, ^2H-^{15}N-^{13}C triply-labeled protein samples.

AutoAssign requires five different types of peak lists but may use up to nine different types of peak lists representing data obtained from a variety of triple resonance experiments and a ^{15}N-H^{N} HSQC spectrum. These nine types of peak lists represent information from the following nine types of experiments: HSQC*, HNCO, HNCACB*, HNcoCACB*, HNCA*, HNcoCA*, HNcaCO, HNcaHA, and HNcocaHA. Those peak lists marked by an asterisk are required by the program; however, using all nine types of data obtains the best performance (Moseley, et al., 2001).

Key components of the processing, peaking picking, and automated assignment software, AutoProc (Monleon, et al., 2002), NMRPipe (Delaglio, et al., 1995), AutoPeak (Monleon, et al., 2002), Sparky (Goddard and Kneller, 2000) and AutoAssign (Moseley, et al., 2001, Zimmerman, et al., 1997), have been integrated together to provide a platform for rapid analysis of resonance assignments from triple resonance data. This prototype "integrated backbone resonance assignment platform" (Monleon, et al., 2002) was applied to data collected from the small protein bovine pancreatic trypsin inhibitor (BPTI) using a first-generation high-sensitivity triple resonance NMR cryoprobe. Seven NMR spectra were recorded in each of two sessions on a 500 MHz NMR system, requiring 36.6 hrs and 5.5 hrs of data collection time, respectively. Fourier transforms were carried out using a cluster of Linux-based computers, and complete analysis of the seven spectra collected in each session was carried out in about 2 hrs. Nearly complete backbone resonance assignments and secondary structures (based on chemical shift data) for a 58-residue protein were determined in less than 30 hours, including data collection, processing *and* analysis time. In this optimum case of this small well-behaved protein providing excellent spectra, extensive backbone resonance assignments could also be obtained using less than 6 hours of data collection and processing time. These results demonstrate the feasibility of high throughput triple resonance NMR for determining resonance assignments and secondary structures of small proteins.

Automated Analysis of Sidechain Resonance Assignments

While several approaches have been found to provide robust automation of backbone resonance assignments, a robust approach to automated sidechain assignments is not yet generally available. The program GARANT (Bartels, et al., 1996) supports automated backbone and sidechain assignments. Recently, a combined approach of using GARANT and AUTOPSY (Koradi, et al., 1998) together demonstrates promising results in automating both peak picking and resonance assignments, including many sidechain aromatic ^{1}H resonance assignments (Malmodin, et al., 2003).

The principal challenge in automated analysis of sidechain resonances is incompleteness in experimental peak lists generally available for this task. Most published efforts in automating sidechain resonance assignments (Bartels, et al., 1997, Coggins and Zhou, 2003, Hyberts and Wagner, 2003) focus on HCCcoNH-TOCSY (Grzesiek, et al., 1993 , Logan, et al., 1992, Montelione, et al., 1992), and use statistical comparisons to ^{13}C sidechain resonance values of amino acid residues to assign the chemical shifts. These H^{N}-detected ^{13}C-^{13}C TOCSY spectra are simple to interpret, but are often quite incomplete. Generally, no single spectrum has all side chain carbon resonances due to differences in

TOCSY transfer efficiencies for short chain and long chain amino acids, although more complete data can sometimes be obtained by co-adding spectra recorded with different isotropic mixing times (Celda and Montelione, 1993). While fairly complete HCCcoNH-TOCSY data can sometimes be obtained for proteins of < 10 KDa, and analyzed automatically with published methods, relaxation effects generally prevent the experiment from working well with larger proteins unless they are partially deuterated (Farmer and Venters, 1995, Gschwind, et al., 1998, Lin and Wagner, 1999). For these reasons, a robust approach for automated sidechain assignments should utlize HCCcoNH-TOCSY recorded with multiple mixing times, as well as other data such as HCCH-COSY (Bax, et al., 1990, Ikura, et al., 1990, Kay, et al., 1990) and/or HCCH-TOCSY (Bax, et al., 1990, Fesik, et al., 1990)

Resonance Assignment Validation Software

As with peak picking, quality assessment of resonance assignments is crucial for robustness in later steps of the automated NMR analysis. For this purpose, we have developed a set of computer utilities called the Assignment Validation Software (AVS) suite (Moseley, et al., 2004) for rigorously evaluating and validating a set of protein resonance assignments before submission to the BMRB and/or use in subsequent structure and/or functional analysis, without the need of a 3D structure. They serve the purpose of providing strict consistency checks for detecting possible errors and identifying 'suspicious' assignments that deserve closer scrutiny prior to NOESY spectral analysis and 3D structure generation.

AutoStructure - Automated Analysis of NOESY data

One of the principle goals of automated structure determination programs involves iterative analysis of multidimensional NOESY data. Several fully automated heuristic approaches for NOESY interpretation and structure calculation have been developed, including NOAH (Mumenthaler and Braun, 1995, Mumenthaler, et al., 1997), ARIA(Nilges, 1995, Nilges, et al., 1997), CANDID (Herrmann, et al., 2002), AutoStructure (Huang, et al., 2003), a self-consist constraint analysis method implemented in XPLOR (Kuszewski, et al., 2004) and other generally less developed programs (Adler, 2000, Grishaev and Llinas, 2002, Gronwald, et al., 2002). The NOAH, ARIA, and CANDID programs utilize an iterative *top-down* data interpretation approach, having the following steps in common: i) Ambiguous proton-proton interactions from unassigned NOESY cross peaks, together with unambiguously assigned proton-proton interactions are incorporated into structure calculations and generate a new set of model structures; ii) ambiguous proton-proton interactions are iteratively trimmed using the resulting model structures if they are far apart in the intermediate model structures. One key difference between NOAH and ARIA/CANDID is how ambiguous peaks are converted into distance constraints: NOAH creates an unambiguous constraint for each ambiguous proton-proton interactions while ARIA/CANDID uses an ambiguous constraint strategy (Nilges, 1995, Nilges, et al., 1997) which only generates one ambiguous distance constraint for each ambiguous peak.

AutoStructure (Huang, et al., 2003) uses a iterative *bottom-up topology-constrained approach* to analyze NOE peak lists and generate protein structures. AutoStructure first builds an initial fold based on intraresidue and sequential NOESY data, together with characteristic NOE patterns of secondary structures, including helical medium-range NOE interactions and interstrand β-sheet NOE interactions, and unique long-range packing NOE interactions based on chemical shift matching and symmetry considerations. Unassigned NOESY cross peaks are not used in structure calculations. Additional NOESY cross peaks are iteratively assigned using intermediate structures and the knowledge of high-order topology constraints of α-helix and β-sheet packing geometries. This protocol, in principal, resembles the methodology that an expert would utilize in manually solving a protein structure by NMR. The program AutoStructure has been combined with the structure generation programs DYANA(Guntert, et al., 1997) or XPLOR/CNS(Brunger, 1992, Brunger, et al., 1998).

The control-flow of AutoStructure

The first step of AutoStructure (Fig. 3) is to match the chemical shifts from the NOESY peak list with the chemical shifts from the resonance assignment table using a loose match tolerance Δ_1 (typical values are 0.05 ppm for ^{1}H and 0.5 ppm for ^{13}C or ^{15}N). Aliased peaks can be directly matched to unaliased chemical shifts; there is no need for manually unfolding aliased peaks or generating aliased chemical shifts. AutoStructure builds an ambiguous distance network (G_{ANOE}) from the chemical shift matching, in which nodes represent protons from resonance assignment table, and edges represent NOE cross peaks linking all possible matched proton pairs. The rest of the steps of AutoStructure involve building a heuristic subgraph (HG_{NOE}) from G_{ANOE}, which is as close to the true distance network (representing the true 3D structures) as possible.

