February 2006

A topology-constrained distance network algorithm for protein structure determination from NOESY data

Yuanpeng Janet Huang
Rutgers University

Roberto Tejero
Rutgers University

Robert Powers
University of Nebraska - Lincoln, rpowers3@unl.edu

Gaetano T. Montelione
guy@cabm.rutgers.edu

Published in *Proteins: Structure, Function, and Bioinformatics* 62:3 (15 February 2006), pp. 587-603. Copyright © 2006 Wiley-Liss, Inc., A Wiley Company. Used by permission. Digital Object Identifier (DOI) 10.1002/prot.20820 http://www3.interscience.wiley.com/cgi-bin/jhome/36176

A topology-constrained distance network algorithm for protein structure determination from NOESY data

Yuanpeng Janet Huang [1], Roberto Tejero [1], Robert Powers [2], Gaetano T. Montelione [1,3]

[1]Center for Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry,
Rutgers University, Piscataway, New Jersey

[2]Department of Chemistry, University of Nebraska–Lincoln, Lincoln, Nebraska

[3]Department of Biochemistry, Robert Wood Johnson Medical School,
University of Medicine and Dentistry of New Jersey, Piscataway, New Jersey.

Corresponding author: Gaetano T. Montelione, e-mail: guy@cabm.rutgers.edu

Submitted May 2005; accepted July 2005; published online December 2005; in print February 2006

Abstract: This article formulates the multidimensional nuclear Overhauser effect spectroscopy (NOESY) interpretation problem using graph theory and presents a novel, bottom-up, topology-constrained distance network analysis algorithm for NOESY cross peak interpretation using assigned resonances. AutoStructure is a software suite that implements this topology-constrained distance network analysis algorithm and iteratively generates structures using the three-dimensional (3D) protein structure calculation programs XPLOR/CNS or DYANA. The minimum input for AutoStructure includes the amino acid sequence, a list of resonance assignments, and lists of 2D, 3D, and/or 4D-NOESY cross peaks. AutoStructure can also analyze homodimeric proteins when X-filtered NOESY experiments are available. The quality of input data and final 3D structures is evaluated using recall, precision, and F-measure (RPF) scores, a statistical measure of goodness of fit with the input data. AutoStructure has been tested on three protein NMR data sets for which high-quality structures have previously been solved by an expert, and yields comparable high-quality distance constraint lists and 3D protein structures in hours. We also compare several protein structures determined using AutoStructure with corresponding homologous proteins determined with other independent methods. The program has been used in more than two dozen protein structure determinations, several of which have already been published.

Keywords: AutoStructure, RPF, constraint network analysis, automated NMR structure determination, graph theory, fold topology constraints

Funded by: National Science Foundation; National Institutes of Health Grant Number: P50 GM62413; New Jersey Commission on Science and Technology Grant Number: 99-2042-007-13; DGS; Grant Number PB98-1455

INTRODUCTION

Protein NMR and X-ray crystallography are the two principal approaches for determining atomic resolution structures of macromolecules. Traditionally, NMR structure determination requires analysis of sequence-specific resonance assignments and interpretation of multidimensional nuclear Overhauser effect (NOESY) spectra using these resonance assignments. High-resolution three-dimensional (3D) structures are calculated based on distance constraints calibrated from interpreted NOESY cross peaks. Due to resonance degeneracy, manual interpretation of NOESY cross peaks is time-consuming and involves significant expertise. This manual analysis process is one of the significant barriers challenging the use of NMR as a routine tool for protein structure analysis in structural biology and in the emerging area of structural genomics. [1]

In this study we use graph theory to formulate the NOESY interpretation problem, and present a novel bottom-up, topology-constrained distance network algorithm for NOESY interpretation. AutoStructure is a software suite that implements this topology-constrained distance network analysis algorithm and automatically generates 3D protein structures using NOESY cross peaks, together with the structure calculation programs XPLOR/CNS [2] [3] or DYANA. [4]

Several fully automated approaches for NOESY interpretation and structure calculation have been developed, including NOAH, [5] [6] ARIA, [7] [8] CANDID, [9] a self-consist constraint analysis method implemented in XPLOR, [10] and other generally less developed programs. [11-14]The ARIA, CANDID, and NOAH programs utilize a *top-down* data interpretation approach, incorporating all of the data simultaneously and often incorporating an ambiguous constraint strategy [7] [8] to help in resolving information from potentially overlapped cross peaks in the NOESY spectrum. AutoStructure uses a topology-constrained *bottom-up* approach, first building structures based on intraresidue and sequential NOESY data, then local structures indicated by medium-range interactions, then β-strand topologies, and finally interpreting information arising from long-range packing interactions. This protocol, in principle, resembles the methodology that an expert would utilize in manually solving a protein structure by NMR.

In this article, we describe for the first time the underlying algorithms of AutoStructure and report the performance of the program on three real protein NMR data sets. These three real protein data sets used for developing and testing AutoStructure are from three different protein fold families (i.e., mainly β, mainly α, and α/β folds), and range in size from 113 to 169 amino acid residues. The NMR spectral data available for these three proteins vary with respect to completeness, resolution, degeneracy, and spurious peaks. With these data, AutoStructure provides high-quality automated interpretation of NOESY cross peaks and generates accurate 3D structures using XPLOR/CNS or DYANA. The AutoStructure program has been used in more than two dozen de novo protein structure determinations, several of which have already been published. [15-22] This program also plays a central role in the NMR structure analysis component of the Northeast Structural Genomics Consortium (http://www.nesg.org) and is an integral part of an evolving process for high-throughput protein NMR structure analysis. [23-25] Several of the structures determined with AutoStructure have subsequently been validated by independent structure determination of homologous protein structures using other methods. [26-28] These examples demonstrate the robustness and reliability of the program AutoStructure.It is also very critical for automated NOESY interpretation and structure determination approaches to use a fast and sensitive structure quality assessment measure to evaluate the quality of the generated structures, and to indicate the correctness of the fold and accuracy of the structure. Here, we use recall, precision, and F-measure (RPF) scores, [29] a statistical method from information retrieval, to evaluate the quality of a protein structure against the NOESY and resonance assignment data from which the structure is derived.

ALGORITHMS

Graph Theory Formulation of the NOESY Interpretation Problem

Given a model 3D protein structure, a complete distance network $G = (V, E)$ can be generated in which vertices (V) represent all protons of the model structure ($V = \{h \mid h$ is any proton from the model structure$\}$) and edges (E) connect the vertices and represent exact distance relations between proton pairs that are separated by at most $d_{\max}$ (Å) [$E = \{(h1, h2, d) \mid d < d_{\max}$ is the distance between nodes $h1$ and $h2\}$]. When $d_{\max}$ is large enough, this complete and exact proton-pair distance network can be used together with the known amino acid covalent geometry and chirality to generate an accurate structure model using projection methods of distance geometry. [30-32] The process of determining sequence-specific resonance assignments provide a set R of protein nuclei resonance frequencies assigned to specific atoms of the protein structure [$R = \{ \delta(h) \mid h$ is any atom of the protein and δ is its resonance chemical shift value$\}$]. NOESY cross peaks (p) of intensity I represent resonances ($\delta1, \delta2$) of proton pairs with close distance relationships [NOE $=\{p = (\delta1, \delta2, I) \mid \exists$ proton pairs $(h1, h2)$, $\delta(h1) = \delta1$ and $\delta(h2) = \delta2\}$]. In the simplest approximation, peak intensity I is correlated with the interproton distance $d(h1, h2)$ by $I \sim d^{-6}$ for $d(h1, h2) < d_{\mathrm{NOE}}$, where d_{NOE} is the maximum distance detected in the NOESY spectrum. If the NOESY cross peaks $p = (\delta1, \delta2, I)$ are each interpreted by resonance assignment to one or more proton pairs $(h1, h2)$, these interpreted NOESY cross peaks can then be converted into a *NOE-linked distance network* $G_{\mathrm{NOE}} = (V, E_{\mathrm{NOE}})$, of vertices V corresponding to hydrogen atoms and edges $E_{\mathrm{NOE}} = \{(h1, h2, p) \mid p = [\delta1, \delta2, I(d)], \delta(h1) = \delta1$ and $\delta(h2) = \delta2\}$. Edges of G_{NOE} represent NOE cross peaks arising from interactions between protons $h1$ and $h2$. Ideally, every proton pair that is separated by $d < d_{\mathrm{NOE}}$ will be linked by a pair of symmetric NOESY cross peaks. Once an extensive, self-consistent, though generally incomplete distance network G_{NOE} has been constructed, 3D protein structures can be generated using structure generation programs such as XPLOR/CNS or DYANA.

From sets R and NOE, an *ambiguous NOE network* $G_{\mathrm{A\text{-}NOE}} = (V, E_{\mathrm{ANOE}})$ is built, where edge $E_{\mathrm{ANOE}} = \{(h1, h2, p) \mid p = [\delta1, \delta2, I(d)], \mid \delta1 - \delta(h1)\mid < \Delta$ and $\mid \delta2 - \delta(h2)\mid < \Delta\}$, and Δ is a *match tolerance* for matching chemical shift values in the resonance assignment list, set R, with values in the NOESY peaks list(s), set NOE. In constructing G_{ANOE}, each NOESY cross peak p may be used to link more than one proton pair [i.e., frq(p) ≥1]. The true solution network G_{NOE} is a subgraph of this G_{ANOE}. Given complete sets R and NOE, for each NOESY cross peak p, at least one of its linked proton pairs belongs to G_{NOE}. Inter-residue contact maps derived from $G_{\mathrm{A\text{-}NOE}}$ are *potential contact maps*, whereas contact maps derived from G_{NOE} are *true contact maps*.

In this formulation, the process of NMR protein structure analysis from NOESY data reduces to the generation of an accurate self-consistent distance network G_{NOE} from the initial ambiguous NOE network G_{ANOE}. This task is complicated by chemical shift degeneracy, resulting in ambiguous interpretation of NOE cross peaks to multiple potential interacting proton pairs, by incompleteness in the resonance assignment list (set R), and by artifacts in the NOESY peak list (set NOE), which can result in erroneous NOESY cross peak interpretations. The problem of finding solution network G_{NOE} from G_{ANOE} is considered to be NP-complete in that enumeration of all possible configurations of assignments requires exponential time. [12] [33] Accordingly, for typical protein NMR data sets, most computational approaches attempt to construct an approximate heuristic solution HG_{NOE} to the true distance network G_{NOE}, and then attempt to refine this approximation to be as close to the true G_{NOE} as possible by various analysis methods.

