Automated Analysis of Protein NMR Assignments and Structures

Michael C. Baran,[†] Yuanpeng J. Huang,[†] Hunter N. B. Moseley,[†] and Gaetano T. Montelione*,[†,‡]

Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers University, Piscataway, New Jersey 08854; and Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Piscataway, New Jersey 08854

Received March 24, 2004

Contents

1. Introduction

A powerful feature of macromolecular structure analysis by NMR spectroscopy is its potential for automation.[1,2] It has been recognized for some time that many of the interactive tasks carried out by an expert in the process of spectral analysis could, in principle, be carried out more efficiently and rapidly by computational systems. Manual methods of protein data analysis often involve using laboratory-specific, or even user-specific, rules of interpretation and validation, which are difficult (if not impossible) to document and reproduce from one laboratory to another. In a sense, until such rules of data interpretation and validation can be standardized, the NMR structure analysis process will retain subjective aspects that limit its reproducibility and compromise its scientific value. For these reasons, a critical next step in evolving a level of scientific maturity in the field of biomolecular NMR is to establish conventions and standards of data interpretation and validation and to instantiate these standards as part of a largely automated process of data analysis. In this review we summarize recent advances in automating the processes of determining 3D structures of proteins from NMR data, with emphasis on those methods of computational and experimental NMR which have been incorporated into automated analysis platforms.

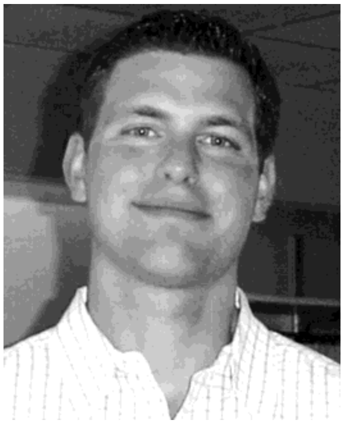
2. Multidimensional, Triple Resonance, and Cryogenic Probe NMR Technologies

Resonance assignments provide the basis for analysis of protein structure and dynamics by NMR spectroscopy.[3,4] The use of multidimensional triple resonance NMR for determination of protein resonance assignments[5−7] has become standard in many laboratories. Indeed, for small proteins (<15 kDa) the process of determining resonance assignments from triple resonance NMR data is, in many cases, a completely routine task. The introduction of RD (reduced-dimensionality)[8−14] and GFT (G-matrix Fourier transformation)[15,16] triple resonance NMR provides powerful approaches for rapid data collection and richer spectral features that are more amenable to automated analysis. For example, these methods can provide 4D and 5D spectral information in 3D and even 2D spectral representations. Other combined experimental and computational methods such as nonlinear sampling with maximum entropy reconstruction,[17,18] Hadamard techniques for selective multichannel excitation and detection,[19,20] and spectral reconstruction from suitably tilted planes[21,22] provide ways to drastically reduce data collection times. Residual dipolar couplings (RDCs)[23−26] and trans hydrogen bond scalar coupling measurements[27−30] also provide extensive and powerful structural constraints that complement information derived from more classical NOESY, scalar coupling, and amide proton exchange experiments.

Motivated by these spectroscopic advances, significant progress has been made in automating the

* To whom correspondence should be addressed. Address: CABM—Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854. Telephone: 732-235-5321. Fax: 732-235-5633. E-mail: guy@cabm.rutgers.edu.
[†] Rutgers University.
[‡] Robert Wood Johnson Medical School.

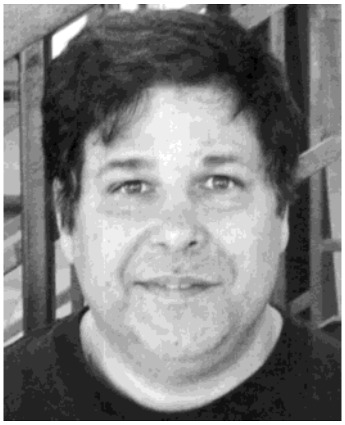
10.1021/cr030408p CCC: $48.50 © 2004 American Chemical Society
Published on Web 07/13/2004



Michael Baran, a native of New Jersey, received his B.S. degree in Biochemistry with a minor in Information Technology from Syracuse University in the year 2000. He is currently enrolled in the Ph.D program in Computational Molecular Biology at Rutgers University. Baran has been a recipient of NIH Biotechnology and NIH Biophysics training fellowship awards. His current research interests lie in the design and development of scientific database systems, the development of bioinformatics software, and high-resolution protein structure determination using NMR.

Yuanpeng Janet Huang, a native of China, received a B.S. degree in Computer Science/Engineering from Zhejiang University in 1994. In 1997, she earned an M.A. in Computer Science from Rutgers University, and in 2001 a Ph.D in Computational Molecular Biology from Rutgers University under the supervision of Prof. Gaetano Montelione. She is currently a Research Assistant Professor in the Center for Advanced Biotechnology and Medicine at Rutgers. Her current research interests lie in automated analysis of NMR data and bionetworks.

Hunter Moseley, a native of Alabama, received a B.S. degree in Chemistry/Computer Science/Mathematics from Huntington College in Montgomery, Alabama, in 1992. In 1998, he received a Ph.D in Biochemistry and Molecular Genetics from the University of Alabama at Birmingham under the mentorship of Professor N. Rama Krishna. During his doctoral studies he was awarded the UAB Samuel B. Barker Award for Excellence in Graduate Studies. Dr. Moseley has carried out postdoctoral studies at Rutgers University in the laboratory of Gaetano Montelione. During this time he was awarded an NSF Postdoctoral Research Fellowship in Biological Informatics. He is currently a Research Assistant Professor in the Center for Advanced Biotechnology and Medicine at Rutgers. His research interests lie in the area of bioinformatics software development in NMR data analysis, and other biophysical techniques and analyses.



Gaetano Montelione, a native of New York, New York, received his B.S. degree from Cornell University in Biochemistry and Cell Biology in 1980. Following some time as a graduate student at The University of Oregon, he received his Ph.D. in Physical Chemistry from Cornell University in 1987 under the joint mentorship of Prof. Harold Scheraga and Prof. Kurt Wüthrich. Montelione then carried out postdoctoral work in Molecular Biophysics with Prof. Gerhard Wagner at the University of Michigan. In 1989, he became Resident Faculty at Center for Advanced Biotechnology and Medicine at Rutgers University, where he is currently Professor of Molecular Biology and Biochemistry, and Director of the Northeast Structural Genomics Consortium. His current research interests focus on the analysis of protein structure and function.

processes of resonance assignment determination, NOESY and RDC data analysis, and 3D structure generation, including for example the algorithms of AutoAssign,[31] ARIA,[32−34] CANDID,[35] and AutoStructure.[36,37] Although such software greatly accelerates the process of going from high-quality peak lists to resonance assignments and 3D structures, the additional processes of data collection, referencing, Fourier transformation, peak picking, and peak list editing now constitute the major portion of the time required for protein structure determination. Moreover, failure to accurately and completely execute these "data processing" tasks results in failures in the automated assignment and structure analysis processes.

Recent developments in cryogenic probe technology for NMR spectroscopy[38,39] provide significant improvements in signal-to-noise ratios for protein samples in aqueous buffered solution and allow much shorter data collection times for most of the triple resonance experiments designed for protein resonance assignments. High-field (800 and 900 MHz) NMR data often provide improved dispersion and well-resolved spectra. Using such modern instrumentation, the data required for analysis of resonance assignments and 3D structure determination of small proteins (<150 residues) can be collected in just a few days. Parallel computer system architectures also provide an important approach to reducing the time required for data processing.[40] Indeed, using current-generation, automated methods of data analysis in favorable cases, resonance assignments and complete

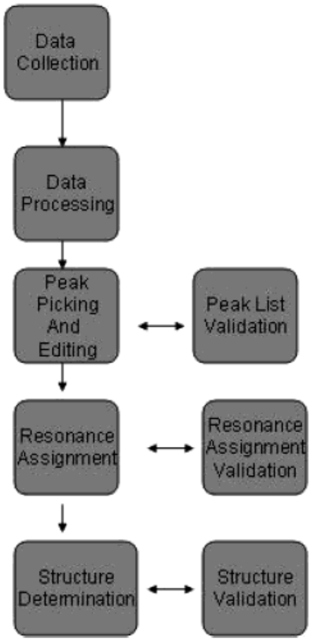


Figure 1. Flow chart of the overall process of protein structure analysis from NMR data.