In step 2, HG_{NOE} is initialized using all well-matched (within a tighter tolerance Δ_2) NOE-linked proton pairs that are connected by only two-, three-, or four- covalent bonds (Wuthrich, et al., 1983), or belong to one of the $H^{\alpha}H^{N}$(i,i+1), $H^{\beta}H^{N}$(i,i+1), or $H^{N}H^{N}$(i,i+1) sequential NOE connections, commonly observed in protein NOESY spectra.(Billeter, et al., 1982) These close proton pair connections are anticipated from the amino acid sequence of the protein. A similar approach of reliably finding identifiable intraresidue and sequential NOESY peaks is often used by experts in the process of manual analysis of NOESY data. At this step, AutoStructure also attempts to minimize site-specific chemical shift differences between resonance assignment table and the NOESY peak list, due to interspectral variations of temperature and sample conditions. If proton h_i is involved in at least three NOE interactions (degree of vertex $h_i \geq 3$), its resonance frequency $\delta(h_i)$ in the refined resonance assignment list R′ is updated with the median value derived from these linked NOE cross peaks. Match tolerances (Δ_1) for those protons with refined chemical shifts are set to a narrower tolerance and linking edges with large mismatches resulting from these protons with updated chemical shift values are removed from G_{ANOE}. This step simulates the expert analysis process of refining chemical shift values to be used in NOESY analysis from the frequencies of interpreted NOESY cross peaks.

After refining the resonance assignment table with intraresidue NOESY data, AutoStructure identifies helices and β-sheets, including inter-strand alignments, by discovering patterns of NMR data that characterize secondary structures. This part of the algorithm uses chemical shift index (CSI) values,(Wishart and Sykes, 1994) $^{3}J(H^{N}\text{-}H^{\alpha}_{.})$ scalar coupling data,(Wuthrich, 1986) and characteristic NOE contact patterns. These NOE contact patterns, characteristic of canonical secondary structures, are identified in G_{ANOE} and then added into the HG_{NOE} heuristic distance network using constraints implied by unique features of these secondary and tertiary structures already identified by the NMR data. At the same time, edges that represent linked proton pairs which are inconsistent with the geometries of identified secondary structures are removed from G_{ANOE}. In these ways, both local and long-range constraints indicated by the secondary structure topology are used to further build HG_{NOE} from G_{ANOE} prior to the actual structure generation process.

At the end of Step 2, AutoStructure identifies unique NOE connections (h1, h2, p) with frq(p) = 1 from G_{ANOE} and selectively adds into HG_{NOE} those that are supported by a large number of potential interresidue contacts in a contact map generated from the G_{ANOE} network that has been interpreted to this point. A well-matched NOE-linked proton pair (h1, h2, p) is identified as a unique connection if the number of possible proton-proton interactions linked to the peak is unique [frq(p) = 1]. At this point, symmetry features of multidimensional NOESY spectra are also considered in order to resolve ambiguities due to chemical shift degeneracy for peaks with frq(p) > 1. Well-matched symmetric NOE-linked proton pairs (h1, h2, p1) and (h2, h1, p2) ($|\delta i-\delta(hi)| < \Delta_{good}{}^{i}$, $|\delta_1(p1)-\delta_2(p2)| < \Delta_{sym}$, and $|\delta_1(p2)-\delta_2(p1)| < \Delta_{sym}$) are also identified as unique connections if, in the subgraph of G_{ANOE} which consists of only symmetric NOE-linked proton pairs, frq(p1) = frq(p2) = 1.

In Step 3, AutoStructure constructs protein model structures. The program generates distance constraints directly from HG_{NOE} by calibrating the peak's intensities assuming a simple two-spin approximation and binning them into upper-bound distance classes as described by Wüthrich and co-workers (Mumenthaler, et al., 1997, Wuthrich, 1986, Wuthrich, et al., 1983). Dihedral angle constraints are generated from local NOE and scalar coupling data using the conformational grid search program HYPER.(Tejero, et al., 1999) Hydrogen bond distance constraints are identified based on the observation of helix and β-sheet NOE contact patterns, together with analysis of amide hydrogen exchange data and 3D structures when available.(Wuthrich, 1986) Potential cis-peptide bonds [i.e. H^{α}-$H^{\alpha}(i, i+1) \in HG_{NOE,}$ and H^{α}-$H^{N}(i, i+1) \notin HG_{NOE}$ or H^{α}-$H^{\delta}(i, Pro(i+1)) \notin HG_{NOE}$] and disulfide bonds (i.e. H^{β}-$H^{\beta}(Cys(i), Cys(j)) \in HG_{NOE}$) are identified and reported to the user for expert validation. After validation, these special structural features are manually added into the constraint list. AutoStructure generates input constraint lists suitable for either XPLOR/CNS or DYANA for protein structure generation calculations. Structures are usually generated using a coarse-grain parallel calculation strategy on a Linux cluster, although the program can also be run on a single processor system, such as a Linux-based laptop computer.

In Step 4, a set of *N* model structures that best satisfy the resulting constraints is used to evaluate and refine the self-consistency of HG_{NOE}. First, distances (of the sum of inverse sixth powers of individual degenerate proton-proton distances) between all NOE-linked proton pairs of HG_{NOE} are calculated. Proton pairs with internuclear distances that violate the corresponding constraints by greater than $dvio_{min}$ in all of these *N* initial structures are removed from HG_{NOE} distance network. The resulting HG_{NOE} is then used to regenerate another set of 3D model structures, which are again used for self-consistency analysis. This process of identifying inconsistent constraints within HG_{NOE} by 3D structure generation and analysis of consistent violations is repeated until no more such inconsistent proton pair interactions remain in HG_{NOE}.

The resulting HG_{NOE} distance network and its corresponding model structures are considered to be self-consistent and are subsequently used as templates to refine and expand HG_{NOE}. First, AutoStructure analyzes the topology of the initial or intermediate structures, and trims G_{ANOE} down based on *topology constraints* implied by helical-packing and β-sheet packing geometries based on the “ridges into grooves model”,(Chothia, 1984, Chothia, et al., 1981, Cohen, et al., 1982, Janin and Chothia, 1980). Next, AutoStructure further expands HG_{NOE} by adding NOE-linked proton pairs from G_{ANOE} that are well supported by the intermediate 3D structures. During this process, HG_{NOE} is further refined by removing any NOE assignments to long-range interactions associated with “orphan contacts” that may have evolved in the structure evolution process. Step 3 and 4 are repeated several times (typically 9 times) to iteratively refine the resulting structures. During this process, AutoStructure continues to refine the resonance assignment table using the resulting self-consistent HG_{NOE}.

Description of input data for AutoStructure

AutoStructure uses the following input data: i) protein amino acid sequence and a list of resonance assignments (set R); ii) a list of the multidimensional (i.e. 2D, 3D, or 4D) NOESY cross peak frequencies (which may be aliased) and intensities (set NOE); iii) a list of scalar coupling constant data (optional); iv) a list of slow amide ^{1}H exchange data (optional); and v) other manually analyzed constraints when available, such as residual-dipolar-coupling (RDC) (Tjandra and Bax, 1997), disulfide-bond, and dihedral-angle (Cornilescu, et al., 1999) constraint data. NOESY peak lists are generated using third-party automatic spectrum peak-picking programs, usually followed by some manual editing. Dimeric proteins can also be analyzed when interchain NOESY cross peak data are available from X-filtered NOESY experiments (Clore, et al., 1994), as demonstrated for coil-coil helix dimers.(Greenfield, et al., 2001, Greenfield, et al., 2003)

Quality control issues of input data for Autostructure

1. Requirements for resonance assignment table

AutoStructure uses a chemical shift index method(Wishart and Sykes, 1994) for secondary structure analysis and therefore requires accurate chemical shift referencing for C^α, C^β and H^α resonances. This chemical shift index method relies on the use of the recommended IUPAC chemical shift referencing method with DSS as the reference compound. High quality AutoStructure calculations require the input resonance assignment table to be more than 85% complete. For each aromatic residue, at least one aromatic side chain proton should be assigned in order for AutoStructure to define its ring packing.