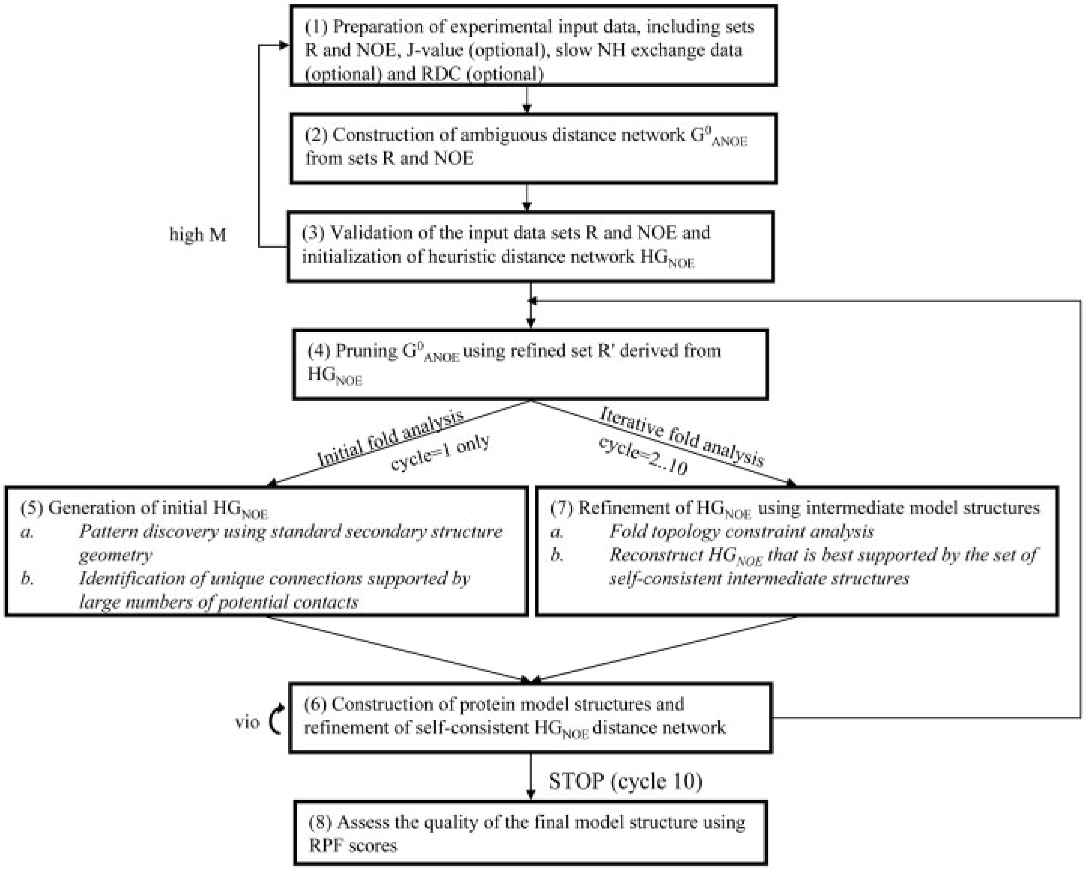
(1) Preparation of experimental input data, including sets R and NOE, J-value (optional), slow NH exchange data (optional) and RDC (optional)
(2) Construction of ambiguous distance network G0ANOE from sets R and NOE
high M
(3) Validation of the input data sets R and NOE and initialization of heuristic distance network HGNOE
(4) Pruning G0ANOE using refined set R' derived from HGNOE
Initial fold analysis
cycle=1 only
Iterative fold analysis
cycle=2..10
(5) Generation of initial HGNOE
a. Pattern discovery using standard secondary structure geometry
b. Identification of unique connections supported by large numbers of potential contacts
(7) Refinement of HGNOE using intermediate model structures
a. Fold topology constraint analysis
b. Reconstruct HGNOE that is best supported by the set of self-consistent intermediate structures
vio
(6) Construction of protein model structures and refinement of self-consistent HGNOE distance network
STOP (cycle 10)
(8) Assess the quality of the final model structure using RPF scores

Figure 1. Architecture of AutoStructure/AutoQF analysis processes. *Initial fold analysis* (cycle 1) includes steps 1-6, and *iterative fold analysis* (cycles 2–10) includes steps 4, 7 and 6. When an initial protein structure model is available as input (e.g., from homology modeling or manually analyzed NMR structure), step 7 is used in cycle 1. The initial ambiguous network G^0_{ANOE} of step 2 is reanalyzed in each iterative cycle. RPF scores [29] (step 8) assesses the quality of NMR structures.

AutoStructure's Heuristic Approach

AutoStructure uses a bottom-up topology-constrained distance network analysis algorithm to build an optimal solution distance network HG_{NOE}. The architecture of the overall iterative processes is shown in Figure 1. The AutoStructure protocol consists of two principal algorithms: *initial fold analysis* (cycle 1) that includes steps 1-6 in Figure 1, and *iterative structure analysis* (cycles 2-*N*) that includes steps 4, 7, and 6. Following several (typically ~10) cycles of iterative structure analysis, the program RPF (step 8 in Fig. 1) evaluates the quality of the resulting structure by measuring its goodness of fit to the NOESY peak list(s) and resonance assignment list. [29] In the following sections, each of the principal steps of Figure 1 is discussed. Table I summarizes definitions and typical values of key parameters described in each step of the algorithm.

Step 1. Preparation of experimental input data

AutoStructure uses the following input data: (1) protein amino acid sequence and a list of resonance assignments (set R); (2) "NOESY peak list" of the multidimensional (i.e., 2D, 3D, or 4D) NOESY cross peak frequencies (may be aliased) and intensities (set NOE); (3) a list of scalar coupling constant data (optional); (4) a list of amide sites exhibiting slow amide ^{1}H exchange (optional); and (5) other manually analyzed constraints when available, such as residual-dipolar-coupling (RDC), [34] disulfide-bond, and dihedral-angle constraint [35] data. NOESY peak lists are generated using third-party automatic spectrum peak-picking programs, usually followed by some manual editing. Prior to analysis, the quality of combined NOESY peak list and chemical shift list data are evaluated using the *M score* statistic for data quality, as described below. Dimeric proteins can also be analyzed when interchain NOESY cross peak data are available from X-filtered NOESY experiments. [17] [18]

Step 2. Construction of initial ambiguous network G^0_{ANOE} from input data sets R and NOE

The initial ambiguous distance network $G^{\,0}_{ANOE}$, with nodes corresponding to all hydrogen atoms and edges corresponding to all possible NOESY cross peak assignments, is generated by matching chemical shift values of NOESY cross peaks (set

TABLE I. Summary of Key Parameters Used in AutoStructure Analysis

Symbol	Definition	Default value
d_{NOE}	The maximum distance observed in NOESY spectrum	5.0 Å
$\Delta^{\mathrm{i}}_{\mathrm{err}}$	Error match tolerances for the *i*th dimension	H: 0.05 ppm ^{15}N/^{13}C: 0.5 ppm
$\Delta^{\mathrm{i}}_{\mathrm{allow}}$	Allowable match tolerances for the *i*th dimension	H: 0.04 ppm ^{15}N/^{13}C: 0.4 ppm
$\Delta^{\mathrm{i}}_{\mathrm{good}}$	Good match tolerances for the *i*th dimension	H: 0.03 ppm ^{15}N/^{13}C: 0.3 ppm
Δ_{sym}	Threshold for symmetric peaks	H: 0.03 ppm ^{15}N/^{13}C: 0.3 ppm
$dvio_{\mathrm{min}}$	Threshold of the minimum distance violation	0.1 Å
ms_{min}	Threshold of the minimum model-support score	0.4
ms_{high}	Threshold of the high model-support score	0.7

NOE) with chemical shift values in the resonance assignment list (set R) using match tolerances defined by: $\{(h1, h2, p) \mid p = (\delta 1, \delta 2, I), |\delta 1 - \delta(h1)| < \Delta\mathrm{err}^1 \text{ and } |\delta 2 - \delta(h2)| < \Delta\mathrm{err}^2\}$, where Δerr is defined in Table I. Because many sets of proton resonances of proteins are degenerate, multidimensional NMR uses the resonance frequencies of covalently bonded heavy atoms (i.e., ^{13}C or ^{15}N) as an additional filtering dimension to help resolve these proton degeneracies. For 3D and 4D NOE data sets, additional covalently bonded heavy atom dimensions are analyzed using a similar match-tolerance analysis (Table I). Spectral aliasing is handled internally by the program, as described in the Supplemental Material.

Step 3. Validation of the input data sets R and NOE, and initialization of heuristic distance network HG_{NOE}

Before proceeding with NOESY cross peak interpretation, AutoStructure first analyzes the quality, completeness, and self-consistency of the input resonance assignment list (set R) and NOESY peak list (set NOE). This is done using a distance network G_{local} of all conformation-independent two- and three-bond connected NOE-linked proton pairs predicted from set R. A data validation score, *M score*, is then calculated as

$$M = \frac{|\{(h1,h2,d)|(h1,h2,d) \in G_{\mathrm{local}},(h1,h2,d) \notin G^0_{\mathrm{ANOE}}\}|}{|\{(h1,h2,d)|(h1,h2,d) \in G_{\mathrm{local}}\}|} \quad (1)$$

This *M score* represents the fraction of NOE-linked proton pairs with short interproton distances that are in the predicted G_{local} network but missing from the $G_{\mathrm{ANOE}}{}^0$ network. The *M score* thus provides a measure of the qualities of sets R and NOE. Poor quality data sets that do not include most of these expected short-range NOESY cross peaks result in higher *M scores*; for example, a high *M score* (i.e., > 25%) suggests that at least one of the input data sets (R and/or NOE) is of inadequate quality and needs to be improved. Two- and three-bond connected NOE-linked proton pairs predicted from set R but not included in $G_{\mathrm{ANOE}}{}^0$ are reported to the user to further validate the corresponding chemical shift assignments, and/or identify 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 (set R). For each frequency dimension, the software computes the overall average chemical shift match difference from these NOE-linked proton pairs of the G_{local} distance network. Consistent spectral referencing is achieved using these differences as global reference correction factors for the target spectrum, providing a tighter match between R and NOE, and allowing the use of smaller matching tolerances (Δ) for further NOESY interpretation.

The heuristic HG_{NOE} distance network is initialized from G^0_{ANOE} using only NOE-linked proton pairs that are (1) well matched within tolerance Δ^i_{good} $[|\delta_i - \delta\ (\mathrm{h}_i)| \le \Delta^i_{\mathrm{good}}]$, and (2) connected by only two, three, or four covalent bonds, [36] 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. [37] These close proton pair connections are anticipated from the amino acid sequence of the protein. Generally, the match tolerance Δ^i_{good} is significantly smaller than the match tolerances Δ^i_{err} used to construct G^0_{ANOE} (Table I). 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.