3D structures of small proteins can now be completed in several days (see, for example, refs 40−42).

3. Organizational Challenge

Despite the technological advances described above, 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, a protein NMR structure determination may require proper collection and analysis of 10−20 individual 2D, 3D, and 4D NMR spectra. These data must be highly consistent, as the input to the structure calculations is a composite generated from across these many data sets. Although it is now possible to collect these data rapidly using cryogenic probes and reduced-dimensionality methods, the large number of data sets required presents logistical challenges. Accordingly, some of the most challenging bottlenecks that remain to high-efficiency protein NMR are organizational rather than scientific.

4. Overview of the Automated Protein Structure Analysis Process

The principal steps of automated protein NMR structure analysis are outlined in Figure 1. These include (i) standardized data collection, (ii) data 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.[43,44] 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 (both intermediate and final) 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.

5. 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 these 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, and efforts must be made to minimize spectrum-to-spectrum variability. Ideally, efforts should be made wherever possible to collect all the data needed for a protein structure analysis back-to-back on the same sample and, where possible, using the same NMR instrument. However, even this strategy does not ensure consistency across spectra, as sample heating effects can depend on decoupler duty cycles, which are different in different NMR experiments. Fortunately, the latest generation NMR probes, and particularly cryogenic probes, exhibit less sample heating from decoupling than previous generation probes.

Another critical organization 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. By defining standard sets of NMR data collection strategies, a robust platform is created with consistent types of input data, guiding users with respect to 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.

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, it is helpful to constrain the digital resolutions in “matching dimensions of complementary spectra” (e.g. the ^{13}C dimensions of HNCA and HNcoCA spectra) to be identical, to maximize accuracy in matching intra-residue and sequential cross-peaks between these spectra. In the activities of the Northeast Structural Genomics Consortium (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.

6. 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, outdated, and error-prone, leading to frequent loss of valuable data that are both hard and expensive to obtain. Archives on tape and optical media are difficult to track and recover, frequently lack adequate organizational and querying facilities, and have limited longevity. On the other hand, disk space is now inexpensive enough to consider using mirrored disk arrays as live disk archives along with regular tape backups. However, whether using archival media or live disk archives, laboratories carrying out multiple protein structure determinations and generating many different data sets create organizational problems that need to be addressed by an appropriate database structure.

The growing demands on data organization and formatting in submitting NMR data and structures to public databases such as the BMRB[45] and the PDB[46] 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[47−52] 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 not only to 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.

Although critical to the process of automated NMR data analysis, there have been only limited efforts to date to develop computer systems for organizing and integrating NMR data analysis software and intermediate results of structure analysis. Examples of recently described NMR laboratory information management system (LIMS) solutions include the SESAME[44] and SPINS[43] databases. The SESAME database[44] provides an easy to use, database-driven, tool for managing NMR data as well as a large-scale structural genomics project. SESAME provides a number of different modules for tracking samples and experiments through a Java client/server architecture. This implementation makes SESAME accessible from anywhere on the Internet, an attribute essential in a large scale genomics project where data may be collected and analyzed from multiple locations. SPINS (standardized protein NMR storage)[43] 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.[45] The SPINS software is implemented in a Java three-tier system architecture (Figure 2). This configuration provides the flexibility necessary to efficiently serve all users by allowing access from anywhere within a laboratory's local Intranet.

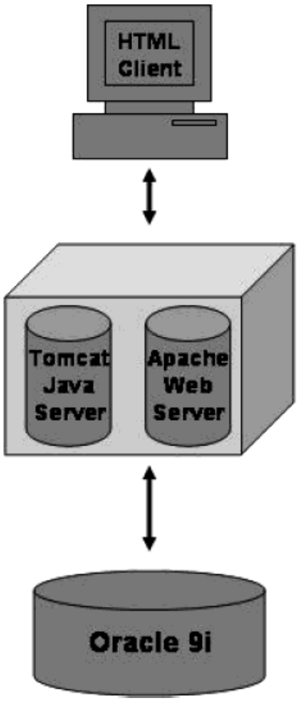


Figure 2. SPINS three-tier system architecture. This figure shows the relationships between the major components implementing SPINS: Oracle 9i database, Tomcat application server, and Apache Web server. We developed SPINS under Oracle 9i v9.0.1 enterprise edition database running on Red Hat Linux 7.1 using Java servlet technology in conjunction with Perl 5.005, and TCL/TK v7.6. Java Database Connectivity (JDBC) drivers and the Perl Database Interface (DBI) drivers provide the direct interface between the Java servlets on the Tomcat server and the Oracle database.

7. 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,[53] 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. Proper chemical shift referencing for aliphatic ^{13}C and ^{1}H resonances is also critical for accurate amino acid typing[31,54,55] and secondary structure analysis,[56] generating information that is used in most automated assignment and structure programs.

Accurate referencing can be done by externally calibrating 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 using these calibrations to define the corresponding chemical shift value of the carrier offset in each dimension of each NMR spectrum.[40]

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.[55,57] This allows for using the tightest possible "match tolerances" in later steps of automated analysis. 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.[40]

Another critical issue for automated analysis is to correctly process spectra in a highly reproducible and timely manner. This can sometimes be a significant bottleneck, even though processing is often viewed as a routine task. Even for expert NMR spectroscopists, the referencing and transformation of several time domain FID data sets into properly phased and referenced frequency domain spectra suitable for analysis typically requires several hours to carry out, and if errors are made in defining processing parameters, significant time can be wasted trouble-shooting processing parameters.

Several high-quality NMR processing programs suitable for incorporation into automated analysis pipelines have been developed over the last several years, including Felix (Molecular Simulations, Inc., San Diego, CA), NMRPipe,[58] PROSA,[59] VNMR (Varian, Inc., Palo Alto, CA), 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 nontrivial amounts of time, knowledge of NMR theory, and familiarity with technical features of the specific data collection methods and/or processing software.

An example of recent approaches to organizing and streamlining NMR data processing is the software package AutoProc.[40] AutoProc is a data dictionary together with a set of software tools designed to allow a nonexpert in NMR spectroscopy to accurately reference multidimensional NMR spectra, generate and run appropriate conversion scripts, and process NMR data using the software package NMRPipe.[58] 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 such as Felix (Molecular Simulations, Inc., San Diego, CA) or PROSA.[59]

8. 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[60−65] 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.[66] Many programs utilize line shape comparisons across spectra or perform restricted peak picking (or filtered peak picking), which is a form of peak editing where one peak list is filtered against another in comparable dimensions.[31,40,63]

The contour approach to peak picking (CAPP)[61] relies primarily on peak shape. After CAPP generates a contour plot, it calculates ellipses that best fit the contours. CAPP then detects potential ridges before finally testing the ellipsoid model of each potential peak against cutoff conditions. Although the results for 2D spectra are generally quite good, 3D spectra still require manual editing. Another popular peak picker, AUTOPSY,[62] has methods to deal with overlap and deviations from ideal Lorentzian line shape. It also takes advantage of symmetry peaks present in some spectra (COSY, NOESY). Multidimensional NMR spectra interpretation (MUNIN)[64] uses a three-way decomposition to decompose a 3D spectrum into a sum of components. Each component can represent one or a group of peaks in the spectrum. ATNOS[65] 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. ATNOS can be used together with NOESY assignment and structure determination software to iteratively identify and assign NOESY cross-peaks.

In our own laboratory, peak picking is usually done using the restrictive peak picking and peak editing facilities in the program Sparky.[63] Additional software, AutoPeak,[40] uses peak lists generated from manually peak picked 2D $^{15}N-^{1}H$ HSQC and $^{13}C-$

^{1}H HSQC spectra as frequency-filters across raw peak lists from 3D spectra. For the peaks which pass these filters, Sparky reports line width, 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., spin system-based peak list quality reports 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.[55]

9. 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) peak list completeness reports, and (iii) spin system-based peak list quality reports generated from the AutoPeak software suite.[55] 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 comparison of chemical shift values for the same resonance in different spectra (calculate_registration[40]) is used to register peak lists from different spectra. This approach has the added benefit of providing standard deviations for matching resonance frequencies between spectra, which is useful in deriving 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 be used to assess the quality of the corresponding data. Interspectral registration data and other peak list quality assessments 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.