2. Requirements for NOE peak lists

Peak lists do not have to be perfect. AutoStructure can handle the presence of artifactual peaks and incompleteness; however, inaccurate or imprecise peak picking can considerably limit the performance of the program. Intense solvent lines, ridges and/or sinc wiggles should be manually inspected and remove from the peak lists. Many NOE peaks may overlap with solvent lines and become hard to peak pick. However, collecting 3D ^{13}C-NOESY in D_20 can minimize such problem. AutoStructure can handle aliased/folded peaks. . High quality AutoStructure calculations require the input peak list (set NOE) to contain at least 90% real cross peaks.

3. Requirements for matching the NOE peak lists and resonance assignments

AutoStructure calculates an M-score which estimates the percent of predicted conformation-independent two- and three-bond connected NOE-linked proton pairs that are missing from the NOE peak lists. Four factors can contribute to high M scores: i) misalignment between chemical shifts from NOE peak lists and the resonance assignment table; ii) significant differences in the digital resolutions between chemical shifts from NOE peak lists and the resonance assignment table; iii) poor quality of NOE peak lists; iv) incorrect resonance assignments. A high M score (i.e. > 25%) suggests that at least one of the input data sets (R and/or NOE) are of inadequate quality and need to be improved. Those predicted two- and three-bond connected NOE-linked proton pairs missing from the NOE peak lists are reported to aid the user in improving the corresponding chemical shift assignments, and/or identifing the expected NOESY cross peaks in the corresponding NOESY spectrum.

AutoStructure requires that all NOESY spectra be accurately referenced relative to the values of chemical shifts reported in the resonance assignment table. For each frequency dimension, the software computes the overall average chemical shift match difference from these predicted NOE-linked proton pairs. Consistent spectral referencing is achieved using these differences as global reference correction factors for the target spectrum, providing a tighter match between NOE peak lists and resonance assignment table, and allowing the use of smaller matching tolerances for further NOESY interpretation.

Using AutoStructure

AutoStructure is implemented in a combination of C/C++ programs, Perl programs, and shell scripts. It can be run in batch model or using the graphical user interfact (GUI). The AutoStructure distribution on Linux platform is freely available to academic users at http://www-nmr.cabm.rutgers.edu. AutoStructure analyzes NOEs and generates constraints for structure calculations. At least one of the structure calculation programs XPLOR/CNS, DYANA is required to be installed before running AutoStructure for iterative NOESY data analysis.

AutoStructure can automatically generate constraints for XPLOR/CNS, DYANA structure calculations. Manual constraints, including RDCs can also be used in structure calculations and the resulting structures used for iterative analysis of AutoStructure; however, individual manual constraints are not directly used in the AutoStructure analysis. Initial structure model or homology models can be used as input for AutoStructure analysis, which can indentify NOE interactions that are consistent with the input model.

AutoStructure can also be used at varies stage of resonance assignments for validation. For example, given backbone resonance assignments and 3D ^{15}N-NOESY peak lists, AutoStructure can assign all backbone related intra and sequential NOEs and identify all secondary structure elements. These backbone related intra and sequential NOE connectivities are commonly used for cross-validation of backbone sequential connectivity derived from triple resonance methods. Given near complete backbone and side-chain resonance assignments and 3D HCCH-COSY peak lists, AutoStructure can assign all peaks in the 3D HCCH-COSY peaks for validation of the two-bond and three-bond connectivity of the side-chain resonances.

Testing AutoStructure

AutoStructure was developed and tested using several different experimental input data sets. For all test proteins, low rmsd's were obtained across the final structures, which by conventional criteria are indicative of high-quality structure determinations. The AutoStructure program has been used in over a dozen protein structure determinations(Aramini, et al., 2004, Aramini, et al., 2003, Greenfield, et al., 2001, Greenfield, et al., 2003, Huang, et al., 2003, Makokha, et al., 2004, Ramelot, et al., 2003, Sahota, et al., 2004). Figure 4 shows AutoStructure results for the human basic fibroblast growth factor (154 amino acid residues), together with a comparison with the structure obtained by manual analysis of the same NMR data (Moy, et al., 1996) and by X-ray crystallography. Figure 4 also presents a de novo structure determination for a homodimeric

33-residue-per-chain coiled-coil protein using AutoStructure (Greenfield, et al., 2001).

Minimal Constraint Approaches to Rapid Automated Fold Determination

Medium-accuracy fold information can often provide key clues about protein evolution and biochemical function(s). Extending ideas originally proposed by Kay and coworkers for determining low-resolution structures of larger proteins (Gardner, et al., 1997), a largely automatic strategy has been developed for rapid determination of medium-accuracy protein backbone structures using deuterated, ^{13}C-, ^{15}N-enriched protein samples with selective protonation of side-chain methyl groups ($^{13}CH_3$) (Zheng, et al., 2003). Data collection includes acquiring NMR spectra for automatically determining assignments of backbone and side-chain ^{15}N, H^N resonances, and side-chain $^{13}CH_3$ methyl resonances. Conformational constraints are automatically derived using these chemical shifts, amide ^{1}H/^{2}H exchange, NOESY spectra, and residual dipolar coupling data. The total time required for collecting and analyzing such NMR data and generating medium-resolution but accurate protein folds can potentially be as short as a few days (Zheng, et al., 2003).

Structure Quality Assessment Tools

One of the most important challenges in modern protein NMR is to develop a fast and sensitive structure quality assessment measure which can evaluate the “goodness-of-fit” of a 3D structure compared with its NOESY peak lists and indicate the correctness of its fold. This is especially critical for automated NOESY interpretation and structure determination approaches. One approach uses an NMR R-factor similar to that used in X-ray crystallography, which often require computationally intensive, complete relaxation matrix calculations(Gonzalez, et al., 1991, Gronwald, et al., 2000, Zhu, et al., 1998). We have developed a set of quality scores Recall, Precision, F-measure (NMR RPF scores) from information retrieval to assess the global "goodness-of-fit". These statistical RPF scores are quite rapid to compute, since NOE assignments and complete relaxation matrix calculations are not required, and are valuable in assessing protein NMR structure accuracy.

The quality of an NMR structure is also defined by a number of structural parameters including fold and packing quality, deviations of bond lengths and bond angles from standard values, backbone and side-chain dihedral angle distributions, hydrogen-bond geometery, and close contacts between atoms. Currently there does not exist a single comprehensive structure validation program which takes all these structural parameters into account to evaluate the overall quality of the structure. However, a number of different individual structure quality software packages exist which report scores quantifying some key structural parameters, such as ProCheck_nmr (Laskowski, et al., 1996), WHAT IF(Vriend, 1990), PDBStat (Bhattacharya, et al.), Verify 3D (Eisenberg, et al., 1997), PDB Validation Software (Westbrook, et al., 2003), MAGE (Word, et al., 2000). In the NESG Consortium, we have developed an overall structure quality report which takes into account output from all of the programs mentioned above, and others, and evaluates their output based on a Z-Score which normalizes the results of all the software against a set of high-resolution X-ray crystal structures. This tool handles all data format conversions required to run the software mentioned above and presents the output as a series of easy to read reports and graphs for one-step structure quality evaluation.