Step 4. Pruning of G^0_{ANOE} using refined resonance assignment list R′ derived from HG_{NOE}

In addition to the global reference correction described above, AutoStructure attempts to correct for site-specific chemical shift differences between resonance assignment R and the NOESY peak list due to interspectral variations of temperature and sample conditions. In HG_{NOE}, if proton h_i is involved in at least three assigned NOE interactions (degree of vertex $h_i \geq 3$), its resonance frequency $\delta(h_i)$ is updated in a refined resonance assignment list R′ with the median value derived from these linked NOE cross peaks. Match tolerances (Δ^i_{err}) for those protons with refined chemical shifts are set to a narrower tolerance $\Delta^i_{\mathrm{allow}}$ (Table I) in G^0_{ANOE}, and linking edges with large mismatches $[|\delta_i - \delta(h_i)| > \Delta^i_{\mathrm{allow}}]$ resulting from these protons with updated chemical shift values are removed from G^0_{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. The resulting "pruned G^0_{ANOE}" is referred to as "G_{ANOE}" for the rest of the steps of the cycle. In each iterative cycle, the initial ambiguous network G^0_{ANOE} is pruned based on a refined chemical shift list, and a new G_{ANOE} is generated.

Step 5. Generation of initial HG$_{NOE}$ and initial fold analysis

Initial fold analysis (Fig. 1), a core process of AutoStructure, builds an initial HG_{NOE} distance network and uses this to construct a preliminary model of the protein structure. This step uses constraints indicated by identified secondary structure elements and long-range fold packing considerations to rule-in and rule-out NOESY cross peak assignments prior to the actual structure generation process.

a. Pattern discovery using standard secondary structure geometry. First (step 5a in Fig. 1), secondary structures (β-sheets and α-helices) are identified based on their characteristic NOE patterns, [38] together with $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift index (CSI) [39] and scalar coupling [38] data. Details of algorithms developed to discover β-sheets and α- or 3_{10}-helices using NOE contact patterns in G_{ANOE} are described elsewhere. [40] 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 structures identified in the NMR data. Therefore, the HG_{NOE} network is expanded by analysis of the G_{ANOE} network. At the same time, many possible NOE-linked proton pairs that 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}.

b. Identification of unique connections supported by large numbers of potential interresidue contacts. Next (Step 5b in Fig. 1), a well-matched NOE-linked proton pair ($h1$, $h2$, p) [$|\delta_i$ - $\delta(h_i)| < \Delta_{\mathrm{good}}{}^{\mathrm{i}}$] 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 ($h2$, $h1$, $p2$), where [$|\delta_i - \delta(h_i)| < \Delta^i_{\mathrm{good}}$, $|\ \delta_1(p1) - \delta_2(p2)| < \Delta_{\mathrm{sym}}$, and $|\ \delta_1(p2) - \delta_2(p1)| < \Delta_{\mathrm{sym}}$] 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. This symmetry filter exploits the fact that 4D NOESY information is encoded in pairs of symmetry peaks in 3D NOESY spectra, and uses the symmetry features of NOESY data to confirm potential assignments of 2D, 3D, or 4D NOESY cross peaks.

In addition, at this point, *potential contact support scores* pct($r1$, $r2$) for each possible spatial contact between residue pairs indicated by the G_{ANOE} network are used to provide an assessment of the confidence in the proposed contact. These algorithms are described in detail in the Supplementary Material. The effect of this analysis is to utilize a *potential interresidue contact map* derived from G_{ANOE} to filter out weakly supported (but "apparently unique") NOE-linked proton pairs from the initial fold analysis. This conservative process aims to avoid using incorrect NOESY cross peak assignments in generating the initial set of 3D structures that will be used in later stages to rule-in/rule-out other assignments. These weakly supported "apparently unique" NOE-linked proton pairs may be added into HG_{NOE} during the subsequent *iterative fold analysis* stage (Fig. 1; cycles 2-10), described below, if they are also well-supported by the tertiary conformations of intermediate structures. The "potential contact support analysis" of AutoStructure is conceptually similar to the process of "network anchoring" used by the program CANDID, [9] although somewhat more sophisticated in that it uses knowledge of expected short distances within and between secondary structure elements. Details of how secondary structure and intersecondary structure packing information are used in this analysis are presented in the Supplementary Material.

Step 6. Construction of protein model structures and refinement of self-consistent HG$_{NOE}$ distance network

Distance constraints are directly generated from HG_{NOE} by calibrating the peak's intensities assuming a simple two-spin approximation, and binned into upper-bound distance classes as described by Wüthrich and coworkers. [6] [36] [38] Dihedral angle constraints are generated from local NOE and scalar coupling data using the conformational grid search program HYPER. [41] Hydrogen bond distance constraints are identified based on analysis of amide hydrogen exchange data together with observations of helix and β-sheet NOE contact patterns, and/or tertiary 3D structure s. [38] Detailed descriptions of criteria used for identifying hydrogen bond constraints are presented in Supplementary Material. Potential *cis*-peptide bonds {i.e., H^{α} - H^{α} (i, i + 1) $\in HG_{\mathrm{NOE},}$ and H^{α} - H^{N}(i, i + 1) $\notin HG_{\mathrm{NOE}}$ or H^{α} - H^{δ} [i, Pro(i + 1)] $\notin HG_{\mathrm{NOE}}$} and disulfide bonds {i.e., H^{β} - H^{β} [Cys(i), Cys(j)] $\in HG_{\mathrm{NOE}}$} are identified and reported to the user for expert analysis and validation. *Cis* and *trans* X-Pro peptide bonds and disulfide bonds can be characterized by specific NOEs. [42] Proline $^{13}C^{\delta}$ chemical shifts can also be used to distinguish *cis* from *trans* X-Pro peptide bonds, and $^{13}C^{\beta}$ chemical shifts can be used to distinguish reduced and oxidized Cys residues. 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 (described in the Supplementary Material), although the program can also be run on a single processor system, such as a Linux-based laptop computer.

A set of N model structures that best satisfy the resulting constraints is next used to evaluate and refine the self-consistency of HG_{NOE}. First, distances between all NOE-linked proton pairs of HG_{NOE} are calculated. For interactions involving two or more degenerate proton resonances (e.g., methyl protons, symmetric aromatic protons, and degenerate methylene protons), r^{-6} summation [41] is used. Proton pairs with internuclear distances that violate the corresponding constraints by greater than $dvio_{\text{min}}$ (Table I) in all of these N initial structures are then 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 thus considered to be self-consistent, completing cycle 1 of the AutoStructure analysis. The aim of this process is to generate a set of initial 3D structures which can be reliably used for further interpretation of HG_{NOE} from G_{ANOE}. The resulting self-consistent HG_{NOE} and initial 3D structures can then be used for next stage of AutoStructure, *iterative fold analysis*.

Step 7. Iterative fold analysis and refinement of HG$_{NOE}$ using intermediate model structures and topology constraints

This step of AutoStructure (step 7 in Fig. 1) utilizes intermediate 3D structures as templates to refine and expand HG_{NOE}.

a. Fold topology constraint analysis. First, AutoStructure analyzes the topology of model structures and trims G_{ANOE} down based on *topology constraints* implied by helical-packing and β-sheet packing geometries. Globular protein molecules are formed by the close packing of α-helix and/or β-sheet secondary structure elements. [43] The amino acids in these secondary structure elements have relatively fixed conformations restricted by the constraints of the main-chain hydrogen bonds. The packing of these segments of secondary structure is also geometrically and energetically restricted. [43–47] At this stage of the AutoStructure process, patterns of residue-to-residue contacts, based on constraints implied by the packing of secondary structure elements [40] [43] [45–47] are used to further expand HG_{NOE} from G_{ANOE}. This process utilizing *protein-structure-based topology constraints* represents a higher order topology-constraint analysis similar to that used in considering canonical secondary structures for NOE interpretation in the *initial fold analysis* process.

b. Reconstruct HG$_{NOE}$ distance network best supported by the set of self-consistent intermediate structures. Next, HG_{NOE} is further expanded by adding NOE-linked proton pairs from G_{ANOE} that are well supported by the intermediate 3D structures. From the N self-consistent intermediate structures available at the end of cycle 1, distances $d(h1, h2)$ for all potential NOE-linked proton pairs $(h1, h2)$ are calculated, as described above using r^{-6} summation as appropriate. These distances are then used to generate N graphs, $G_{\text{min}}(i)$, in which edges correspond to the shortest among these potential NOE-linked proton pairs or $d(h1, h2) < 3$ Å. For each potential NOE-linked proton pair $(h1, h2, p)$ in G_{ANOE}, a model-support score ms$(h1, h2, p)$ (Table I) is then computed based on the frequency of observing the corresponding short distance in these G_{min} graphs:

$$\begin{cases} ms(h1,h2,p) & = & \frac{1}{n}\sum_{i=1}^{n} min(h1,h2,p)_i \\ min(h1,h2,p)_i & = & 1, \text{ if distance } d(h1,h2)_i \text{ is the shortest among all} \\ & & \quad \text{potential (ambiguous) proton} \\ & & \quad \text{pairs linked by peak } p \text{ in model } i \\ min(h1,h2,p)_i & = & 1, \text{ if } d(h1,h2)_i < 3\text{Å} \\ min(h1,h2,p)_i & = & 0, \text{ otherwise.} \end{cases} \tag{2}$$

Proton pairs whose average (i.e., median value) distance is $< d_{\text{NOE}}$, $ms(h1, h2, p) > ms_{\text{high}}$ and $|\delta_i - \delta(h_i)| < \Delta_{\text{allow}}{}^{\text{i}}$ (as defined in Table I) are considered to be well-supported by these intermediate structures, and added into HG_{NOE}. Well-supported NOE-linked proton pairs $(h1, h2, p)$, whose largest distance in *all N* intermediate structures is $< d_{\text{NOE}}$ and model support score $ms(h1, h2, p) \geq ms_{\text{min}}$, are also added into HG_{NOE}. This process ensures that only the most well-defined interproton distances are considered when using intermediate structures to expand HG_{NOE}. The resulting constraints are used to generate an ensemble of 3D model structures, refined by removing consistently violated constraints (as described for step 6 above), and the process of structure generation and consistent constraint analysis iterated to provide a self-consistent HG_{NOE} and structure ensemble, representing the results of cycle 2 (Fig. 1). These N structures from this ensemble that best fit the data are then used to further expand HG_{NOE} using the model support scores, as described above for well-ordered regions and below for less-well-ordered regions. This process is iterated for several (typically eight more) cycles of HG_{NOE} refinement and structure calculations (Fig. 1). In order to make smooth changes between these iterations of HG_{NOE}, old links from the HG_{NOE} of the previous cycle are retained only if $ms \geq ms_{min}$, while the remaining links of the HG_{NOE} from the previous cycle are removed. In this way, the intermediate structure is used to confirm, and in some cases correct, NOESY cross peak interpretations made in earlier stages of analysis.

In less well-defined or loosely packed regions of intermediate structures, we generally observe few proton pairs with consistent close distances < d_{NOE}. To improve the structure in these regions of intermediate structures, it is important to have means of ruling in correct NOESY cross peak assignments even for interactions separated by distances > d_{NOE} in the intermediate models. In the intermediate cycles of AutoStructure analysis (typically cycles 2-7), AutoStructure allows proton pairs with distances > d_{NOE} that are consistent with the fold (ms = 1) to be added into HG_{NOE}, but only under certain stringent conditions. The symmetry subgraph of G_{ANOE} described above is also pruned using intermediate structures. Well-matched proton pairs that were not ruled-in by the peak symmetry rules of *initial fold analysis* can also be added into the HG_{NOE} network in the *iterative fold analysis* if certain stringent criteria are met. For example, for each ambiguous NOE cross peak in the symmetry subgraph, the edges that have shortest distances in the intermediate structures ($h1$, $h2$, $p1$) are identified and added into HG_{NOE} only if there is another nonsymmetric linkage ($h3$, $h4$, $p2$) in G_{ANOE} for the same residue contact pair.