10. Pattern-Based Spectral Peak Picking

With the development of RD and GFT NMR experiments, new approaches to peak picking/editing are being realized that circumvent the problems that traditional peak pickers encounter. RD, GFT, and some scalar and dipolar coupling experiments present spectral data in groups of peaks with characteristic relationships between components that provide additional constraints to verify the veracity of each peak in a group. A good example is the peak pattern obtained in RD-TR NMR spectra (Figure 3). With such data, instead of selecting one peak at a time, one can select a group of peaks that, together, fulfill the pattern and, hence, mutually support each other. The PatternPicker[67] program then edits a raw list of peaks by selecting groups of peaks that fit a defined pattern. This algorithm is applicable to any experiment containing a characterizable pattern of peaks. The program is designed to be very flexible with respect to the peak patterns it can recognize and includes facilities to easily craft new patterns. PatternPicker exploits experiments that encode information in "patterns of peaks", even where the "patterns" are spread over multiple spectra, thus promoting the use of these experiments that are typically harder for a human to analyze. This opens up new areas to explore in experiment design, since the complexity of the pattern is now a benefit, and not a drawback, to analysis.

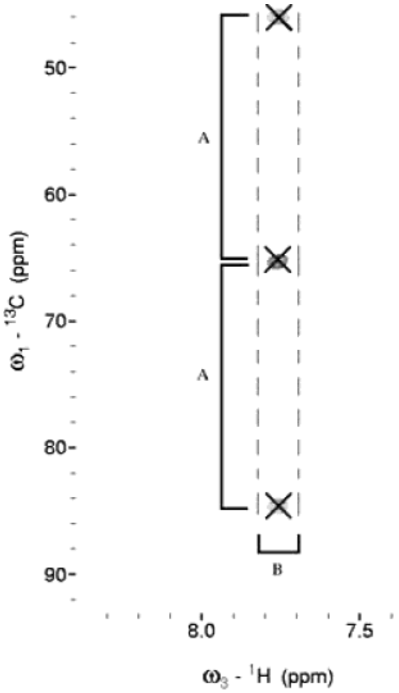


Figure 3. RD-TR NMR pattern model. The following information is used to represent an RD peak pattern (three peak pattern). The center peak is the central peak of the pattern. The outer two peaks are the doublet peaks. Together these peaks have detectable relationships that can be used in automated peak list editing. For example, the distances between the doublet peaks and the central peak along the vertical dimension are identical within a certain matching tolerance. Several such constraints on relative peak positions provide a pattern that is used to distinguish real peaks from noise peaks.[67]

11. 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[68] (reviewed in refs 1, 68, and 69). A summary of some

Table 1. Programs for Automated Protein NMR Resonance Assignments

assignment program	mapping method	backbone[a]	side chain[b]
Andrec and Levy[79]	exhaustive search	yes	no
AutoAssign[31,82]	heuristic best-first	yes	no
Buchler et al.[72]	simulated annealing/Monte Carlo	yes	yes
CAMRA[81]	comparison to predicted shifts	yes	yes
GARANT[75,76]	genetic algorithm	yes	yes
IBIS[84]	heuristic best-first	yes	yes
Li and Sanctuary[83]	heuristic best-first	yes	yes
MAPPER[78]	exhaustive search	yes	no
MONTE[73]	simulated annealing/Monte Carlo	yes	no
PACES[80]	exhaustive search	yes	yes
PASTA[74]	simulated annealing/Monte Carlo	yes	no
TATAPRO[77]	exhaustive search	yes	no

[a] Backbone resonances include H^N, N^H, C′, C^β, and C^α resonances. [b] Resonances further down the side chain than C^β.

of the programs available for automated analysis of resonance assignments is provided in Table 1.

Most (though not all) automated programs use the same general analysis scheme which originates from the classical strategy developed by Wüthrich and coworkers.[3,70,71] Commonly used algorithms for automated analysis of resonance assignments generally include some of the following steps:[1] (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 (Table 1) include simulated annealing/Monte Carlo algorithms[72−74] such as MONTE[73] and PASTA;[74] genetic algorithms such as GARANT;[75,76] exhaustive search algorithms[77−80] such as TATAPRO,[77] MAPPER,[78] and PACES;[80] CAMRA,[81] which performs a heuristic comparison to predicted chemical shifts derived from homologous proteins; and heuristic best-first algorithms[31,82−84] such as IBIS[84] and AutoAssign.[31,82] Another distinguishing characteristic between the methods is the differing types of experimental data used in the analysis. All the programs mentioned use heteronuclear experimental data. A few programs such as GARANT,[75,76] CAMRA,[81] and the Li and Sanctuary algorithm[83] use homonuclear experimental data as well.

AutoAssign[31,55] 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).[31] The current version of AutoAssign is written in C++ with a Java-based GUI.[55] 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.[40,55]

Key components of specific processing (AutoProc,[40] NMRPipe[58]), peak picking (AutoPeak,[40] Sparky[63]), and automated assignment (AutoAssign[31,55]) software 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"[40] 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 and 5.5 h 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 h. Several different subsets of these data collection strategies were compared. This benchmark study demonstrated that nearly complete backbone resonance assignments and secondary structures (based on chemical shift data) for a 58-residue protein can be determined in less than 30 h, 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 h 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 using an integrated platform for automated NMR data analysis.

12. Automated Analysis of Side Chain Resonance Assignments

While several approaches have been found to provide robust automation of backbone resonance assignments, a robust approach to automated side chain assignments is not yet generally available. Several programs listed in Table 1 support automated analysis of side chain resonances with different degrees of robustness. For example, a combined approach using GARANT[75]and AUTOPSY[62] together has demonstrated excellent success in automating both peak picking and resonance assignments, including many side chain aromatic ^{1}H resonance assignments.[85]

The principal challenge in automated analysis of side chain resonances is incompleteness in experimental peak lists generally available for this task. Most published efforts in automating side chain resonance assignments[76,80,84] focus on HCCcoNH-TOCSY,[86,87,88] and use statistical comparisons to expected ^{13}C side chain resonance values of amino acid residues to assign the carbon 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.[89] 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.[90−92] Other methods[72,83] use HCCH-COSY[5,93,94] and/or HCCH-TOCSY[95,96] data to assign side chain carbon and hydrogens. However, these approaches are often challenged by chemical shift degeneracy and incompleteness of input data. For these reasons, we suggest that a robust automated side chain assignment strategy might utilize a combination of HCCcoNH-TOCSY recorded with multiple mixing times, together with data such as HCCH−COSY[5,93,94] and/or HCCH-TOCSY.[95,96]

13. 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. Aside from efforts by the BioMagResDatabase (http://www.bmrb.wisc.edu), there have been few efforts to develop tools for validating resonance assignments when the 3D structure is unknown. One exception is the SHIFTY[97] program, which predicts chemical shift assignments from known assignments of homologous proteins. Other methods, which use chemical shifts calculated from a 3D structure to validate chemical shift assignments, include the SHIFTS[98] and SHIFTX[99] programs. SHIFTX has been used to create a reference corrected version of the BMRB called RefDB.[100]

As part of an integrated platform for protein NMR structure analysis, we have developed a set of computer utilities called the Assignment Validation Software (AVS) suite[101] 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.