An Integrated Platform for Automated NMR Structure Analysis

Protein NMR spectroscopists depend on a number of software packages to facilitate the analysis of data. For this reason, the computational challenge of solving a protein structure by NMR presents a formidable technical challenge to scientists. While a number of software packages have been developed for the analysis of NMR data, a comprehensive solution for the complete automated analysis of NMR data from FIDs to three-dimensional structures is not yet available. Users choose between a number of different software programs each specialized in a certain step of the structural determination process. As a result, a dramatic learning curve exists for a scientist to become proficient enough with all the necessary software in order to do his or her job. Furthermore, invaluable time is often wasted on trivial tasks such as preparing the output of one program to be usable for the next. Also, inter and in some cases even intra laboratory data exchange becomes extremely difficult when people are using a number of different formats required by the various pieces of software available. To add to this complexity, with data passing between so many sources organization quickly becomes a problem. Precious data is often lost due to disorganization. This disorganization can lead to irreproducible results and curb the development of future technologies.

The CCPN effort (Fogh, et al., 2002) (http://www.bio.cam.ac.uk/nmr/ccp/) is attempting to address these problems in data organization and pipelining by developing a detailed data model to capture the complete NMR structure determination process. The data model is not only a standard solution for NMR databases to be implemented under but also an application programming interface (API) to unify the development of future NMR software. The ANSIGv3.3 (Helgstrand, et al., 2000) spectral visualization software is an example of software developed over the CCPN data model.

The SPINS (Baran, et al., 2002) software provides an alternative solution to the integration problem. The SPINS data model is designed to easily accommodate any software available to the community. Rather than designing a data model for the world to adopt, the SPINS data model is intended for internal use by SPINS as a means to easily integrate any software. The SPINS data model was designed to be compatible with the BMRB NMRStar format, thus ensuring compatibility with other public domain efforts.

The current implementation of SPINS integrates several pieces of third party pieces of software (Fig. 5), presenting them as a single application to the user. The SPINS software makes use of the following programs, (i) the SPINS (Baran, et al., 2002) database for storage and organization of raw FIDs, peak lists, chemical shift lists, constraint lists, 3D structures, and other intermediate results; (ii) AutoProc (Monleon, et al., 2002), a spectral referencing and processing script generating program; (iii) NMRPipe (Delaglio, et al., 1995) for executing multidimensional Fourier transformations using scripts generated by AutoProc; (v) NMRDraw (Delaglio, et al., 1995) spectral visualization software for evaluating spectral quality; (vi) SPARKY (Goddard and Kneller, 2000) spectral visualization software, launched out of SPINS, for peak picking and interactive peak list editing; (vii) AutoPeak software

(Monleon, et al., 2002, Moseley, et al., 2001) for interspectral registration, automated peak list editing, and peak data validation; (viii) AutoAssign (Moseley, et al., 2001, Zimmerman, et al., 1997) automated backbone assignment software; (ix) Assignment Validation Suite software (AVS) (Moseley, et al., 2004), providing statistical and graphical tools for validating the quality of the assignments; and (x) AutoStructure, along with DYANA (Guntert, et al., 1997), XPLOR-nih (Schwieters, et al., 2003) or CNS (Brunger, et al., 1998) to iteratively assign NOESY peak lists and generate 3D structures.

The SPINS software provides an integrated process and user interface for using the software packages described above without having to worry about the numerous I/O complexities associated with data analysis using multiple software packages. Furthermore, the process is warehoused by the underlying SPINS database, making it completely reproducible. The completed process can be automatically exported in a standard format (NMRStar 3.1) for submission to the BMRB (Seavey, et al., 1991).

Conclusions

Recent developments provide automated analysis of NMR assignments and 3D structures. These approaches are generally applicable to proteins ranging from about 50 to 150 amino acids. While progress over the last few years is encouraging, even for small proteins more work is required before automated structural analysis is routine. In particular, general methods for automated analysis of sidechain resonance assignments are not yet well developed, though current efforts in this area are quite promising. Moreover, little work has focused on the specific problems associated with nucleic acid structures. The critical area of quality assessment has evolved significantly over the last few years and involves evaluation of both intermediate and final peak lists, resonance assignments, and structural information derived from the NMR data. However, while various resonance assignment and 3D structure "R factors" are beginning to be used, no community-wide consensus has been reached on how to evaluate the accuracy and precision of a protein NMR structure. Despite these significant challenges, when good quality data are available, automated analysis of protein NMR assignment and structures is both fast and reliable. Moreover, automation methods are beginning to have a broad impact on the structural NMR community.

Acknowledgments

We thank J. Aramini, A. Bhattacharya, G. Sahota, D. Snyder, G.V.T. Swapna, R. Tejero, and D. Zheng for useful discussions and for their efforts over the last several years in developing automated NMR data analysis algorithms and software. The authors' recent work on automated NMR data analysis has been supported by the NIH Protein Structure Initiative (P50-GM62413).

Figure Legends

Fig. 1. Flowchart of the overall process of protein structure analysis from NMR data.

Fig. 2. Flowchart outlining the use of the RDpack. Steps required solely for the *de novo* implementation of RD NMR experiments are shown in red boxes, while steps required for rapid adjustment of parameter sets are displayed in green boxes. First, shaped pulses are generated by use of a shell scripts, and the power levels for pulsed field gradients are adjusted to the available hardware configuration (using the macro RD_gscale). Second, the 3D <u>HC</u>CH parameter set is updated by providing proton and carbon high power pulse widths, power levels and carrier positions. Execution of the macro 'RD_setup' transfers these parameters to the entire suite of RD experiments. Third, the 3D HACAcoNH parameter set is updated by providing nitrogen high power pulse width, power level and carrier position. These parameters are the transferred to nitrogen resolved RD experiments by use of RD_setup. Finally, the macro RD_1d starts the acquisition of the first FID of all 11 parameter sets, while also allowing rapid assessment of the relative sensitivity of the various experiments.

Fig. 3. The control-flow of AutoStructure. AutoStructure uses a bottom-up iterative approach. It has four major steps. Step 1 construct an ambiguous distsance network G_{ANOE}, in which all vertices represent protons and proton pairs are connected when their chemical shift values are matched with a NOE peak's chemical shift values within a loose match tolerance. A heuristic HG_{NOE} is initialized from G_{ANOE} at step 2. After HG_{NOE} is initialized, an initial fold is generated at step 3. Step 4 iteratively refine HG_{NOE} from the structures generated from

step 3.

Fig. 4. Results of automatic analysis of protein structures from NMR data. (a) Comparison of backbone structures of human basic fibroblast growth factor (FGF) determined by manual analysis of NMR data (PDB code 1bld), by automated analysis of the same NMR data using AutoStructure / XPLOR , or by X-ray crystallography (PDB code 1bas). The superposition of 10 NMR structures of human basic fibroblast growth factor (FGF) computed by AutoStructure with XPLOR is also shown. Backbone conformations are shown only for residues 29 to 155, since the N-terminal polypeptide segment is not well defined in either the automated or manual analysis. For this portion of the structure, the backbone r.m.s.d.'s within the families of structures determined by AutoStructure are ~0.7 Å and the backbone r.m.s.d. between the AutoStructure and the X-ray crystal structure or manually-determined NMR structures is ~0.8 Å. (b) Solution NMR structure of TM1bZip N-terminal segment of human α-tropomyosin determined by AutoStructure with DYANA(Greenfield, et al., 2001). The top panels show superpositions of backbone (left) and all heavy (right) atoms, respectively. Secondary structures are colored in red. The bottom panel shows ribbon diagrams of one representative structure.