Intermediate structures are also used to refine inaccuracies in assignments arising from *orphan interresidue contacts* in HG_{NOE}. After four cycles of *iterative fold analysis*, suspect long-range *orphan contacts* between residue pairs ($r1$, $r2$) are identified from a contact map generated from HG_{NOE}. Although these *orphan contacts* may be strongly supported by contacts observed in the *potential contact map* generated for G_{ANOE}, unless they are well-supported from a contact map generated from HG_{NOE} (i.e., by the intermediate structures), they are considered at this stage to be unreliable. Suspect long-range *orphan contacts* are defined as those between residue pairs ($r1$, $r2$) that have no neighboring residue pairs ($r1 \pm 0 \ldots 5$, $r2 \pm 0 \ldots 5$) in the contact map of HG_{NOE} and for which no neighboring residue has long range contact with any other residue in the contact map of HG_{NOE}. Based on these criteria, potentially incorrect proton pair interactions associated with these orphan contacts are eliminated from HG_{NOE}, and thus from the derived constraint list. This criterion is generally satisfied only for incorrect long-range contacts in poorly defined loop regions.

Step 8. Assessment of the quality of the final 3D structure

Having completed the 3D protein structure determination, we use RPF scores to evaluate the quality of the resulting ensemble of structures given the experimental NMR data. [29] The RPF scores includes (1) recall score, which measures the fraction of NOE cross peaks that are consistent with the resulting structures, (2) precision score, which measures the fraction of back-calculated close proton-pair interactions from the resulting structure that are also observed in the peak list, and (3) F-measure score, which provides an assessment of the overall fit between the resulting structures and the experimental data, assuming that the input data set are near complete. Also reported is a normalized F-measure score, the discriminating power (DP) score, which measures how the query structure is distinguished from a freely-rotating chain model. [29]

MATERIALS AND METHODS

Experimental NMR Data Sets Used for the Validation of Automated Structure Determination With AutoStructure

The manual solution structures and NMR assignments for FGF-2, [48] [49] MMP-1, [50] [51] and IL-13 [52] have been described in detail previously. Most data were collected on 600 MHz NMR systems. Briefly, the assignments of the ^{1}H, ^{15}N, and ^{13}C resonances were typically based on the following 3D NMR experiments: CBCA(CO)NH, CBCANH, C(CO)NH, HC(CO)NH, HBHA(CO)NH, HNCO, HCACO, HNHA, HNCA, HCCH-COSY, and HCCH-TOCSY. [53] [54] The accuracy of the NMR assignments was further confirmed by sequential NOEs in the ^{15}N-edited NOESY-HMQC spectra. In some cases, stereospecific assignments were obtained for many β-methylene protons, and methyl groups of Val and Leu residues. These solution structures were based primarily on the experimental distance and torsion angle restraints determined from the following series of spectra: HNHA, HNHB, HACAHB-COSY, 3D ^{15}N- and ^{13}C-edited NOESY.

The manual FGF-2 structure was calculated on the basis of 2865 experimental NMR restraints consisting of 2486 approximate interproton distance restraints, 50 distance restraints for 25 backbone hydrogen bonds, and 329 torsion angle restraints consisting of 118 φ, 99 ψ, 84 χ_1, and 28 χ_2 torsion angle restraints. The manual MMP-1 structures were calculated on the basis of 3333 experimental NMR restraints consisting of 2493 approximate interproton distance restraints, 84 distance restraints for 42 backbone hydrogen bonds, 426 torsion angle restraints comprised of 155 φ, 134 ψ, 103 χ_1, and 34 χ_2 torsion angle restraints, 125 $^3J(H^N - H^\alpha)$ restraints, and 153 Cα and 136 Cβ chemical shift restraints. The manual IL-13 structures were calculated on the basis of 2848 experimental NMR restraints consisting of 2248 approximate interproton distance restraints, 100 distance restraints for 50 backbone hydrogen bonds, 299 torsion angle restraints comprised of 104 φ, 105 ψ, 66 χ_1, and 24 χ_2 torsion angle restraints, 96 $^3J(H^N - H^\alpha)$ restraints, and 104 C^α and 101 C^β chemical shift restraints. These three structures were calculated using the hybrid distance geometry-dynamical simulated annealing method of Nilges et al., [55] with minor modifications [56] using the program XPLOR. [3] For the MMP-1 and IL-13 manual structures, the method was adapted to incorporate pseudopotentials for $^3J(H^N - H^\alpha)$ coupling constants, [57] secondary ^{13} $C^\alpha/^{13}C^\beta$ chemical shift restraints, [58] and a conformational database potential. [59] [60] Additionally, for the IL-13 manual structure, a pseudopotential for the radius of gyration [61] was incorporated into the structure calculations.

TABLE II. Summary of Experimental NMR Data Sets Used for Validation of AutoStructure

Protein	FGF-2	MMP-1	IL-13
Fold class	β	α/β	α
Size (residues)	154	169	113
Assigned chemical shifts[a] (set R)			
All (% completeness)	94.7	90.5	93.7
Side-chain atoms (% completeness)	91.5	85.4	90.3
Aromatic atoms (% completeness)	93.8	83.6	92.7
Completeness of stereospecific isopropyl methyl proton assignments	16/84 (19.1%)	6/80 (7.5%)	56/88 (63.6%)
Completeness of stereospecific β methylene proton assignments	156/444 (35.1%)	86/408 (21.1%)	80/284 (28.2%)
NOESY spectra (set NOE)			
Peaks picked from 3D(^{15}N) spectrum	1353	2409	1994
Assignable[b]	1255	2003	1509
Peaks picked from 3D(^{13}C) spectrum	3702	3043	4143
Assignable[b]	3588	2970	3943
Input validation—M score			
3D(^{15}N) spectrum (%)	12	4	7
3D(^{13}C) spectrum (%)	15	24	6
Overall (%)	14	21	6

[a] Percentage of the total number of nonlabile protons and their attached C/N atoms observed in routine NMR studies that were assigned.

[b] Number of peaks (p) with frq(p) ≥ 1 in G^0_{ANOE}.

RESULTS

Analysis of FGF-2, MMP-1, and IL-13 Using AutoStructure

AutoStructure was developed and tested using experimental input data sets for three distinctly different proteins: human basic fibroblast growth factor (FGF-2), [48] [49] the inhibitor-free catalytic fragment of human fibroblast collagenase (MMP-1), [50] [51] and human interleukin-13 (IL-13). [52] The completeness of resonance assignments for these proteins are between 90% and 95%, and the input *M scores* of overall data quality range from 0.06 to 0.21 (Table II). For each AutoStructure calculation, the raw uninterpreted NOESY peak lists were used, and 10 iterative cycles of AutoStructure were performed. Table III (A) provides a summary of results for these 10-cycle AutoStructure/XPLOR calculations. Between 83% and 86% of all peaks in these NOESY peak lists were assigned by AutoStructure. F-measure scores for the three data sets range from 89% to 93%, and the DPs range from 79% to 85%, indicating good agreement between these final structures and the input NMR data. For all proteins, low root-mean-square deviations (RMSDs) across the final structures were obtained, which, by conventional criteria, are indicative of high-precision structure determinations.

The evolution of various quality parameters determined during the course of 10 AutoStructure cycles is illustrated in Figure 2 for the three proteins tested. In cycle 1, > 40% of NOESY cross peaks were assigned without using any intermediate 3D model structures to guide the assignment process. At this point in the analysis, 3–7% of the resulting NOE distance constraints are long-range constraints. The initial folds for the three data sets, prior to using the initial 3D structure for further NOESY cross peak interpretation, have overall fitness F-measure scores > 80%, and discriminating power DP score > 40% (Fig. 2). Stereoimages showing convergence of these initial folds are presented in Figure 3(a). By the end of cycle 2, and using initial structures to guide additional assignments, more than 65% of NOESY peaks were assigned, the overall fitness F scores are > 85%, and the DP scores are > 60% (Fig. 2). At this point, the structures are reasonably well converged with RMSDs within each ensemble of < 3 Å for all heavy atoms in secondary structure regions [Figs. 2 (b) and 3b)]. During cycles 3 through 10, additional peaks were interpreted based on the intermediate structures, resulting in a monotonic decrease in RMSD (Fig. 2). The rates of change of the F-measure and DP scores are slower after cycle 2, as the resulting refinement involves only small changes in the protein structures. Figure 3 (c) shows the well-converged structures from the final cycle. All three of the resulting ensembles have good quality assessment scores, with F-measure scores ranging from 89% to 93% and DP scores > 75%.

Comparison of FGF-2, MMP-1, and IL-13 Structures Analyzed by AutoStructure and Other Methods

FGF-2 contains 11 antiparallel β-strands comprising three β-sheets (Fig. 4, top). [49] AutoStructure identified 10 of the 11 β-strands and the proper alignments of the β-sheets during cycle 1, initially missing the small three-residue β-strand (strand 10) and three small β-sheet alignments involving two or three interstrand hydrogen bonds. However, all of the structural features of FGF-2 were accurately characterized by the end of the iterative analysis process. The mean coordinate differences between the final AutoStructure analysis [FGF-2a, Fig. 4 (a), top] and the published manual analysis [FGF-2b, Fig. 4 (b), top] are 0.5 Å for backbone atoms and 0.7 Å for all heavy atoms of secondary structure elements [Table III (C)]. As summarized in Table III, the NOE distance constraints obtained with AutoStructure are overall in good agreement with the structures determined by manual analysis. About 4% of distance constraints interpreted by AutoStructure have violations > 1.0 Å compared to structures determined by manual

TABLE III. Comparison of 3D Structures Calculated by AutoStructure With XPLOR Versus Manual Analysis, Using the Same Resonance Assignment List (R) and NOESY Peak Lists (NOE)

Quality	FGF-2	MMP-1	IL13
A. AutoStructure models			
NOE crosspeaks assigned			
^{15}N-NOESY	1147	1613	1214
^{13}C-NOESY	3160	2522	3254
Constraints			
NOE distance constraints[a]	2058	2088	2705
Hydrogen bond distance constraints	76	86	76
Dihedral angle constraints	177	221	165
RPF scores[b]	94.2/92.1/93.2/85.4	88.6/89.2/88.9/79.5	87.3/96.9/91.9/79.8
RMSD[c]			
Sec. Strut. Region[d] (b.b./heavy)	0.3/0.9	0.5/1.0	0.2/0.7
Ordered region[e] (b.b./heavy)	0.7/1.3	1.3/1.8	0.6/1.1
Ramachandran plot summary from PROCHECK[f]	75.0/23.9/1.1/0.0	81.7/18.0/0.3/0.0	90.4/9.3/0.1/0.1
B. Manual analysis models			
Constraints			
NOE distance constraints[a,g]	2105	2078	1979
Hydrogen bond distance constraints	50	84	50
Dihedral angle constraints	330	426	302
RPF scores[b]	94.5/92.6/93.5/86.2	88.9/89.6/89.2/80.7	82.5/97.1/89.2/72.3
RMSD[c]			
Sec. Strut. region[d] (b.b./heavy)	0.3/0.7	0.3/0.6	0.2/0.7
Ordered region[e] (b.b./heavy)	0.4/0.8	0.4/0.8	0.4/0.8
Ramachandran plot summary from PROCHECK[f]	77.5/21.5/1.0/0.0	90.1/9.8/0.31/0.0	91.1/8.0/0.9/0.0
C. Comparisons between AutoStructure and manual analysis			
% Distance constraints[h] of AutoStructure that are violated by > 1.0 Å in manual analysis models	3.3	3.7	5.8
% Distance constraints[g,h] of manual analysis that are violated by > 1.0 Å in AutoStructure analysis models	4.2	5.1	3.9
Mean Coordinate Differences[c]			
Sec. Strut. Region[d] (b.b./heavy)	0.5/0.6	0.6/0.8	0.8/0.9
Ordered region[e] (b.b./heavy)	0.8/1.0	1.2/1.5	1.4/1.6

[a] Reported are the total numbers of conformational-restricting constraints.