The AVS suite includes both new software tools and extensions of the graphical user interface (GUI) component of the AutoAssign software package. They are designed to perform the following tasks: (i) statistical evaluation of individual chemical shifts and their associated amino acid spin system classification against the database of protein chemical shift data, (ii) evaluation of the quality of information used to create segments of linked spin systems in the assignment process, and the uniqueness of their mapping into the protein amino acid sequence, and (iii) visual representation of assignment completeness and consistency with other spectral data not used in the assignment process but useful as additional validation of the assignment results. Figure 4 shows an example of AVS suite output, documenting assignment completeness and consistency of NMR assignments generated for a polypeptide segment of the BRCT domain from *Thermus thermophilus* DNA ligase. This image visually demonstrates the wealth of sequential connectivity and other NMR data supporting these assignments.[101]

14. NOESY Interpretation and Structure Determination

In protein NMR, 3D structures are generated mainly using the following data: (i) distance constraints based on analysis of multidimensional NOESY spectra, (ii) constraints on dihedral angles derived from experimental and/or statistical data, including NOESY, chemical shift, and scalar coupling constant data, (iii) residual dipolar couplings, and (iv) hydrogen bond, and/or disulfide bond, distance constraints derived from other experimental data. Commonly used structure generation calculation programs include DYANA,[102] XPLOR,[103,104] and CNS.[105] DYANA utilizes a dihedral angle representation of protein structure, suitable for fast structure calculations. Both XPLOR and CNS can use dihedral angle and Cartesian space representations. DYANA, XPLOR, and CNS all use dynamical simulated annealing methods for structure calculations. In addition, RDC, pseudopotentials for representing scalar coupling data, secondary $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shift restraints, a conformational database potential, and molecular dynamics simulation in explicit water are often incorporated into XPLOR and CNS calculations for energy minimization and structure refinements.[105,106]

Several approaches have been described for identifying backbone and/or side chain dihedral angle constraints by simultaneous analysis of NOE, scalar coupling, and/or chemical shift data. The program HYPER[107] generates dihedral angle constraints using a conformational grid searching method, which calculates the set of ϕ and ψ dihedral angles and stereospecific assignments of β methylene protons that are consistent with a combined analysis of

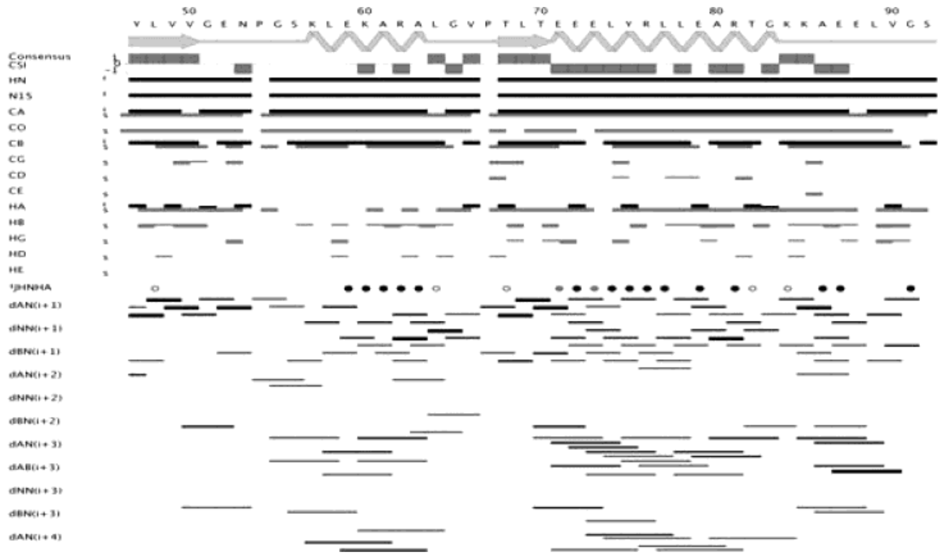


Figure 4. Portion of CMap image generated by the AVS software suite[101] showing assignment completeness and consistency for the BRCT domain from *Thermus thermophilus* DNA ligase. The first row is the protein sequence. The next row annotates the secondary structure. The third row is the consensus chemical shift index (CSI) calculated from the H^{α}, C^{α}, C^{β}, and C′chemical shifts, on the basis of the method of Wishart et al.[56] The next 13 rows summarize triple resonance connectivity data. The next row summarizes ^{3}J ($H^{N}-H^{\alpha}$) scalar coupling data. The final 11 rows summarize sequential and medium-range NOE data derived with the program AutoStructure[36,120] validating the assignments and secondary structure. The software also provides other graphical tools for evaluating inconsistencies of resonance assignments relative to the experimental data. Reprinted with permission from ref 101. Copyright 2004 Kluwer Academic Publishers.

vicinal scalar coupling constants and local intra-residue and sequential NOE data calibrated using the isolated two-spin pair approximation. Gippert et al.[108] have described two complementary approaches involving a systematic searching in torsion angle space for generation of all conformations of polypeptides which satisfy the local conformational constraints. In the TALOS program,[109] protein backbone ϕ and ψ constraints are derived by comparing experimental chemical shifts with a database of high-resolution crystal structures, for which resonance assignments are available. These methods provide robust automated approaches for generating dihedral angle constraints and starting conformations consistent with these local constraints.

One of the principal goals of automated structure determination programs involves iterative analysis of multidimensional NOESY data. Several automated approaches for NOESY interpretation and structure calculation have been developed, including NOAH,[110,111] ARIA,[32,33] CANDID,[35] AutoStructure,[36,37] a self-consist constraint analysis method implemented in XPLOR,[112] and other generally less developed programs.[113−115]

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 interaction, reassigning constraints that are internally inconsistent (self-correcting) in the course of the structure calculation, while ARIA uses an ambiguous constraint strategy,[32,33] involving multiple ambiguous distance constraints for each ambiguous NOESY peak. The program NOAH has been combined with the structure generation programs DYANA and DIAMOD.[116] The program ARIA[32,33] has been combined with the structure generation program CNS. Initial structures are first built using ambiguous constraint strategies and then iteratively refined.

Underlying the ambiguous constraint strategies of ARIA is a key *correctness assumption*: that, for each NOE cross-peak, at least one of its potentially linked proton pairs is a true proton−proton interaction.[32,33] Noise peaks in the NOESY peak lists and missing resonance assignments generally violate this assumption. The program CANDID,[35] combined with DYANA, also uses top-down ambiguous constraint strategies but, in addition, employs network anchoring and constraint-combination methods, minimizing deleterious effects when this *correctness assumption* is not satisfied. Though providing improved robustness, CANDID's network anchoring and constraint-combination methods require some 90% complete resonance assignments (corresponding to ∼87% complete side chain resonances),[117] almost complete aromatic side chain assignments, low percentage of noise peaks, and small chemical shift variations.[118,119] For both ARIA and CANDID, it is also important to obtain a well-converged initial fold (rmsd < 3.0 Å).

AutoStructure[36,37,120] uses a distinct *bottom-up topology-constrained approach*, which distinguishes it from NOAH, ARIA, and CANDID. AutoStructure first builds an initial fold based on intra-residue and sequential NOESY data, together with characteristic NOE patterns of secondary structures, including

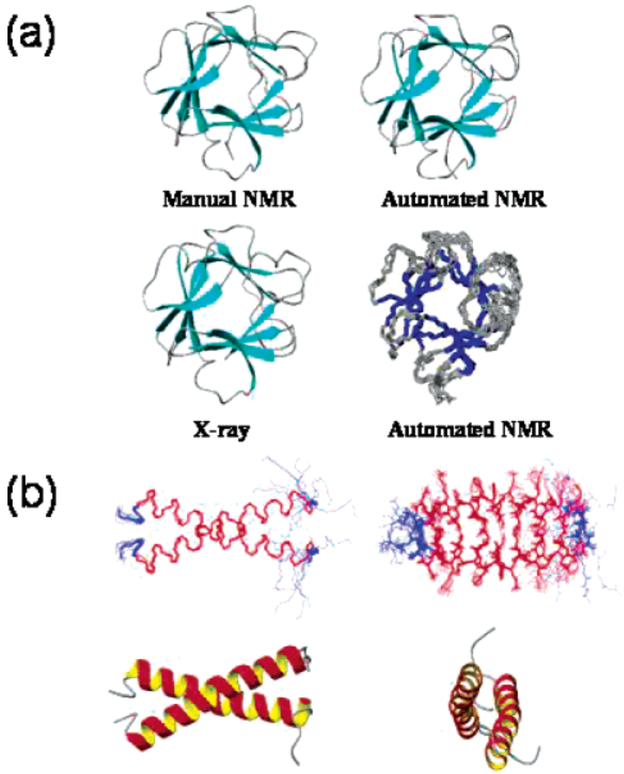


Figure 5. Results of automatic analysis of protein structures from NMR data. (a) Comparison of backbone structures of human basic fibroblast growth factor (FGF) determined by manual analysis of NMR data (PDB code 1bld), by automated analysis of the same NMR data using AutoStructure/XPLOR,[120] or by 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 29−155, 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 rmsd's within the families of structures determined by AutoStructure are ∼0.7 Å and the backbone rmsd between the AutoStructure and the X-ray crystal structure or manually determined NMR structure is ∼0.8 Å. (b) Solution NMR structure of the TM1bZip N-terminal segment of human α-tropomyosin determined by AutoStructure with DYANA.[122] The top panels show superpositions of backbone (left) and all heavy (right) atoms, respectively. Secondary structures are in red. The bottom panel shows ribbon diagrams of one representative structure.