Fig. 5. The Integrated SPINS Platform for Automated Analysis of NMR Data. This figure depicts the flow of data through the SPINS software from raw FIDs to backbone assignments. (i) The raw FID data are housed in the SPINS database. (ii) AutoProc queries the SPINS database for auto-referencing and processing of experimental data using NMRPipe. (iii)

Sparky software used for manual peak picking and peak list editing. (iv) AutoPeak software used to validate peak lists as well as prepare AutoAssign input. (v) AutoAssign software is used for automated backbone resonance assignments. The SPINS platform also integrates AutoStructure software for NOESY data analysis, together with DYANA / CNS / XPLOR software for 3D structure generation and AutoQF software providing estimates of structure quality NMR RPF scores.

Literature Cited

Adler, M. (2000) Modified genetic algorithm resolves ambiguous NOE restraints and reduces unsightly NOE violations. *Proteins* **39,** 385-392.

Andrec, M., and Prestegard, J. H. (1998) A Metropolis Monte Carlo implementation of bayesian time-domain parameter estimation: application to coupling constant estimation from antiphase multiplets. *J Magn Reson* **130,** 217-232.

Andrec, M., and Levy, R. M. (2002) Protein sequential resonance assignments by combinatorial enumeration using 13C alpha chemical shifts and their (i, i-1) sequential connectivities. *J Biomol NMR* **23,** 263-270.

Aramini, J., Xiao, R., Huang, Y. J., Acton, T., Wu, M. J., Mills, J. L., Tejero, R., Szyperski, T., and Montelione, G. T. (2004) Solution structure of the hypothetical protein Yggu from E. Coli. **in preparation**.

Aramini, J. M., Huang, Y. J., Cort, J. R., Goldsmith-Fischman, S., Xiao, R., Shih, L. Y., Ho, C. K., Liu, J., Rost, B., Honig, B., Kennedy, M. A., Acton, T. B., and Montelione, G. T. (2003) Solution NMR structure of the 30S ribosomal protein S28E from Pyrococcus horikoshii. *Protein Sci* **12,** 2823-2830.

Atreya, H. S., Sahu, S. C., Chary, K. V., and Govil, G. (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). *J Biomol NMR* **17,** 125-136.

Baran, M. C., Moseley, H. N., Sahota, G., and Montelione, G. T. (2002) SPINS: standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra. *J Biomol NMR* **24,** 113-121.

Baran, M. C., Huang, Y. J., Moseley, H., and Montelione, G. T. (2004) Automated analysis of protein NMR assignments and structures. *Chemical Reviews* **in press**.

Bartels, C., Billeter, M., Guntert, P., and Wuthrich, K. (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *J Biomol NMR* **7,** 207-213.

Bartels, C., Guntert, P., Billeter, M., and Wuthrich, K. (1997). *J Comput Chem* **18,** 139-149.

Bax, A., Clore, G. M., Driscoll, P. C., Gronenborn, A. M., Ikura, M., and Kay, L. E. (1990). *J Magn Reson* **87,** 620-627.

Bax, A., Clore, G. M., and Gronenborn, A. M. (1990) 1H-1H correlation via isotropic mixing of 13C magnetisation, a new three-dimensional approach for assigning 1H and 13C spectra of 13C-enriched proteins. *J Magn Reson* **88,** 425-431.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res* **28,** 235-242.

Bhattacharya, A., Tejero, R., and Montelione, G. T. Protein Structure Validation Software (PSVS) Suite, and its Applications in Evaluating Protein Structures Generated by Structural Genomics Consortia. *in-preparation*.

Billeter, M., Braun, W., and Wuthrich, K. (1982) Sequential resonance assignments in protein 1H nuclear magnetic resonance spectra. Computation of sterically allowed proton-proton distances and statistical analysis of proton-proton distances in single crystal protein conformations. *J Mol Biol* **155,** 321-346.

Brunger, A. T. (1992) X-PLOR, Version 3.1 : A system for X-ray crystallography and NMR. Yale University Press, New Haven.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54,** 905-921.

Buchler, N. E., Zuiderweg, E. R., Wang, H., and Goldstein, R. A. (1997) Protein heteronuclear NMR assignments using mean-field simulated annealing. *J Magn Reson* **125,** 34-42.

Celda, B., and Montelione, G. T. (1993) Total correlation spectroscopy (TOCSY) of proteins using co-addition of spectra recorded with several mixing times. *J Magn Reson* **B101,** 189-193.

Chance, M. R., Bresnick, A. R., Burley, S. K., Jiang, J. S., Lima, C. D., Sali, A., Almo, S. C., Bonanno, J. B., Buglino, J. A., Boulton, S., Chen, H., Eswar, N., He, G., Huang, R., Ilyin, V., McMahan, L., Pieper, U., Ray, S., Vidal, M., and Wang, L. K. (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci* **11,** 723-738.

Chothia, C., Levitt, M., and Richardson, D. (1981) Helix to helix packing in proteins. *J Mol Biol* **145,** 215-250.

Chothia, C. (1984) Principles that determine the structure of proteins. *Annu Rev Biochem* **53,** 537-572.

Clore, G. M., Omichinski, J. G., Sakaguchi, K., Zambrano, N., Sakamoto, H., Appella, E., and Gronenborn, A. M. (1994) High-resolution structure of the oligomerization domain of p53 by multidimensional NMR. *Science* **265,** 386-391.

Coggins, B. E., and Zhou, P. (2003) PACES: Protein sequential assignment by computer-assisted exhaustive search. *J Biomol NMR* **26,** 93-111.

Cohen, F. E., Sternberg, M. J., and Taylor, W. R. (1982) Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. *J Mol Biol* **156,** 821-862.

Cornilescu, G., Delaglio, F., and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* **13,** 289-302.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6,** 277-293.

Eccles, C., Guntert, P., Billeter, M., and Wuthrich, K. (1991) Efficient analysis of protein 2D NMR spectra using the software package EASY. *J Biomol NMR* **1,** 111-130.

Eisenberg, D., Luthy, R., and Bowie, J. U. (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* **277,** 396-404.

Farmer, B. T., and Venters, R. A. (1995) Assignment of Side-Chain 13C Resonances in Perdeuterated Proteins. *J Am Chem Soc* **117,** 4187-4188.

Fesik, S. W., Eaton, H. L., Olejniczak, E. T., and Zuiderweg, E. R. (1990) 2D and 3D NMR Spectroscopy Employing 13C-13C Magnetization Transfer by Isotropic Mixing. Spins Sytem Identification in Large Proteins. *J Am Chem Soc* **112**.

Fogh, R., Ionides, J., Ulrich, E., Boucher, W., Vranken, W., Linge, J. P., Habeck, M., Rieping, W., Bhat, T. N., Westbrook, J., Henrick, K., Gilliland, G., Berman, H., Thornton, J., Nilges, M., Markley, J., and Laue, E. (2002) The CCPN project: an interim report on a data model for the NMR community. *Nat Struct Biol* **9,** 416-418.

Gardner, K. H., Rosen, M. K., and Kay, L. E. (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry* **36,** 1389-1401.

Garrett, D. S., Powers, R., Gronenborn, A. M., and Clore, G. M. (1991) A Common Sense Approach to Peak Picking in Two-, Three-, and Four Dimensional Spectra Using Automatic Computer Analysis of Contour Diagrams. *J. Magn. Reson* **95,** 214-230.