[b] These scores (%) include Recall/Precision/F-measure/DP scores.

[c] Throughout the article, RMSDs across conformation ensembles are computed relative to the mathematical-average coordinates. When comparing ensembles of structures, or ensembles of structures to a single structure, the RMSD between the mathematical-average coordinates of each ensemble is reported as the mean coordinate difference.

[d] Residues included in FGF-2 secondary structures are 31–34, 39–43, 49–53, 62–67, 71–76, 81–85, 91–94, 103–107, 113–116, 124–126, and 148–151.[49] Residues included in MMP-1 secondary structures are 13–19, 27–43, 48–53, 59–65, 82–85, 94–99, 112–124, and 150–160.[51] Residues included in IL-13 secondary structures are 6–22, 33–35, 43–52, 59–70, 89–91, and 92–108.[52]

[e] Disordered regions are excluded for RMSD calculations. For FGF-2, residues from 29 to 152 are used for calculation.[49] For MMP-1, residues from 7 to 137 and 145 to 163 are used for calculation.[51] For IL-13, all residues are used for calculation.[52] Sec. struct = secondary structure; b.b. = backbone atoms; heavy = heavy atoms.

[f] These scores are percentage of residues in most favored regions/additionally allowed regions/generously allowed regions/disallowed regions.

[g] Reported is the average number of distance constraints violated by > 1.0 Å. Distance constraints include both NOE and H-bond distance constraints.

[h] Center-averaging pseudoatom corrections are added to distance constraints for slowly rotating aromatic side-chain atoms.

analysis and vice versa. Most of these differences are in regions of the structure that are not well defined and do not have a significant impact on the 3D structure. However, more than twice the number of (manually defined) dihedral angle constraints were used in the manual analysis. Accordingly, residues in loop regions are somewhat better defined in the structures determined by manual analysis than by AutoStructure analysis. Comparison of the 1.9-Å resolution X-ray crystal structure of FGF-2 [FGF-2c, Fig. 4 (c), top] [62] with both the AutoStructure and manually determined NMR structures, respectively [Fig. 4(a and b), top], indicate that the NMR structures determined manually and by AutoStructure are about equally similar to the X-ray crystal structure, with backbone mean coordinate differences to this crystal structure of 0.5 ± 0.1 Å [Fig. 4(d), top]. Both the RPF scores and Ramachandran plot analysis from Procheck [63] summarized in Table III indicate that FGF-2 structures determined by the careful manual analysis fit the NOESY data slightly better and have slightly better stereochemical qualities compared to the structure automatically generated by AutoStructure.

The MMP-1 protein structure has three helices and eight β-strands as identified by manual anaysis. [51] AutoStructure

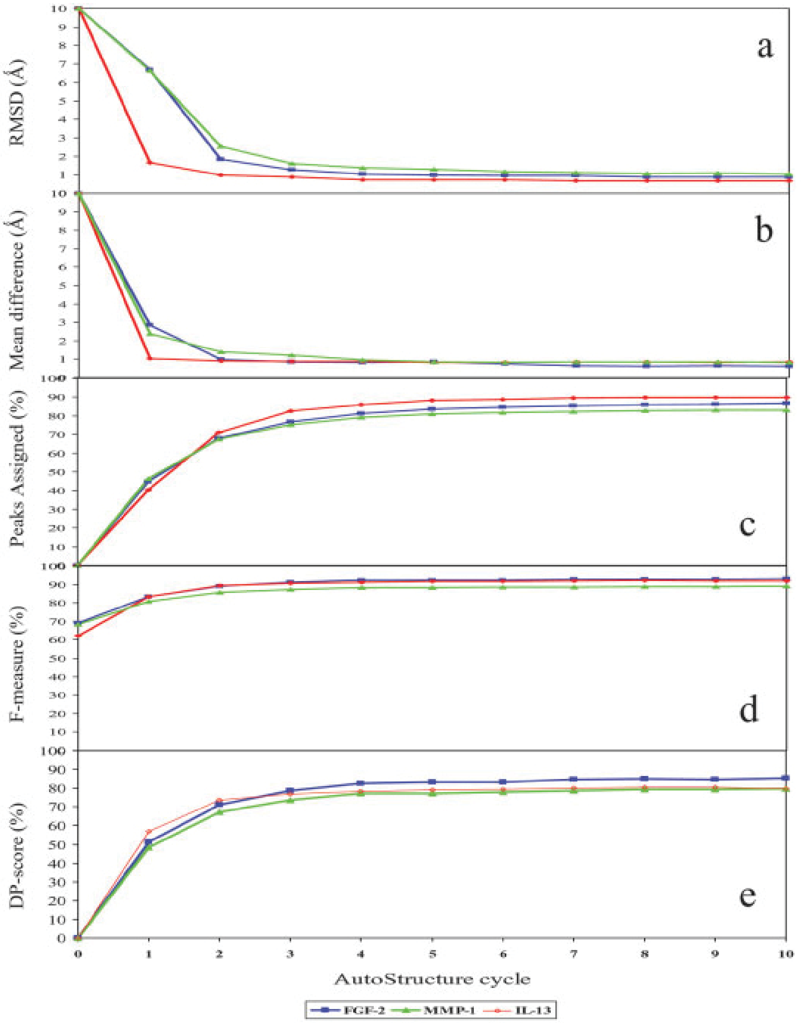

Figure 2. Evolution of characteristic parameters for NMR structures in the course of the 10 cycles of structure calculation for FGF-2 (blue), MMP-1 (green), and IL-13 (red). (**a**) RMSD for heavy atoms in the secondary structure. (**b**) Mean difference from manually determined NMR structures for heavy atoms in the secondary structure. (**c**) Percentage of peaks assigned. (**d**) F-measure score. (**e**) DP score. Cycle 0 is the conformational state before starting AutoStructure analysis, corresponding to an ensemble of random-coil chains whose RMSD within the ensemble is 10 Å and mean coordinate difference to the final structure is 10 Å.

correctly identified all secondary structure elements of MMP-1 during cycle 1. The mean backbone coordinate differences for residues in secondary structures between the final cycle of AutoStructure analysis [MMP-1a, Fig. 4(a), center] and the manual analysis [MMP-1b, Fig. 4(b), center] is < 1.0 Å [Table III(C)]. MMP-1 protein has one calcium and two zinc binding sites, and a nearby ligand-binding site [MMP-1c, Fig. 4(c), center]. [51] [64] These regions are not well defined in the structures determined by AutoStructure [Fig. 3(c), center], due largely to the exclusion of these calcium and zinc ions in the automated structure calculation process. Manual structure analysis also observed only sequential or short-range NOEs for these residues; the addition of zinc and calcium ions and the associated distance constraints are important in establishing the proper local structures with a resulting lower RMSD values for these regions during the manual structure calculation [Fig. 3(d), center]. [51] Dynamic studies of inhibitor-free MMP-1 shows that the loop region of residues 138-144 is mobile, with dynamic order parameters $S^2 < 0.6$. [50] This is consistent with the structure determined by AutoStructure [Fig. 3(c). center, indicated by an arrow]. About 4% of constraints determined by AutoStructure have violations > 1.0 Å compared to structures determined by manual analysis, though most of these differences are in the regions of the structure that are not well defined. A slightly higher number of the manually derived constraints (~5%) are violated by > 1.0 Å in the structures determined by AutoStructure, again mostly in not well-defined regions. As for FGF-2, more than twice the number of dihedral angle constraints were used in the manual analysis. Residues in loop regions are also better defined in

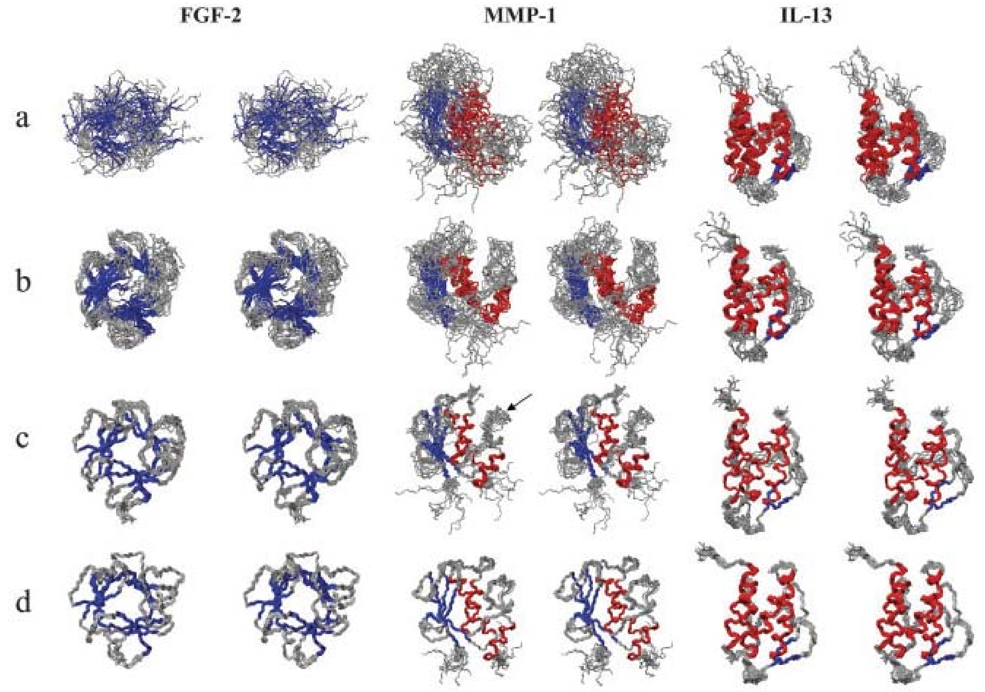

Figure 3. Structures of FGF-2, MMP-1 and IL-13 proteins generated by the AutoStructure/XPLOR process. (**a**) AutoStructure/XPLOR cycle 1, (**b**) cycle 2, and (**c**) cycle 10. (**d**) Manually analyzed NMR structures available from the PDB. β-strands are indicated in blue, helices in red.

the structures determined by manual analysis than by AutoStructure analysis [Fig. 3(d), center]. Comparison of the 1.56-Å resolution X-ray structure of MMP-1 complexed with a hydroxamate inhibitor [Fig. 4(c), center] [64] with both the AutoStructure and manually determined NMR structures MMP-1, respectively [Fig. 4(a and b), center], indicate that the NMR structures determined manually and by AutoStructure are about equally similar to the X-ray crystal structure, with backbone mean coordinate differences to this crystal structure of 0.5 ± 0.1 Å [Fig. 4 (d), center]. As for FGF-2, the RPF scores and Procheck Ramachandran analysis summarized in Table III indicate that MMP-1 structures determined by careful manual analysis fit the data slightly better and have better stereochemical qualities compared to the structures generated automatically by AutoStructure.