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. This protocol, in principle, 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 and XPLOR/CNS. Figure 5 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[121] and by X-ray crystallography. Figure 5 also presents a *de novo* structure determination for a homodimeric 33-residue-per-chain coiled-coil protein.[122]

By incorporating rules of structural and topological constraints that are similar to those used by a human expert in the structure determination process, the correctness assumption described above is less critical for most algorithms in AutoStructure. General input requirements for reliable performance of AutoStructure include (i) NOESY peak lists containing at least 90% real cross-peaks and (ii) at least 85% complete resonance assignments (corresponding to ∼80% complete side chain resonance assignments).[1] The requirement for high percentage completeness of input data is necessary to accurately define protein core side chain packing in high-resolution structure determinations, especially for aromatics side chains and methyl groups. Given a partially complete input data set, AutoStructure also provides an initial fold analysis that is used for refinement of input data. In general, we expect CYANA, ARIA, and AutoStructure to each exhibit advantages and disadvantages with different NMR data sets. Ideally, it should be possible to routinely compare results for all three methods in a given structural study.

A fully automated robust approach for automated structure analysis has recently been implemented within the NIH-XPLOR package.[112] The approach takes in a large list of NOE restraints created in a simplistic fashion from direct all-to-all matching of NOE peaks to resonance assignments and uses a probabilistic method to turn on and off NOE restraints as the simulated annealing progresses. The approach is very fault tolerant and robust but also very computationally intensive. It should be noted that this approach, involving dynamic analysis of constaint consistency in the course of the structure calculation, is highly complementary to methods described above that attempt to generate correct constraint lists prior to initial 3D structure calculations; the combination of these methods should provide an even more robust and complete solution to the challenge of going from NOESY peak lists to accurate 3D structures.

15. NMR “*R*-Factors”

One of the most important challenges in modern protein NMR is to develop a fast and sensitive structure quality assessment measure which can be used to evaluate the “goodness-of-fit” of the 3D structure compared with NOESY peak lists and to indicate the correctness of the fold. This is especially critical for automated NOESY interpretation and structure determination approaches. One approach uses an *R*-factor definition similar to that used in X-ray crystallography, in which the NOESY spectrum is compared with a simulated NOESY spectrum back-calculated from 3D structure ensembles. However, direct adaptation of the X-ray *R*-factor to NMR data is challenging for several reasons. In the most direct analogy, the distance of every atom (or chemical shift pair) is treated as a lattice point, and the NOE intensity at each point on this lattice is back-calculated from the structure under evaluation. Such a matrix is dominated by the numerous numbers of true negative data, which are not detected in both the experimental and back-calculated NOESY spectra. Such a quality score will not be sensitive and meaningful if all these true negative points are included for quality assessments.

An alternative improved approach is to compare only the intensity differences for peaks observed from experimental and back-calculated spectra.[123−125] However, effects of spin diffusion, internal dynamics, and differential heteronuclear polarization transfer efficiencies make it difficult to make accurate estimates of NOESY cross-peak intensities from interproton distances of 3D structures, even when using complete relaxation matrix calculations.[126,127] The program R-FAC[125] provides a set of NMR *R*-factor scores, including a global *R*-factor and different *R*-factors for the intra-residue NOEs, the inter-residue NOEs, sequential NOEs, medium-range NOEs, and long-range NOEs. R-FAC uses NOESY cross-peaks to amide H^N protons for comparison in a complete relaxation matrix formalism. Recent work[125] suggests that one particular *R*-factor calculated by R-FAC (monitoring long-range NOEs, and referred to as R5) is most useful in measuring the quality of an NMR structure.

Alternative approaches for calculating a statistical global performance score avoid the true negative domination problem while preventing the inaccuracies in peak intensities from dominating the structure quality assessment. The field of information retrieval has encountered a similar true negative domination problem; recall, precision, and *F*-measure are statistical quality scores commonly used in information retrieval applications that are potentially sensitive to a large number of true negatives.[128,129] The AutoQF[120] algorithm uses an analogous *F*-measure quality factor from information retrieval. Recall measures the percentage of peaks in the NOESY peak lists that are consistent with the resonance assignments and are also consistent with the average interproton distances of the query structures. Precision measures the percentage of close distance proton pairs in the query structures whose back-calculated NOE interactions are detected in the NMR data. Both recall and precision quantify how well the 3D model structures agree with resonance assignment and NOESY cross-peak data. Recall and precision are types of NMR *R*-factor measurements but place emphasis on the presence or absence of distance relationships as opposed to the exact distance values, which require accurate complete relaxation matrix calculations. The *F*-measure score is the overall performance score calculated from the recall and precision. The *F*-measure method provides a global measure of the goodness-of-fit of the 3D structures with the NOESY spectra in minutes. AutoQF also uses an *M*-score to measure the completeness of the two- or three-bond connected NOESY cross-peaks, an indicator of the quality of the NOESY cross-peaks with strong intensities, assuming the resonance assignments are all correct.

Recently, a new approach to quantitative evaluation of each experimental NMR restraint (QUEEN method) has been reported.[130] This method is based on a description of the structure in distance space and concepts derived from information theory. The QUEEN method has been shown to be able to successfully identify the crucial (i.e. important and unique) restraints in a structure determination for various examples. This approach to characterizing “critical constraints” should have great value in evaluating the accuracy and robustness of a protein structure derived from NMR data.

16. Structure Quality Assessment Tools

In addition to the “NMR *R*-factors” described above, the quality of an NMR structure is defined by a number of 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 geometry, and close contacts between atoms. Currently there does not exist a single comprehensive structure validation program which takes all these 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. The most commonly used program for NMR structures is ProCheck_nmr,[131] which reports statistics on overall stereochemistry. The program also provides a highly useful graphical representation of the Ramachandran plot, as well as statistics on bond lengths, bond angles, and secondary structures. Another common protein structure validation program, the WHAT IF server, provides several tools for protein structure analysis, validation, and modeling.[132] PDBStat[133] is also useful for computing various statistical analyses given the Cartesian coordinates of a protein. The program is capable of handling many of the complexities associated with data conversion between different standard formats (CHARMM,[134] CONGEN,[135,136] XPLOR,[104] CNS,[105] PDB,[46] and DYANA[102]/CYANA[35]). PDBStat evaluates distance and dihedral angle violations, produces contact maps based on coordinates or constraints, calculates atomic superimpositions and rmsd’s, evaluates order parameters for ϕ and φ dihedral angles,[137] summarizes constraints and coordinates, and computes optimal superimposition transformations, hydrogen bond analysis, close contact distributions, and chirality analysis. Verify 3D[138] evaluates the environment of each individual amino acid within the structure and assigns a probability of that amino acid existing in the given environment. The PDB Validation Software[139] is capable of reporting close contacts as well as other structural inaccuracies such as nonideal bond lengths and bond angles. The MAGE software[140] also analyzes atomic overlaps of protein structures with protons, identifies distortions of the polypeptide backbone, and provides tools for analysis of both backbone and side chain dihedral angle distributions. MAGE also provides a highly useful interface for visualization of the bad contacts and structural distortions in the context of the 3D structure of the protein.

In our integrated structure analysis platform, we generate 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. The tool has been developed to handle all data format conversions required to run the aforementioned software as well as present the output as a series of easy to read reports and graphs which can be used to evaluate structural quality.