Goddard, T. D., and Kneller, D. G. (2000), Universtiy of California, San Francisco, CA.

Gong, W. M., Liu, H. Y., Niu, L. W., Shi, Y. Y., Tang, Y. J., Teng, M. K., Wu, J. H., Liang, D. C., Wang, D. C., Wang, J. F., Ding, J. P., Hu, H. Y., Huang, Q. H., Zhang, Q. H., Lu, S. Y., An, J. L., Liang, Y. H., Zheng, X. F., Gu, X. C., and Su, X. D. (2003) Structural genomics efforts at the Chinese Academy of Sciences and Peking University. *J Struct Funct Genomics* **4,** 137-139.

Gonzalez, C., Rullmann, J. A. C., Bonvin, A. M. J. J., Boelens, r., and Kaptein, R. (1991) Toward an NMR R Factor. *J. Magn. Reson* **91,** 659-664.

Greenfield, N. J., Huang, Y. J., Palm, T., Swapna, G. V., Monleon, D., Montelione, G. T., and Hitchcock-DeGregori, S. E. (2001) Solution NMR structure and folding dynamics of the N terminus of a rat non-muscle alpha-tropomyosin in an engineered chimeric protein. *J Mol Biol* **312,** 833-847.

Greenfield, N. J., Swapna, G. V., Huang, Y., Palm, T., Graboski, S., Montelione, G. T., and Hitchcock-DeGregori, S. E. (2003) The structure of the carboxyl terminus of striated alpha-tropomyosin in solution reveals an unusual parallel arrangement of interacting alpha-helices. *Biochemistry* **42,** 614-619.

Grishaev, A., and Llinas, M. (2002) CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proc Natl Acad Sci U S A* **99,** 6707-6712.

Gronwald, W., Willard, L., Jellard, T., Boyko, R. F., Rajarathnam, K., Wishart, D. S., Sonnichsen, F. D., and Sykes, B. D. (1998) CAMRA: chemical shift based computer aided protein NMR assignments. *J Biomol NMR* **12,** 395-405.

Gronwald, W., Kirchhofer, R., Gorler, A., Kremer, W., Ganslmeier, B., Neidig, K. P., and Kalbitzer, H. R. (2000) RFAC, a program for automated NMR R-factor estimation. *J Biomol NMR* **17,** 137-151.

Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J., Kremer, W., Neidig, K. P., and Kalbitzer, H. R. (2002) Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE). *J Biomol NMR* **23,** 271-287.

Grzesiek, S., Anglister, J., and Bax, A. (1993) Correlation of Backbone Amide and Aliphatic Side-Chain Resonances in 13C/15N-Enriched Proteins by Isotropic Mixing of 13C Magnetization. *J Magn Reson* **101,** 114-119.

Grzesiek, S., and Bax, A. (1993) Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. *J Biomol NMR* **3,** 185-204.

Gschwind, R. M., Gemmecker, G., and Kessler, H. (1998) A spin system labeled and highly resolved ed-H(CCO)NH-TOSCY experiment for the facilitated assignment of proton side chains in partially deuterated samples. *J Biomol NMR* **11,** 191-198.

Guntert, P., Dotsch, V., Wider, G., and Wuthrich, K. (1992). *J. Magn. Reson* **2,** 395-405.

Guntert, P., Mumenthaler, C., and Wuthrich, K. (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* **273,** 283-298.

Guntert, P., Salzmann, M., Braun, D., and Wuthrich, K. (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. *J Biomol NMR* **18,** 129-137.

Heinemann, U., Frevert, J., Hofmann, K., Illing, G., Maurer, C., Oschkinat, H., and Saenger, W. (2000) An integrated approach to structural genomics. *Prog Biophys Mol Biol* **73,** 347-362.

Helgstrand, M., Kraulis, P., Allard, P., and Hard, T. (2000) Ansig for Windows: an interactive computer program for semiautomatic assignment of protein NMR spectra. *J Biomol NMR* **18,** 329-336.

Herrmann, T., Guntert, P., and Wuthrich, K. (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR* **24,** 171-189.

Herrmann, T., Guntert, P., and Wuthrich, K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* **319,** 209-227.

Huang, Y. J., Swapna, G. V., Rajan, P. K., Ke, H., Xia, B., Shukla, K., Inouye, M., and Montelione, G. T. (2003) Solution NMR structure of ribosome-binding factor A (RbfA), a cold-shock adaptation protein from Escherichia coli. *J Mol Biol* **327,** 521-536.

Hyberts, S. G., and Wagner, G. (2003) IBIS--a tool for automated sequential assignment of protein spectra from triple resonance experiments. *J Biomol NMR* **26,** 335-344.

Ikura, M., Kay, L. E., and Bax, A. (1990) A novel approach for sequential assignment of 1H, 13C, and 15N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry* **29,** 4659-4667.

Janin, J., and Chothia, C. (1980) Packing of alpha-helices onto beta-pleated sheets and the anatomy of alpha/beta proteins. *J Mol Biol* **143,** 95-128.

Kay, L. E., Ikura, M., and Bax, A. (1990). *J Am Chem Soc* **112,** 888-889.

Kennedy, M. A., Montelione, G. T., Arrowsmith, C. H., and Markley, J. L. (2002) Role for NMR in structural genomics. *J Struct Funct Genomics* **2,** 155-169.

Koehl, P. (1999). *Prog. NMR Spec.* **34,** 257.

Koradi, R., Billeter, M., Engeli, M., Guntert, P., and Wuthrich, K. (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J Magn Reson* **135,** 288-297.

Kuszewski, J., Schwieters, C. D., Garrett, D. S., Byrd, R. A., Tjandra, N., and Clore, G. M. (2004). *J Am Chem Soc* **26,** 6258-6273.

Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* **8,** 477-486.

Leutner, M., Gschwind, R. M., Liermann, J., Schwarz, C., Gemmecker, G., and Kessler, H. (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *J Biomol NMR* **11,** 31-43.

Li, K. B., and Sanctuary, B. C. (1997) Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. *J Chem Inf Comput Sci* **37,** 467-477.

Lin, Y., and Wagner, G. (1999) Efficient side-chain and backbone assignment in large proteins: application to tGCN5. *J Biomol NMR* **15,** 227-239.

Logan, T. M., Olejniczak, E. T., Xu, R. X., and Fesik, S. W. (1992) Side chain and backbone assignments in isotopically labeled proteins from two heteronuclear triple resonance experiments. *FEBS Lett* **314,** 413-418.

Lukin, J. A., Gove, A. P., Talukdar, S. N., and Ho, C. (1997) Automated probabilistic method for assigning backbone resonances of (13C,15N)-labeled proteins. *J Biomol NMR* **9,** 151-166.

Makokha, M., Huang, Y. J., Montelione, G. T., Edison, A. S., and Barbar, E. (2004) The solution structure of the pH-induced monomeric dyein light chain LC8 from Drosophilla. *Protein Sci* **in press**.

Malmodin, D., Papavoine, C. H., and Billeter, M. (2003) Fully automated sequence-specific resonance assignments of hetero- nuclear protein spectra. *J Biomol NMR* **27,** 69-79.

Monleon, D., Colson, K., Moseley, H. N., Anklin, C., Oswald, R., Szyperski, T., and Montelione, G. T. (2002) Rapid analysis of protein backbone resonance assignments using cryogenic probes, a distributed Linux-based computing architecture, and an integrated set of spectral analysis tools. *J Struct Funct Genomics* **2,** 93-101.

Montelione, G. T., Lyons, B. A., Emerson, S. D., and Tashiro, M. J. (1992) An efficient triple resonance experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically enriched proteins. *J Am Chem Soc* **114,** 10974 - 10975.