The IL-13 protein structure contains four α-helices and two β-strands. [52] AutoStructure correctly identified all secondary structure elements during cycle 1. The two disulfide bonds were identified during the course of the structure analysis of IL-13 by AutoStructure and, as in the manual structure analysis, were incorporated in the further structure calculations. AutoStructure identified more NOE distance constraints per residue for IL-13 than for FGF-2 and MMP-1 (Table III). There are three potential contributing factors for this performance (Table II): (1) IL-13 has more complete resonance assignments (94%), and a much larger number of stereospecific isopropyl methyl resonance assignments; (2) despite IL-13's smaller size, the input data set has more NOE peaks in the peak lists, especially for ^{13}C-NOESY; (3) the quality of the input data (*M score* = 6–7%) is much better for IL13 than for the other proteins tested. Interestingly, the total number of NOE distance constraints from AutoStructure analysis is also much higher than the total number from manual analysis (Table III). The mean coordinate differences for well-ordered heavy atoms of IL-13 structures determined by AutoStructure and manual analyses are < 1.0 Å. About 4% of the constraints from manual analysis have violations > 1.0 Å, compared to the structures generated by automated analysis. A slightly higher number of constraints (~6%) identified by automated analysis are violated by > 1.0 Å when compared with structures determined by manual analysis. As in the other systems tested, most of these differences are in the regions of the structure that are not well defined, and they do not significantly affect the overall structure. Again, more than twice the number of dihedral angle constraints were used in the manual analysis. Comparison with another IL-13 NMR structure [IL-13c, Fig. 4(c), bottom] [65] indicates that the NMR structures determined manually and by AutoStructure using the same NOESY data are about equally similar to a second independently determined manual NMR structure, with backbone mean coordinate differences to this independent NMR structure of 1.0 ± 0.1 Å (Fig. 4, bottom). RPF scores (Table III) indicate that the structures generated by automated methods fit the data better than the structures determined by manual analysis. However, PROCHECK

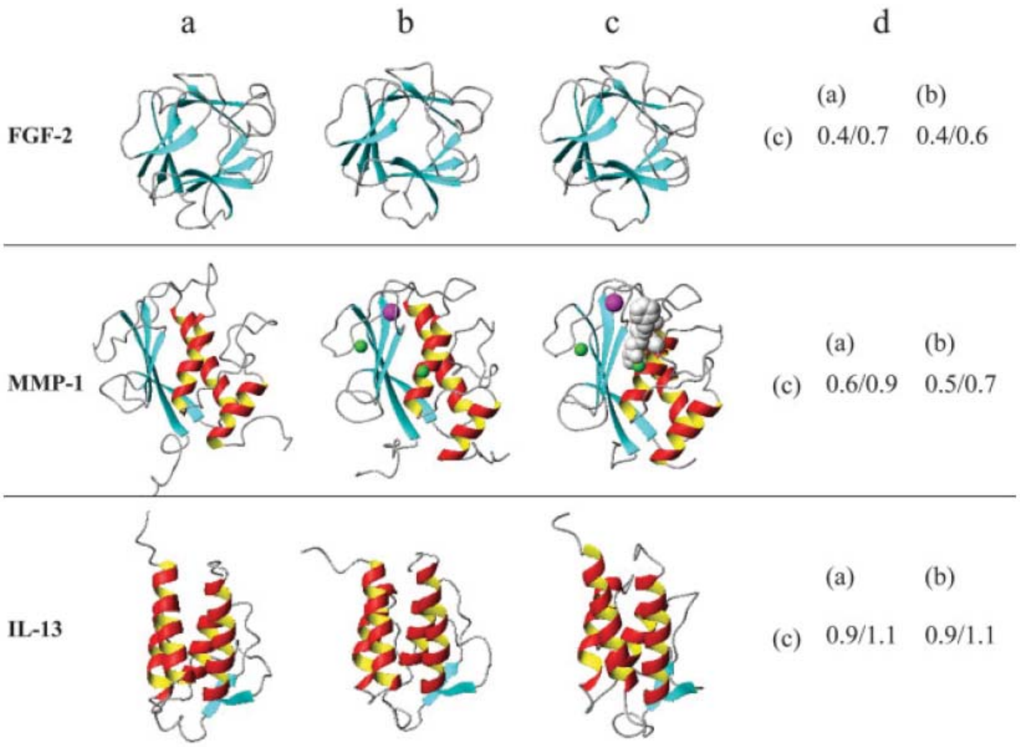

Figure 4. Ribbon diagrams of representative structures of FGF-2, MMP-1, and IL-13 proteins used for the validation of the AutoStructure/XPLOR process: (**a**) final structures from cycle 10; (**b**) structures deposited in PDB, analyzed using the same NMR data set; (**c**) structures determined by X-ray crystallography or third NMR group; (**d**) mean coordinate differences (Å) in the secondary structure region between structures (c) and structures (a) and (b). In each structure comparison, first value is the RMSD differences for backbone atoms and value following the / symbol is the RMSD differences for heavy atoms.

Ramachandran analysis indicates that IL-13 structures determined by the careful manual analysis have slightly better stereochemical qualities compared to the structure generated by AutoStructure.

We also tested these three input data sets using AutoStructure with DYANA for structure generation, in place of XPLOR. Similar results were obtained in these AutoStructure/DYANA calculations, with small mean coordinate differences between automated and manually analyzed structures. The AutoStructure/DYANA final structures have slighter smaller RMSDs within the ensemble, compared to AutoStructure/XPLOR. While it took 3-4 h to run a 10-cycle AutoStructure/DYANA calculation on a 14-node 1600 MHz Linux Athlon CPU cluster, the 10-cycle AutoStructure/XPLOR simulations each required ~ 20-25 h on the same Linux cluster.

Examples of De Novo Structure Determinations With AutoStructure

The AutoStructure program has been used in more than two dozen de novo protein structure determinations that have been deposited in the Protein Data Bank (PDB), several of which have already been published. [15–22] Figure 5(a) shows examples of these protein structures analyzed using AutoStructure [15] [20] [22] (group I), which could subsequently be validated by independently determined NMR or X-ray crystallographic structures of homologous proteins (group II). [26–28] At the time when group I structures were first determined and deposited into PDB, the group II structures were not yet released from the PDB and were not available to provide guidance for de novo structure determination of group I proteins. After the group II homolog structures were released by the PDB, we compared the differences between these two groups using the program CE. [66] The backbone RMSDs between the two groups are between 2.0 and 3.0 Å. [66] These good agreements demonstrate the robustness and reliability of the program AutoStructure.

AutoStructure has also been used for homodimeric protein structure analysis. [17] [18] Figure 5(b) shows a representative example, the solution NMR structures of TM1bZip N-terminal segment of human α-tropomyosin determined with AutoStructure. [17] The superposition of all heavy atoms [Fig. 5(b), top-right] reveals classic coil-coil side-chain contact patterns.

DISCUSSION

Fold Analysis

AutoStructure uses a *bottom-up* algorithmic approach, incorporating expert knowledge to guide the search for NOESY

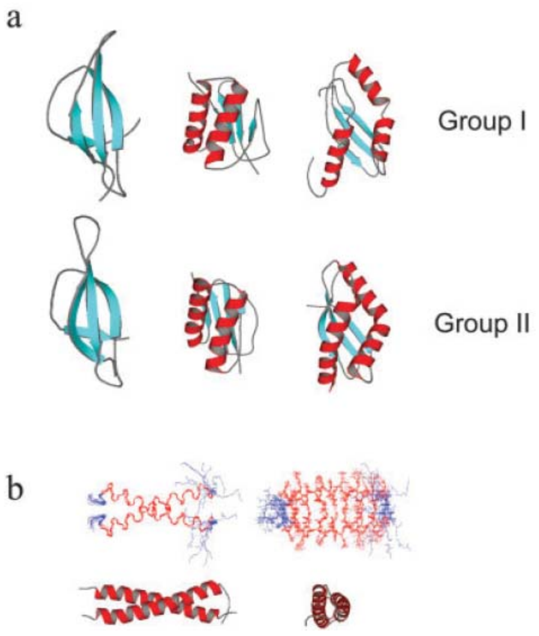

Figure 5. Examples of de novo structure determinations with AutoStructure. (**a**) Top (group I): ribbon diagrams of representative solution NMR structures analyzed with AutoStructure: 30S ribosomal protein S28E from *Pyrococcus horikoshii* (PDB: 1NY4), [22] dynein light chain Lc8 from *Drosophila* (PDB: 1RHW), [20] and ribosome-binding factor A (RbfA) from *Escherichia coli* (PDB: 1KKG) [15] (left to right). Bottom (group II): ribbon diagrams of representative homologus protein structures analyzed subsequentialy and independently using other methods: Solution NMR structure of 30S ribosomal protein S28E from *Methanobacterium thermoautotrophicum* (PDB: 1NE3), [26] solution NMR structure of Lc8 from *Rattus norvegicus* (PDB: 1PWJ), [27] and X-ray crystallography structure of RbfA from *Haemophilus influenzae* (PDB: 1JOS). (**b**) Solution NMR structures of TM1bZip N-terminal segment of human α-tropomyosin determined with AutoStructure [17] (PDB: 1IHQ). 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.

cross peak assignments and a self-consistent network of distance constraints. Rules for fold analysis are derived from knowledge of regular helix and β-sheet geometries, and standard models of packing between secondary structure elements. [43] For example, interstrand alignments within the β-sheet are identified directly from backbone subgraphs of G_{ANOE} in cycle 1 prior to the generation of an initial fold. Short β-sheet alignments may be missed, as their backbone subgraphs are too small to be detected reliably, but these are generally identified in later stages of iterative analysis using intermediate 3D structures. This β-sheet alignment method is very important for analysis of β and α/β protein structures. In the examples of FGF-2 and MMP-1 used in this study, all of the major β-sheet alignments are correctly identified in cycle 1 prior to initial structure-generation calculations. On the other hand, packing at helix-helix interfaces generally use ± 1n, ± 3n, and ± 4n rows and involves mainly side-chain atoms. [43] These helix-helix packing interactions are identified by unique NOESY crosspeak assignments made in cycle 1 and from constraints implied by intermediate 3D structures during iterative cycle analysis. It is hard to reliably identify helix packing directly from G_{ANOE} prior to generating an initial structure, due to side-chain packing variability.