17. 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 co-workers for determining low-resolution structures of larger proteins,[141] a largely automatic strategy has been described 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$_3$).[41] 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.[41] Other approaches for rapid fold determination focus on using residual dipolar coupling data to craft medium-resolution backbone folds. One method uses specific analysis tools to reconstruct the protein structure in fragments.[142,143] Another method uses a bounded tree search algorithm to search through a structural database using self-consistency between protein fragments to filter out false positives in the search.[144−146]

18. 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 specializing in a certain step of the structural determination process. As a result, a dramatic learning curve has emerged in which a true expert must be proficient with a number of different pieces of 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 intralaboratory 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 can curb the development of future technologies.

The CCPN effort[148] (http://www.bio.cam.ac.uk/nmr/ccp/) is attempting to address these problems of 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. Many NMR projects are now using the CCPN specification, including new versions of ANSIG, SPARKY, and ARIA.

The SPINS[43] 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 complete compatibility with other public domain efforts.

The current implementation of SPINS integrates several pieces of third party pieces of software (Figure 6), presenting them as a single application to the user. The SPINS software makes use of the following programs: (i) the SPINS[43] database for storage and organization of raw FIDs, peak lists, chemical shift lists, constraint lists, 3D structures, and other intermediate results; (ii) AutoProc,[40] a spectral referencing and processing script-generating program; (iii) NMRPipe[58] for executing multidimensional Fourier transformations using scripts generated by AutoProc; (iv) NMRDraw[58] spectral visualization software for evaluating spectral quality; (v) SPARKY[63] spectral visualization software, launched out of SPINS, for peak picking and interactive peak list editing; (vi) AutoPeak software[40,55] for interspectral registration, automated peak list editing, and peak data validation; (vii) AutoAssign[31,55] automated backbone assignment software; (viii) Assignment Validation Suite software (AVS),[101] providing statistical and graphical tools for validating the quality of the assignments; and (ix) AutoStructure, along with DYANA,[102] XPLOR-nih,[104] or CNS,[105] to iteratively assign NOESY peak lists and generate 3D structures. While still under development, the integrated SPINS NMR structure analysis and validation platform has already been used for a few complete small protein structure determinations.[42,147]

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. Further-

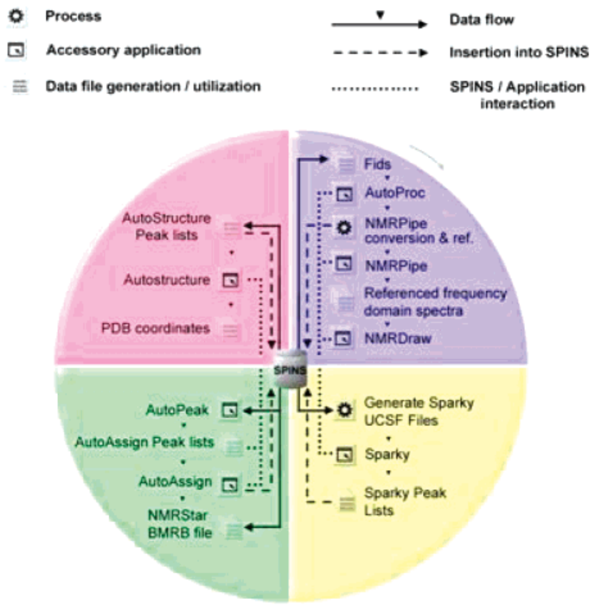


Figure 6. Integrated SPINS platform for automated analysis of NMR data. This figure depicts the flow of data through the SPINS software from raw FIDs to backbone assignments. (i) The raw FID data are housed in the SPINS database. (ii) AutoProc queries the SPINS database for auto-referencing and processing of experimental data using NMRPipe. (iii) Sparky software is used for manual peak picking and peak list editing. (iv) AutoPeak software is 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 "*R*-factors".

more, 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.[45] Furthermore, SPINS provides a set of tools to aid in the structural determination process.

19. 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 past 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 side chain 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. A critical area that has evolved significantly over the past few years involves quality assessment 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 can be both fast and reliable. Moreover, automation methods are beginning to have a broad impact on the structural NMR community.

20. 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 in developing automated NMR data analysis software. We would also like to acknowledge John Everett for his work on the cover art and figures. The authors' recent work on automated NMR data analysis has been supported by the NIH Protein Structure Initiative (P50-GM62413) and by a grant from Geneformatics, Inc.

21. References

(1) Moseley, H. N.; Montelione, G. T. *Curr. Opin. Struct. Biol.* **1999**, *9*, 635.
(2) Montelione, G. T.; Zheng, D.; Huang, Y. J.; Gunsalus, K. C.; Szyperski, T. *Nature Struct. Biol.* **2000**, *7*, 982.
(3) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley & Sons: New York, 1986.
(4) Clore, G. M.; Gronenborn, A. M. Annu. Rev. Biophys. Biophys. Chem. **1991**, *20*, 29.
(5) Ikura, M.; Kay, L. E.; Bax, A. *Biochemistry* **1990**, *29*, 4659.
(6) Montelione, G. T.; Wagner, G. *J. Magn. Reson.* **1990**, *83*, 183.
(7) Kay, L. E. *Curr. Opin. Struct. Biol.* **1995**, *5*, 674.
(8) Szyperski, T.; Wider, G.; Bushweller, J. H.; Wüthrich, K. *J. Am. Chem. Soc.* **1993**, *115*, 9307.
(9) Simorre, J. P.; Brutscher, B.; Caffrey, M. S.; Marion, D. *J. Biomol. NMR* **1994**, *4*, 325.
(10) Brutscher, B.; Simorre, J. P.; Caffrey, M. S.; Marion, D. *J. Magn. Reson.* **1994**, *105B*, 77.
(11) Szyperski, T.; Braun, D.; Banecki, B.; Wüthrich, K. *J. Am. Chem. Soc.* **1996**, *118*, 8146.
(12) Szyperski, T.; Yeh, D. C.; Sukumaran, D. K.; Moseley, H. N.; Montelione, G. T. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8009.
(13) Szymczyna, B. R.; Pineda-Lucena, A.; Mills, J. L.; Szyperski, T.; Arrowsmith, C. H. *J. Biomol. NMR* **2002**, *22*, 299.
(14) Ding, K.; Gronenborn, A. M. *J. Magn. Reson.* **2002**, *156*, 262.
(15) Kim, S.; Szyperski, T. *J. Am. Chem. Soc.* **2003**, *125*, 1385.
(16) Kim, S.; Szyperski, T. *J. Biomol. NMR* **2004**, *28*, 117.
(17) Schmieder, P.; Stern, A. S.; Wagner, G.; Hoch, J. C. *J. Biomol. NMR* **1994**, *4*, 483.
(18) Schmieder, P.; Stern, A. S.; Wagner, G.; Hoch, J. C. *J. Biomol. NMR* **1993**, *3*, 569.
(19) Kupce, E.; Freeman, R. *J. Magn. Reson.* **2003**, *162*, 300.
(20) Kupce, E.; Freeman, R. *J. Biomol. NMR* **2003**, *25*, 349.
(21) Kupce, E.; Freeman, R. *J. Biomol. NMR* **2003**, *27*, 383.
(22) Kupce, E.; Freeman, R. *J. Biomol. NMR* **2004**, *28*, 391.
(23) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9279.
(24) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111.
(25) Tjandra, N.; Omichinski, J. G.; Gronenborn, A. M.; Clore, G. M.; Bax, A. *Nature Struct. Biol.* **1997**, *4*, 732.
(26) Prestegard, J. H. *Nature Struct. Biol.* **1998**, *5*, 517.
(27) Cordier, F.; Grzesiek, S. *J. Am. Chem. Soc.* **1999**, *121*, 1601.
(28) Cornilescu, G.; Hu, J.; Bax, A. *J. Am. Chem. Soc.* **1999**, *121*, 2949.
(29) Cornilescu, G.; Kirsten, M.; Ramirez, B. E.; Frank, K.; Clore, G. M.; Gronenborn, A. M. *J. Am. Chem. Soc.* **1999**, *121*, 6275.
(30) Wang, Y. X.; Jacob, J.; Cordier, F.; Wingfield, P.; Stahl, S. J.; Lee-Huang, S.; Torchia, D.; Grzesiek, S.; Bax, A. *J. Biomol. NMR* **1999**, *14*, 181.
(31) Zimmerman, D. E.; Kulikowski, C. A.; Huang, Y.; Feng, W.; Tashiro, M.; Shimotakahara, S.; Chien, C.; Powers, R.; Montelione, G. T. *J. Mol. Biol.* **1997**, *269*, 592.
(32) Nilges, M. *J. Mol. Biol.* **1995**, *245*, 645.
(33) Nilges, M.; Macias, M. J.; O'Donoghue, S. I.; Oschkinat, H. *J. Mol. Biol.* **1997**, *269*, 408.
(34) Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M. *Bioinformatics* **2003**, *19*, 315.
(35) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319*, 209.
(36) Huang, Y. J. Rutgers University, New Brunswick, NJ, 2001.