Montelione, G. T., Rios, C. B., Swapna, G. V. T., and Zimmerman, D. E. (1999) Biological Magnetic Resonance. *in* "NMR pulse sequences and computational approaches for automated analysis of sequence-specific backbone resonance assignments in proteins." (E. Berliner and N.R. Krishna, Ed.), Vol. v17, pp. 81-130.

Moseley, H. N., and Montelione, G. T. (1999) Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol* **9,** 635-642.

Moseley, H. N., Monleon, D., and Montelione, G. T. (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. *Methods Enzymol* **339,** 91-108.

Moseley, H. N., Sahota, G., and Montelione, G. T. (2004) Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. *J Biomol NMR* **28,** 341-355.

Moy, F. J., Seddon, A. P., Bohlen, P., and Powers, R. (1996) High-resolution solution structure of basic fibroblast growth factor determined by multidimensional heteronuclear magnetic resonance spectroscopy. *Biochemistry* **35,** 13552-13561.

Mumenthaler, C., and Braun, W. (1995) Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J Mol Biol* **254,** 465-480.

Mumenthaler, C., Guntert, P., Braun, W., and Wuthrich, K. (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J Biomol NMR* **10,** 351-362.

Nilges, M. (1995) Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J Mol Biol* **245,** 645-660.

Nilges, M., Macias, M. J., O'Donoghue, S. I., and Oschkinat, H. (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. *J Mol Biol* **269,** 408-422.

Orekhov, V. Y., Ibraghimov, I. V., and Billeter, M. (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. *J Biomol NMR* **20,** 49-60.

Pervushin, K., Riek, R., Wider, G., and Wuthrich, K. (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci U S A* **94,** 12366-12371.

Ramelot, T. A., Ni, S., Goldsmith-Fischman, S., Cort, J. R., Honig, B., and Kennedy, M. A. (2003) Solution structure of Vibrio cholerae protein VC0424: a variation of the ferredoxin-like fold. *Protein Sci* **12,** 1556-1561.

Sahota, G., Goldsmith-Fischman, S., Dixon, B., Huang, Y. J., Aramini, J., Yin, C., Xiao, R., Bhattacharya, A., Monleon, D., Swapna, G. V. T., Anderson, S., Honig, B., Monteiro, A. N. A., Montelione, G. T., and Tejero, R. (2004) Solution NMR structure of the BRCT domain from Thermus thermophilus DNA ligase: Surface features suggest novel intermolecular interactions. *Proteins: Struct. Funct. Genetics* **submitted**.

Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, M. G. (2003) The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* **160,** 65-73.

Seavey, B. R., Farr, E. A., Westler, W. M., and Markley, J. L. (1991) A relational database for sequence-specific protein NMR data. *J Biomol NMR* **1,** 217-236.

Szyperski, T., Wider, G., Bushweller, J. H., and Wuthrich, K. (1993) Reduced dimensionality in triple resonance experiments. *J Am Chem Soc* **115,** 9307-9308.

Szyperski, T., Braun, D., Banecki, B., and Wuthrich, K. (1996). *J Am Chem Soc* **118,** 8147-8148.

Szyperski, T., Yeh, D. C., Sukumaran, D. K., Moseley, H. N., and Montelione, G. T. (2002) Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. *Proc Natl Acad Sci U S A* **99,** 8009-8014.

Tejero, R., Monleon, D., Celda, B., Powers, R., and Montelione, G. T. (1999) HYPER: a hierarchical algorithm for automatic determination of protein dihedral-angle constraints and stereospecific C beta H2 resonance assignments from NMR data. *J Biomol NMR* **15,** 251-264.

Terwilliger, T. C. (2000) Structural genomics in North America. *Nat Struct Biol* **7,** 935-939.

Tjandra, N., and Bax, A. (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278,** 1111-1114.

Vriend, G. (1990) WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **8,** 52-56.

Wagner, G., and Wuthrich, K. (1982) Sequential resonance assignments in protein 1H nuclear magnetic resonance spectra. Basic pancreatic trypsin inhibitor. *J Mol Biol* **155,** 347-366.

Westbrook, J., Feng, Z., Burkhardt, K., and Berman, H. M. (2003) Validation of protein structures for the protein data bank. *Methods Enzymol* **374,** 370-385.

Wishart, D. S., and Sykes, B. D. (1994) The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. *J Biomol NMR* **4,** 171-180.

Wishart, D. S., Bigam, C. G., Yao, J., Abildgaard, F., Dyson, H. J., Oldfield, E., Markley, J. L., and Sykes, B. D. (1995) 1H, 13C and 15N chemical shift referencing in biomolecular NMR. *J Biomol NMR* **6,** 135-140.

Word, J. M., Bateman, R. C., Jr., Presley, B. K., Lovell, S. C., and Richardson, D. C. (2000) Exploring steric constraints on protein mutations using MAGE/PROBE. *Protein Sci* **9,** 2251-2259.

Wuthrich, K., Billeter, M., and Braun, W. (1983) Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-proton distance constraints with nuclear magnetic resonance. *J Mol Biol* **169,** 949-961.

Wuthrich, K. (1986) NMR of Proteins and Nucleic Acids. John Wiley & Sons, New York.

Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Terada, T., Ito, Y., Matsuo, Y., Kuroda, Y., Nishimura, Y., Kyogoku, Y., Miki, K., Masui, R., and Kuramitsu, S. (2000) Structural genomics projects in Japan. *Nat Struct Biol* **7,** 943-945.

Zhang, H., Neal, S., and Wishart, D. S. (2003) RefDB: a database of uniformly referenced protein chemical shifts. *J Biomol NMR* **25,** 173-195.

Zheng, D., Huang, Y. J., Moseley, H. N., Xiao, R., Aramini, J., Swapna, G. V., and Montelione, G. T. (2003) Automated protein fold determination using a minimal NMR constraint strategy. *Protein Sci* **12,** 1232-1246.

Zhu, L., Dyson, H. J., and Wright, P. E. (1998) A NOESY-HSQC simulation program, SPIRIT. *J Biomol NMR* **11,** 17-29.

Zimmerman, D., Kulikowski, C., Wang, L., Lyons, B., and Montelione, G. T. (1994) Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. *J Biomol NMR* **4,** 241-256.

Zimmerman, D. E., and Montelione, G. T. (1995) Automated analysis of nuclear magnetic resonance assignments for proteins. *Curr Opin Struct Biol* **5,** 664-673.

Zimmerman, D. E., Kulikowski, C. A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R., and Montelione, G. T. (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. *J Mol Biol* **269,** 592-610.

Zolnai, Z., Lee, P. T., Li, J., Chapman, M. R., Newman, C. S., Phillips, G. N., Jr., Rayment, I., Ulrich, E. L., Volkman, B. F., and Markley, J. L. (2003) Project management system for structural and functional proteomics: Sesame. *J Struct Funct Genomics* **4,** 11-23.