About 5–10% of helix conformations in proteins are 3_{10}-helices, which generally tend to occur at the termini of α-helices. Loose helical dihedral angle constraints, which support both 3_{10}-helix and α-helix conformations, are used for residues identified as helical from chemical shift, scalar coupling, and other NMR data. Hydrogen bond constraints $\mathrm{O}(i)$ - $\mathrm{H}^{\mathrm{N}}(i + 4)$ for α-helix residues are added only if $\mathrm{H}^{\alpha}\mathrm{H}^{\mathrm{N}}(i,\ i + 4)$ interactions are present in HG_{NOE} or the hydrogen bond is detected from the intermediate structures. Thus, AutoStructure can properly distinguish 3_{10}- and α-helix conformations when (and if) the NOESY data distinguish these structures.

The presence of minor species, or of alternate conformations in slow equilibrium, is a general challenge not only for automated NMR structure analysis but also for manual analysis. Identification of minor species or alternate conformations requires combined analysis of NOE and resonance assignments. AutoStructure is not able to identify the presence of minor species or alternate conformations. AutoStructure is designed to find an optimum self-consistent set of NOE assignments matching to the resonance assignment list, and edges with large mismatches (which may arise because of the presence of minor species or alternate conformations) may be considered to be inconsistent with the resonance assignment list, and will therefore not be assigned.

One major assumption used by AutoStructure is that for most protein structures, a "low resolution" initial fold can be built from spectral data providing secondary structure information and a small portion of "unique" long-range NOE-linked proton pair interactions. As we have shown elsewhere, [16] this assumption is valid for small proteins even for minimum constraint approaches in which only the H^{N}-H^{N}, H^{N}-$\mathrm{H}^{\mathrm{methyl}}$, and $\mathrm{H}^{\mathrm{methyl}}$-$\mathrm{H}^{\mathrm{methyl}}$ NOEs are assigned. [16] [67] For proteins with few secondary structure elements, a higher proportion of "unique" NOE-linked proton pair interactions are generally required.

Quality Control of the NMR Data, AutoStructure Trajectories, and the Derived Structures

The input data (both set R and set NOE) quality for AutoStructure calculations is assessed by the *M score*. Reliable AutoStructure calculations require *M score* values < 25%; that is, > 75% of the two- and three-bond connected peaks predicted from set R should be observed in G_{ANOE}. Best performance is observed with data providing *M scores* < 10%. The input peak list (set NOE) should contain at least 90% real cross peaks, and the input resonance assignments (set R) must 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 ring packing.

Good initial folds have a DP score > 0.40 at the end of cycle 1, and > 0.60 at the end of cycle 2; intermediate structures providing lower DP values may have incorrect local or global fold topology, and require better quality input data. Structures exhibiting lower quality scores at this stage require refinement of NOESY peaks lists and/or resonance assignment lists. Once structures can be generated in cycles 1 and 2 with reasonable DP scores and other statistics which indicate they are self-consistent and reliable, we typically run a full AutoStructure calculation of 10 cycles. The quality of NOESY peak lists and resonance assignment lists can often be improved further by examining the output of AutoStructure. The final structures calculated from AutoStructure should have F-measure score > 0.90 and DP score > 0.70, [29] given near complete input data sets ($M < 25\%$). Other factors used to judge the quality of final-cycle AutoStructure models include backbone RMSD values for well-defined segments in the final ensemble of structures ≤ 1Å, > 10 conformationally restricting constraints per residue, > 80% of NOESY peaks assigned in the final cycle, and low energies from XPLOR/CNS calculations or small target function values from DYANA (< 10 Å^2).

Utility of AutoStructure for Model-Based Peak Picking and Structure Refinement

Peak lists do not have to be perfect. AutoStructure can handle the presence of artifactual peaks and incompleteness to some degree; 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. At step 3, AutoStructure reports list of expected peaks that are separated by two-bond or three-bond connectives, but missing from the peak list for manual validation of both the qualities of the peak picking and the resonance assignments. At step 8, RPF scores [29] are used to compare the 3D structure with the NOESY peak list data, and to assess the quality of the final 3D structures. RPF scores report to the user: (1) NOESY peak list entries inconsistent with the 3D structure, and (2) NOESY peaks that are expected because the corresponding proton pairs are close in the 3D structure, but which are missing from the NOESY peak lists. This information can be used to improve the peak picking process and refine the NOESY peaks lists.

Refinement by restrained molecular dynamics in explicit solvent [68] can also improve the sterochemical quality of the final 3D structures generated by AutoStructure/XPLOR. We have further refined the structures of FGF-2, MMP-1, and IL-13 generated by AutoStructure using constrained energy minimization in a water environment. These results are summarized in a Supplementary Table S1. The stereochemical qualities of refined structures are improved; in fact, the resulting refined structures of FGF-2 and IL-13 have slightly better sterochemical qualities than the structures determined by the manual analysis. RPF scores, which compare the structures to the NOESY peak lists, are also slightly improved with energy refinement, indicating that these refined structures equally/or better fit with the input data set. However, despite these excellent results with FGF-2, MMP-1, and IL-13, protein NMR structures generated with AutoStructure should be considered to be a good starting point for manual refinement and validation of NOESY crosspeak assignments, rather than a final result.

Utility of AutoStructure for Facilitating Analysis of Resonance Assignments

AutoStructure can also be used to facilitate and validate resonance assignments. 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 provide a 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 peak lists for validation of the two-bond and three-bond connectivity of the side-chain resonances.

Comparison With Other Automated NOESY Analysis Software

AutoStructure uses a *bottom-up* approach to NOESY data analysis, building HG_{NOE} piece-by-piece from G_{ANOE} using topology constraint networks, which distinguishes it from other successful NOESY interpretation programs. For example, the program ARIA [7] [8] uses a *top-down* approach to find an optimal solution HG_{NOE}, directly initiating HG_{NOE} from G_{ANOE} ($HG_{\mathrm{NOE}} = G_{\mathrm{ANOE}}$). Model structures are then built from HG_{NOE} using ambiguous constraint [7] [8] strategies. HG_{NOE} is iteratively trimmed using the resulting model structures by removing edges whose linking proton pairs are far apart in the intermediate model structure. Underlying the top-down 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 belongs to the correct solution G_{NOE}. [7] [8] Noise peaks in the NOESY peak lists and missing resonance assignments generally violate this assumption. The program CANDID [9] 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. The correctness assumption and top-down approaches generally require high-quality input R and NOE sets (e.g., the R set needs to be nearly complete and most NOE cross peaks should be real). [9] [69] CANDID's network anchoring and constraint-combination methods still require some 90% complete resonance assignments, almost complete aromatic side-chain assignments, low percentage of noise peaks, and small chemical shift variations. [69] [70] For both ARIA and CANDID, it is also important to obtain a well-converged initial fold (RMSD < 3.0 Å) directly from G_{ANOE}. [69] [70]

AutoStructure's novel bottom-up topology-constrained approach distinguishes it from ARIA and CANDID. 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. During the AutoStructure *initial fold analysis*, most NOE-linked proton pairs are identified by consistency analysis of polypeptide geometry and fold topology. Although the above correctness assumption is critical in interpreting "unique connections" [step 5b in Fig. 1], the *potential contact support analysis* filters out weakly supported, but otherwise unique, connections. Moreover, these "unique connections" identified in step 5b (Fig. 1) account for only 5–10% of total edges in HG_{NOE} at the end of cycle 1. AutoStructure also provides tools to manually validate against the frequency-domain spectra the "correctness assumption" for a small list of critical "uniquely connected" peaks that are consistently violated in the initial fold analysis. HG_{NOE} is then built up by iteratively adding linkages that are consistent with the intermediate fold topology and 3D models. It is possible to rule-in false interactions that may generate local distortions, but only if they are well supported by the intermediate structures. Noise and other artifacts in the NOESY peak list minimally affect AutoStructure analysis, as these noise peaks are generally not supported by the topology constraint networks and/or intermediate structures. Additional experimental information, including dihedral angle, hydrogen bond, and RDC constraints, can also be used by the program to avoid local structure distortions and to identify inconsistent "noise peaks" in the NOESY peak lists. In these ways, AutoStructure uses constraint satisfaction methods [71] to provide self-consistent analysis of the NMR data. At any given point of execution, the search engine of AutoStructure rules-in only those candidate NOE assignments that are highly consistent with the topology constraint networks or current partial solution. This approach makes the program less sensitive to the effects of spectral artifacts and incompleteness of the resonance assignments.

From assigned resonances, AutoStructure identifies secondary structures, including β-sheet alignments, using a graph-based pattern discovery method derived from secondary structure networks first characterized by Wüthrich. [38] The program JIGSAW also utilizes a novel algorithm to identify graph-based secondary structure patterns from unassigned resonances in order to determine sequence-specific resonance assignments within these secondary structures. [14] Both AutoStructure and JIGSAW use similar secondary structure patterns for constraint propagation, but the pattern discovery methods and the objectives of the two programs are different.

CONCLUSIONS

This article presents a novel bottom-up topology-constrained distance network analysis algorithm for NOE interpretation. AutoStructure incorporates a new statistical method, RPF scores, [29] for comparing 3D protein structures against the NMR input data, and for quality assessment of the assignment trajectories and final NMR structures. Using these algorithms, AutoStructure has been evaluated using three different human protein NMR test data sets: FGF-2, IL-13, and MMP-1, ranging in size from 113 to 169 amino acid residues. The mean coordinate differences between structures determined by AutoStructure and by manual analysis (0.5–0.8 Å for backbone atoms of ordered residues) demonstrate good accuracy of these automated methods. While protein structures generated by careful manual analysis generally exhibit somewhat better stereochemical quality and RPF structure quality scores, [29] as we have shown elsewhere, [25]automatically generated NOESY peak list assignments and 3D structures by AutoStructure provide an excellent starting point for careful structure refinement. The AutoStructure/RPF output also provides rich information in the form of site-specific recall (flagging NOESY peaks inconsistent with the 3D structure) and precision (flagging close contacts in the 3D structure that are not supported by data in the NOESY peak lists) scores that can be used to improve the interpretation of NOESY spectra and to refine the NOESY peak and resonance assignment lists. This process of iterative structure analysis, RPF analysis, refinement of NOESY peaks lists by visual inspection, and reassignment of NOESY cross peaks with AutoStructure provides a means of using the AutoStructure and RPF software together to refine the protein NMR structure. Moreover, as the *bottom-up* algorithms used by AutoStructure are quite different from the top-down methods used by CANDID and ARIA, these different methods will exhibit different strengths and weaknesses, and complementary use of multiple automated analysis methods in parallel can also provide a consensus approach for validating NOESY peak list assignments. [25] The AutoStructure program is currently being used by several NMR groups, and over the last few years, more than two dozen protein structures have been determined using AutoStructures. [15–22]

Software Availability

AutoStructure for automated NOESY data interpretation and structure calculation with DYANA or XPLOR/CNS is available from the authors upon request.