(37) Huang, Y. J.; Swapna, G. V.; Rajan, P. K.; Ke, H.; Xia, B.; Shukla, K.; Inouye, M.; Montelione, G. T. *J. Mol. Biol.* **2003**, *327*, 521.

(38) Styles, P.; Soffe, N. F.; Scott, C. A.; Cragg, D. A.; White, D. J.; White, P. C. *J. Magn. Reson.* **1984**, 391.

(39) Marek, D. RF Receiver Coil Arrangement for NMR Spectrometers. U.S. Patent #5,247,256, 1993.

(40) Monleon, D.; Colson, K.; Moseley, H. N.; Anklin, C.; Oswald, R.; Szyperski, T.; Montelione, G. T. *J. Struct. Funct. Genomics* **2002**, *2*, 93.

(41) Zheng, D.; Huang, Y. J.; Moseley, H. N.; Xiao, R.; Aramini, J.; Swapna, G. V.; Montelione, G. T. *Protein Sci.* **2003**, *12*, 1232.

(42) Baran, M. C.; Aramini, J.; Huang, Q. H.; Xiao, R.; Acton, T. B.; Liang-yu, S.; Goldsmith-Fischman, S.; Liu, J.; Rost, B.; Honig, B.; Montelione, G. T. Submitted.

(43) Baran, M. C.; Moseley, H. N.; Sahota, G.; Montelione, G. T. *J. Biomol. NMR* **2002**, *24*, 113.

(44) 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.; Markley, J. L. *J. Struct. Funct. Genomics* **2003**, *4*, 11.

(45) Seavey, B. R.; Farr, E. A.; Westler, W. M.; Markley, J. L. *J. Biomol. NMR* **1991**, *1*, 217.

(46) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.

(47) Heinemann, U.; Frevert, J.; Hofmann, K.; Illing, G.; Maurer, C.; Oschkinat, H.; Saenger, W. *Prog. Biophys. Mol. Biol.* **2000**, *73*, 347.

(48) 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.; Kuramitsu, S. *Nature Struct. Biol.* **2000**, *7*, 943.

(49) Terwilliger, T. C. *Nature Struct. Biol.* **2000**, *7*, 935.

(50) 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.; Wang, L. K. *Protein Sci.* **2002**, *11*, 723.

(51) Kennedy, M. A.; Montelione, G. T.; Arrowsmith, C. H.; Markley, J. L. *J. Struct. Funct. Genomics* **2002**, *2*, 155.

(52) 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.; Su, X. D. *J. Struct. Funct. Genomics* **2003**, *4*, 137.

(53) Wishart, D. S.; Bigam, C. G.; Yao, J.; Abildgaard, F.; Dyson, H. J.; Oldfield, E.; Markley, J. L.; Sykes, B. D. *J. Biomol. NMR* **1995**, *6*, 135.

(54) Grzesiek, S.; Bax, A. *J. Biomol. NMR* **1993**, *3*, 185.

(55) Moseley, H. N.; Monleon, D.; Montelione, G. T. *Methods Enzymol.* **2001**, *339*, 91.

(56) Wishart, D. S.; Sykes, B. D. *J. Biomol. NMR* **1994**, *4*, 171.

(57) Montelione, G. T.; Rios, C. B.; Swapna, G. V.; Zimmerman, D. E. Biological Magnetic Resonance. In *NMR pulse sequences and computational approaches for automated analysis of sequence-specific backbone resonance assignments in proteins*; Berliner, Krishna, N. R., Eds.; 1999; Vol. 17, p 81.

(58) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277.

(59) Güntert, P.; Dotsch, V.; Wider, G.; Wüthrich, K. *J. Magn. Reson.* **1992**, *2*, 395.

(60) Eccles, C.; Güntert, P.; Billeter, M.; Wüthrich, K. *J. Biomol. NMR* **1991**, *1*, 111.

(61) Garrett, D. S.; Powers, R.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson.* **1991**, *95*, 214.

(62) Koradi, R.; Billeter, M.; Engeli, M.; Güntert, P.; Wüthrich, K. *J. Magn. Reson.* **1998**, *135*, 288.

(63) Goddard, T. D.; Kneller, D. G. *SPARKY 3*; University of California, San Francisco, CA: 2000.

(64) Orekhov, V. Y.; Ibraghimov, I. V.; Billeter, M. *J. Biomol. NMR* **2001**, *20*, 49.

(65) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **2002**, *24*, 171.

(66) Andrec, M.; Prestegard, J. H. *J. Magn. Reson.* **1998**, *130*, 217.

(67) Moseley, H. N.; Riaz, N.; Aramini, J.; Montelione, G. T. *J. Magn. Reson.*, in press.

(68) Zimmerman, D. E.; Montelione, G. T. *Curr. Opin. Struct. Biol.* **1995**, *5*, 664.

(69) Güntert, P. *Prog. Nucl. Magn. Reson. Spectrosc.* **2003**, *19*, 105.

(70) Billeter, M.; Braun, W.; Wüthrich, K. *J. Mol. Biol.* **1982**, *155*, 321.

(71) Wagner, G.; Wüthrich, K. *J. Mol. Biol.* **1982**, *155*, 347.

(72) Buchler, N. E.; Zuiderweg, E. R.; Wang, H.; Goldstein, R. A. *J. Magn. Reson.* **1997**, *125*, 34.

(73) Lukin, J. A.; Gove, A. P.; Talukdar, S. N.; Ho, C. *J. Biomol. NMR* **1997**, *9*, 151.

(74) Leutner, M.; Gschwind, R. M.; Liermann, J.; Schwarz, C.; Gemmecker, G.; Kessler, H. *J. Biomol. NMR* **1998**, *11*, 31.

(75) Bartels, C.; Billeter, M.; Güntert, P.; Wüthrich, K. *J. Biomol. NMR* **1996**, *7*, 207.

(76) Bartels, C.; Güntert, P.; Billeter, M.; Wüthrich, K. *J. Comput. Chem.* **1997**, *18*, 139.

(77) Atreya, H. S.; Sahu, S. C.; Chary, K. V.; Govil, G. *J. Biomol. NMR* **2000**, *17*, 125.

(78) Güntert, P.; Salzmann, M.; Braun, D.; Wüthrich, K. *J. Biomol. NMR* **2000**, *18*, 129.

(79) Andrec, M.; Levy, R. M. *J. Biomol. NMR* **2002**, *23*, 263.

(80) Coggins, B. E.; Zhou, P. *J. Biomol. NMR* **2003**, *26*, 93.

(81) Gronwald, W.; Willard, L.; Jellard, T.; Boyko, R. F.; Rajarathnam, K.; Wishart, D. S.; Sonnichsen, F. D.; Sykes, B. D. *J. Biomol. NMR* **1998**, *12*, 395.

(82) Zimmerman, D.; Kulikowski, C.; Wang, L.; Lyons, B.; Montelione, G. T. *J. Biomol. NMR* **1994**, *4*, 241.

(83) Li, K. B.; Sanctuary, B. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 467.

(84) Hyberts, S. G.; Wagner, G. *J. Biomol. NMR* **2003**, *26*, 335.

(85) Malmodin, D.; Papavoine, C. H.; Billeter, M. *J. Biomol. NMR* **2003**, *27*, 69.

(86) Montelione, G. T.; Lyons, B. A.; Emerson, S. D.; Tashiro, M. J. *J. Am. Chem. Soc.* **1992**, *114*, 10974

(87) Logan, T. M.; Olejniczak, E. T.; Xu, R. X.; Fesik, S. W. *FEBS Lett.* **1992**, *314*, 413.