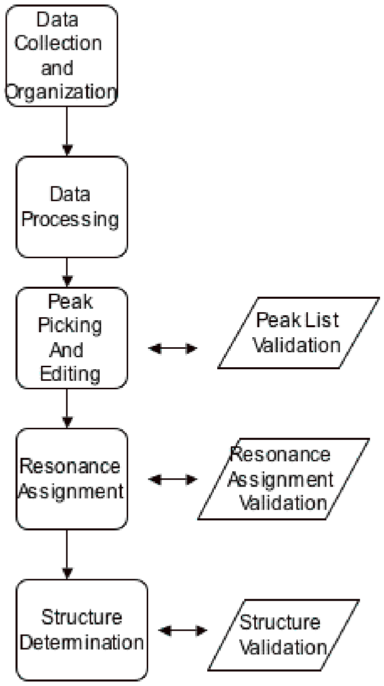
Data Collection and Organization
Data Processing
Peak Picking And Editing
Peak List Validation
Resonance Assignment
Resonance Assignment Validation
Structure Determination
Structure Validation

Figure 1

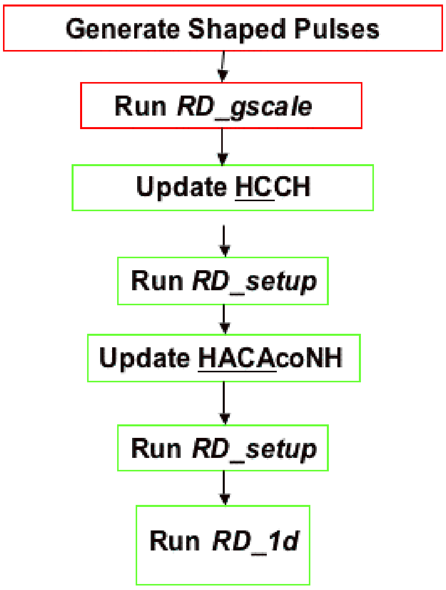
Generate Shaped Pulses
Run *RD_gscale*
Update HCCH
Run *RD_setup*
Update HACAcoNH
Run *RD_setup*
Run *RD_1d*

Figure 2

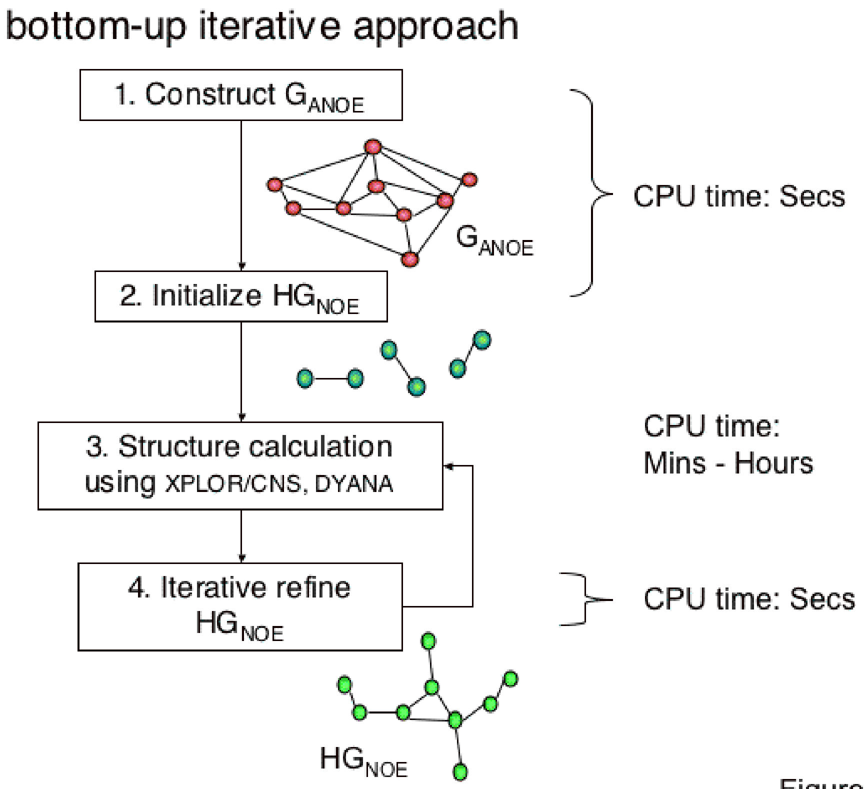
bottom-up iterative approach
1. Construct G_{ANOE}
G_{ANOE}
CPU time: Secs
2. Initialize HG_{NOE}
3. Structure calculation using XPLOR/CNS, DYANA
CPU time: Mins - Hours
4. Iterative refine HG_{NOE}
CPU time: Secs
HG_{NOE}

Figure 3

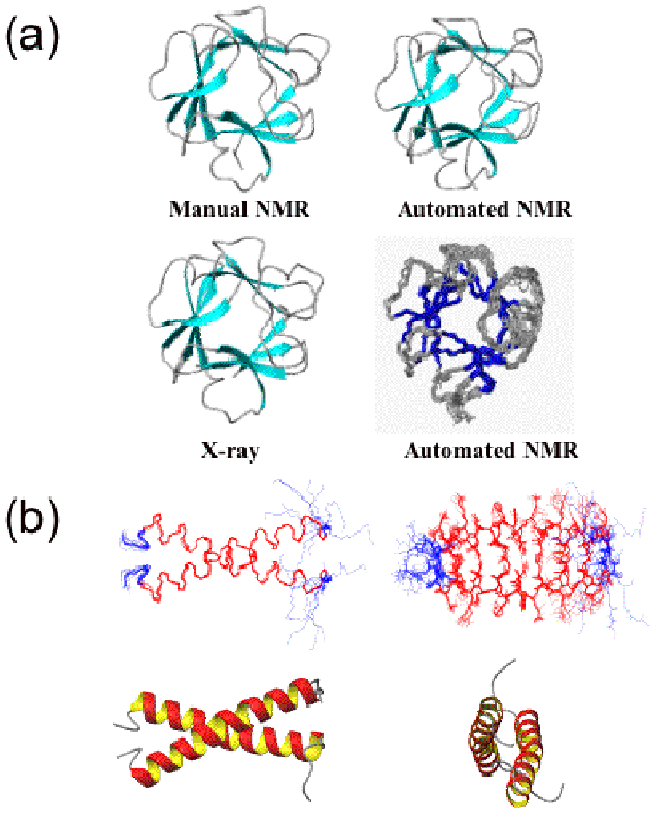

Figure 4

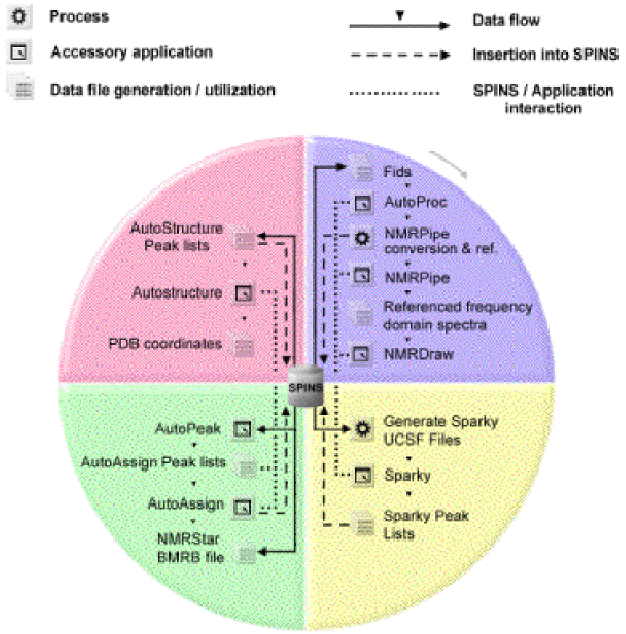
Process
Accessory application
Data file generation / utilization
Data flow
Insertion into SPINS
SPINS / Application interaction
Fids
AutoProc
NMRPipe conversion & ref.
NMRPipe
Referenced frequency domain spectra
NMRDraw
SPINS
Generate Sparky UCSF Files
Sparky
Sparky Peak Lists
AutoPeak
AutoAssign Peak lists
AutoAssign
NMRStar BMRB file
AutoStructure Peak lists
Autostructure
PDB coordinates

Figure 5