(88) Grzesiek, S.; Anglister, J.; Bax, A. *J. Magn. Reson.* **1993**, *101*, 114.

(89) Celda, B.; Montelione, G. T. *J. Magn. Reson.* **1993**, *B101*, 189.

(90) Farmer, B. T.; Venters, R. A. *J. Am. Chem. Soc.* **1995**, *117*, 4187.

(91) Gschwind, R. M.; Gemmecker, G.; Kessler, H. *J. Biomol. NMR* **1998**, *11*, 191.

(92) Lin, Y.; Wagner, G. *J. Biomol. NMR* **1999**, *15*, 227.

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

(94) Kay, L. E.; Ikura, M.; Bax, A. *J. Am. Chem. Soc.* **1990**, *112*, 888.

(95) Bax, A.; Clore, G. M.; Gronenborn, A. M. *J. Magn. Reson.* **1990**, *88*, 425.

(96) Fesik, S. W.; Eaton, H. L.; Olejniczak, E. T.; Zuiderweg, E. R.; McIntosh, L. P.; Dahlquist, F. W. *J. Am. Chem. Soc.* **1990**, *112*, 886.

(97) Wishart, D. S.; Watson, M. S.; Boyko, R. F.; Sykes, B. D. *J. Biomol. NMR* **1997**, *10*, 329.

(98) Xu, X. P.; Case, D. A. *J. Biomol. NMR* **2001**, *21*, 321.

(99) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215.

(100) Zhang, H.; Neal, S.; Wishart, D. S. *J. Biomol. NMR* **2003**, *25*, 173.

(101) Moseley, H. N.; Sahota, G.; Montelione, G. T. *J. Biomol. NMR* **2004**, *28*, 341.

(102) Güntert, P.; Mumenthaler, C.; Wüthrich, K. *J. Mol. Biol.* **1997**, *273*, 283.

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

(104) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, M. G. *J. Magn. Reson.* **2003**, *160*, 65.

(105) Brünger, 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.; Warren, G. L. *Acta Crystallogr., D: Biol. Crystallogr.* **1998**, *54*, 905.

(106) Clore, G. M.; Gronenborn, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5891.

(107) Tejero, R.; Monleon, D.; Celda, B.; Powers, R.; Montelione, G. T. *J. Biomol. NMR* **1999**, *15*, 251.

(108) Gippert, G. P.; Wright, P. E.; Case, D. A. *J. Biomol. NMR* **1998**, *11*, 241.

(109) Cornilescu, G.; Delaglio, F.; Bax, A. *J. Biomol. NMR* **1999**, *13*, 289.

(110) Mumenthaler, C.; Braun, W. *J. Mol. Biol.* **1995**, *254*, 465.

(111) Mumenthaler, C.; Güntert, P.; Braun, W.; Wüthrich, K. *J. Biomol. NMR* **1997**, *10*, 351.

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

(113) Adler, M. *Proteins* **2000**, *39*, 385.

(114) Gronwald, W.; Moussa, S.; Elsner, R.; Jung, A.; Ganslmeier, B.; Trenner, J.; Kremer, W.; Neidig, K. P.; Kalbitzer, H. R. *J. Biomol. NMR* **2002**, *23*, 271.

(115) Grishaev, A.; Llinas, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6707.

(116) Xu, Y.; Wu, J.; Gorenstein, D.; Braun, W. *J. Magn. Reson.* **1999**, *136*, 76.

(117) Assuming backbone atoms are ~95% complete and in the complete resonance assignment list 40% are backbone atoms and the other 60% are side chain atoms.

(118) Jee, J.; Güntert, P. *J. Struct. Funct. Genomics* **2003**, *4*, 179.

(119) Ferentz, A. E.; Wagner, G. *Q. Rev. Biophys.* **2000**, *33*, 29.

(120) Huang, J. Y.; Tejero, R.; Powers, R.; Montelione, G. T. Submitted.

(121) Moy, F. J.; Seddon, A. P.; Bohlen, P.; Powers, R. *Biochemistry* **1996**, *35*, 13552.

(122) Greenfield, N. J.; Huang, Y. J.; Palm, T.; Swapna, G. V.; Monleon, D.; Montelione, G. T.; Hitchcock-DeGregori, S. E. *J. Mol. Biol.* **2001**, *312*, 833.

(123) Gonzalez, C.; Rullmann, J. A.; Bonvin, J. J.; Boelens, R.; Kaptein, R. *J. Magn. Reson.* **1991**, *91*, 659.
(124) Zhu, L.; Dyson, H. J.; Wright, P. E. *J. Biomol. NMR* **1998**, *11*, 17.
(125) Gronwald, W.; Kirchhofer, R.; Gorler, A.; Kremer, W.; Ganslmeier, B.; Neidig, K. P.; Kalbitzer, H. R. *J. Biomol. NMR* **2000**, *17*, 137.
(126) Borgias, B. A.; James, T. L. *J. Magn. Reson.* **1988**, *79*, 493.
(127) Borgias, B. A.; James, T. L. *Methods Enzymol.* **1989**, *176*, 169.
(128) Witten, I. H.; Frank, E. Data mining: practical machine learning tools and techniques with Java implementations; Morgan Kaufmann: San Francisco, CA, 2000.
(129) Hand, D. J.; Mannila, H.; Smyth, P. *Principles of data mining*; MIT Press: Cambridge, MA, 2001.
(130) Nabuurs, S. B.; Spronk, C. A.; Krieger, E.; Maassen, H.; Vriend, G.; Vuister, G. W. *J. Am. Chem. Soc.* **2003**, *125*, 12026.
(131) Laskowski, R. A.; Rullmann, J. A.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. *J. Biomol. NMR* **1996**, *8*, 477.
(132) Vriend, G. *WHAT IF: A molecular modeling and drug design program*; 1990.
(133) Bhattacharya, A.; Tejero, R.; Montelione, G. T. In preparation.
(134) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
(135) Bruccoleri, R. E. Department of Macromolecular Modeling, Bristol-Myers Squibb Pharmaceutical Research Institute, 1995.
(136) Bassolino-Klimas, D.; Tejero, R.; Krystek, S. R.; Metzler, W. J.; Montelione, G. T.; Bruccoleri, R. E. *Protein Sci.* **1996**, *5*, 593.
(137) Hyberts, S. G.; Goldberg, M. S.; Havel, T. F.; Wagner, G. *Protein Sci.* **1992**, *1*, 736.
(138) Eisenberg, D.; Luthy, R.; Bowie, J. U. *Methods Enzymol.* **1997**, *277,* 396.
(139) Westbrook, J.; Feng, Z.; Burkhardt, K.; Berman, H. M. *Methods Enzymol.* **2003**, *374*, 370.
(140) Word, J. M.; Bateman, R. C., Jr.; Presley, B. K.; Lovell, S. C.; Richardson, D. C. *Protein Sci*. **2000**, *9*, 2251.
(141) Gardner, K. H.; Rosen, M. K.; Kay, L. E. *Biochemistry* **1997**, *36*, 1389.
(142) Tian, F.; Valafar, H.; Prestegard, J. H. *J. Am. Chem. Soc.* **2001**, *123*, 11791.
(143) Valafar, H.; Prestegard, J. H. *J. Magn. Reson.* **2004**, *167*, 228.
(144) Andrec, M.; Du, P.; Levy, R. M. *J. Biomol. NMR* **2001**, *21*, 335.
(145) Andrec, M.; Jacobson, M. P.; Friesner, R.; Levy, R. M. *J. Struct. Funct. Genomics* **2002**, *2*, 103.
(146) Andrec, M.; Du, P.; Levy, R. M. *J. Am. Chem. Soc.* **2001**, *123*, 1222.
(147) Bayro, M. J.; Mukhopadhyay, J.; Swapna, G. V.; Huang, J. Y.; Ma, L. C.; Sineva, E.; Dawson, P. E.; Montelione, G. T.; Ebright, R. H. *J. Am. Chem. Soc.* **2003**, *125*, 12382.
(148) 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.; Laue, E. *Nat. Struct. Biol.* **2002**, *9*, 416.

CR030408P