Acknowledgements

We thank J. Aramini, C. Kulikowski, H. Moseley, P. Rossi, D. Snyder, G. V. T. Swapna, and T. Szyperski for helpful discussion and comments on the manuscript. We also thank G. M. Clore and N. Tjandra for kindly providing NIH version 1.1.2 of X-PLOR.

References

1 Montelione GT, Zheng D, Huang YJ, Gunsalus KC, Szyperski T. Protein NMR spectroscopy in structural genomics. *Nat Struct Biol* 2000; **7**: 982-985.

2 Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* 1998; **54**: 905-921.

3 Brünger AT. *X-PLOR, Version 3.1: a system for X-ray crystallography and NMR*. New Haven: Yale University Press; 1992.

4 Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* 1997; **273**: 283-298.

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

6 Mumenthaler C, Güntert P, Braun W, Wüthrich K. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J Biomol NMR* 1997; **10**: 351-362.

7 Nilges M. Calculation of protein structures with ambiguous distance restraints: automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J Mol Biol* 1995; **245**: 645-660.

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

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

10 Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM. Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear Overhauser enhancement spectra and chemical shift assignments. *J Am Chem Soc* 2004; **126**: 6258-6273.

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

12 Adler M. Modified genetic algorithm resolves ambiguous NOE restraints and reduces unsightly NOE violations. *Proteins* 2000; **39**: 385-392.

13 Grishaev A, Llinas M. CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proc Natl Acad Sci USA* 2002; **99**: 6707-6712.

14 Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR. The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J Comput Biol* 2000; **7**: 537-558.

15 Huang YJ, Swapna GV, Rajan PK, Ke H, Xia B, Shukla K, Inouye M, Montelione GT. Solution NMR structure of ribosome-binding factor A (RbfA), a cold-shock adaptation protein from *Escherichia coli*. *J Mol Biol* 2003; **327**: 521-536.

16 Zheng D, Huang YJ, Moseley HN, Xiao R, Aramini J, Swapna GV, Montelione GT. Automated protein fold determination using a minimal NMR constraint strategy. *Protein Sci* 2003; **12**: 1232-1246.

17 Greenfield NJ, Huang YJ, Palm T, Swapna GV, Monleon D, Montelione GT, Hitchcock-DeGregori SE. 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* 2001; **312**: 833-847.

18 Greenfield NJ, Swapna GV, Huang Y, Palm T, Graboski S, Montelione GT, Hitchcock-DeGregori SE. The structure of the carboxyl terminus of striated alpha-tropomyosin in solution reveals an unusual parallel arrangement of interacting alpha-helices. *Biochemistry* 2003; **42**: 614-619.

19 Bayro MJ, Mukhopadhyay J, Swapna GV, Huang JY, Ma LC, Sineva E, Dawson PE, Montelione GT, Ebright RH. Structure of antibacterial peptide microcin J25: a 21-residue lariat protoknot. *J Am Chem Soc* 2003; **125**: 12382-12383.

20 Makokha M, Huang YJ, Montelione GT, Edison AS, Barbar E. The solution structure of the pH-induced monomeric dyein light chain LC8 from *Drosophila*. *Protein Sci* 2004; **13**: 727-734.

21 Ramelot TA, Ni S, Goldsmith-Fischman S, Cort JR, Honig B, Kennedy MA. Solution structure of *Vibrio cholerae* protein VC0424: a variation of the ferredoxin-like fold. *Protein Sci* 2003; **12**: 1556-1561.

22 Aramini JM, Huang YJ, Cort JR, Goldsmith-Fischman S, Xiao R, Shih LY, Ho CK, Liu J, Rost B, Honig B, Kennedy MA, Acton TB, Montelione GT. Solution NMR structure of the 30S ribosomal protein S28E from *Pyrococcus horikoshii*. *Protein Sci* 2003; **12**: 2823-2830.

23 Huang YJ, Moseley HN, Baran MC, Arrowsmith C, Powers R, Tejero R, Szyperski T, Montelione GT. An integrated platform for automated analysis of protein NMR structures. *Methods Enzymol* 2005; **394**: 111-141.

24 Baran MC, Huang YJ, Moseley HN, Montelione GT. Automated analysis of protein NMR assignments and structures. *Chem Rev* 2004; **104**: 3541-3556.

25 Liu GS, Shen Y, Atreya HS, Parish D, Shao Y, Sukumaran DK, Xiao R, Yee A, Acton TB, Arrowsmith CH, Montelione GT, Szyperski T. NMR data collection and analysis protocol for high-throughput protein structure determination. *Proc Natl Acad Sci USA* 2005; **102**: 10487-10492.

26 Wu B, Yee A, Pineda-Lucena A, Semesi A, Ramelot TA, Cort JR, Jung JW, Edwards A, Lee W, Kennedy M, Arrowsmith CH. Solution structure of ribosomal protein S28E from *Methanobacterium thermoautotrophicum*. *Protein Sci* 2003; **12**: 2831-2837.

27 Wang W, Lo KW, Kan HM, Fan JS, Zhang M. Structure of the monomeric 8-kDa dynein light chain and mechanism of the domain-swapped dimer assembly. *J Biol Chem* 2003; **278**: 41491-41499.

28 Rubin SM, Pelton JG, Yokota H, Kim R, Wemmer DE. Solution structure of a putative ribosome binding protein from *Mycoplasma pneumoniae* and comparison to a distant homolog. *J Struct Funct Genomics* 2003; **4**: 235-243.

29 Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J Am Chem Soc* 2005; **127**: 1665-1674.

30 Wako H, Scheraga HA. Visualization of the nature of protein folding by a study of a distance constraint approach in two-dimensional models. *Biopolymers* 1982; **21**: 611-632.

31 Sippl MJ, Scheraga HA. Solution of the embedding problem and decomposition of symmetric matrices. *Proc Natl Acad Sci USA* 1985; **82**: 2197-2201.

32 Havel TF, Kuntz ID, Crippen GM. The combinatorial distance geometry method for the calculation of molecular conformation: I. A new approach to an old problem. *J Theor Biol* 1983; **104**: 359-381.

33 Cormen TH, Leiserson CE, Rivest RL. *Introduction to algorithms*. Cambridge, MA/New York: MIT Press/McGraw-Hill; 1990.

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

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

36 Wüthrich K, Billeter M, Braun W. 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* 1983; **169**: 949-961.

37 Billeter M, Braun W, Wüthrich K. Sequential resonance assignments in protein ^{1}H 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* 1982; **155**: 321-346.

38 Wüthrich K. *NMR of proteins and nucleic acids*. New York: Wiley; 1986.

39 Wishart DS, Sykes BD. The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J Biomol NMR* 1994; **4**: 171-180.

40 Huang YJ. *Automated determination of protein structures from NMR data by iterative analysis of self-consistent contact patterns*. New Brunswick, NJ: Rutgers University; 2001.

41 Tejero R, Monleon D, Celda B, Powers R, Montelione GT. 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* 1999; **15**: 251-264.

42 Wüthrich K, Billeter M, Braun W. Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton-proton distances. *J Mol Biol* 1984; **180**: 715-740.

43 Chothia C. Principles that determine the structure of proteins. *Annu Rev Biochem* 1984; **53**: 537-572.

44 Chou KC, Nemethy G, Rumsey S, Tuttle RW, Scheraga HA. Interactions between an alpha-helix and a beta-sheet: energetics of alpha/beta packing in proteins. *J Mol Biol* 1985; **186**: 591-609.

45 Cohen FE, Sternberg MJ, Taylor WR. Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. *J Mol Biol* 1982; **156**: 821-862.

46 Chothia C, Levitt M, Richardson D. Helix to helix packing in proteins. *J Mol Biol* 1981; **145**: 215-250.

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

48 Moy FJ, Seddon AP, Campbell EB, Bohlen P, Powers R. ^{1}H, ^{15}N, ^{13}C and ^{13}CO assignments and secondary structure determination of basic fibroblast growth factor using 3D heteronuclear NMR spectroscopy. *J Biomol NMR* 1995; **6**: 245-254.

49 Moy FJ, Seddon AP, Bohlen P, Powers R. High-resolution solution structure of basic fibroblast growth factor determined by multidimensional heteronuclear magnetic resonance spectroscopy. *Biochemistry* 1996; **35**: 13552-13561.

50 Moy FJ, Pisano MR, Chanda PK, Urbano C, Killar LM, Sung M-L, Powers R. Assignments, secondary structure and dynamics of the inhibitor-free catalytic fragment of human fibroblast collagenase. *J Biomol NMR* 1997; **10**: 9-19.

51 Moy FJ, Chanda PK, Cosmi S, Pisano MR, Urbano C, Wilhelm J, Powers R. High-resolution solution structure of the inhibitor-free catalytic fragment of human fibroblast collagenase determined by multidimensional NMR. *Biochemistry* 1998; **37**: 1495-1504.

52 Moy FJ, Diblasio E, Wilhelm J, Powers R. Solution structure of human IL-13 and implication for receptor binding. *J Mol Biol* 2001; **310**: 219-230.

53 Clore GM, Gronenborn AM. Multidimensional heteronuclear nuclear magnetic resonance of proteins. *Methods Enzymol* 1994; **239**: 349-362.

54 Bax A, Vuister GW, Grzesiek S, Delaglio F. Measurement of homo- and heteronuclear J couplings from quantitative J correlation. *Methods Enzymol* 1994; **239**: 79-105.

55 Nilges M, Gronenborn AM, Brüenger AT, Clore GM. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints: application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng* 1988; **2**: 27-38.

56 Clore GM, Appella E, Yamada M, Matsushima K, Gronenborn AM. Three-dimensional structure of interleukin 8 in solution. *Biochemistry* 1990; **29**: 1689-1696.

57 Garrett DS, Kuszewski J, Hancock TJ, Lodi PJ, Vuister GW, Gronenborn AM, Clore GM. The impact of direct refinement against three-bond HN-C.alpha.H coupling constants on protein structure determination by NMR. *J Magn Reson B* 1994; **104**: 99-103.

58 Kuszewski J, Qin J, Gronenborn AM, Clore GM. The impact of direct refinement against ^{13}Cα and ^{13}Cβ chemical shifts on protein structure determination by NMR. *J Magn Reson B* 1995; **106**: 92-96.

59 Kuszewski J, Gronenborn AM, Clore GM. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci* 1996; **5**: 1067-1080.

60 Kuszewski J, Gronenborn AM, Clore GM. Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. *J Magn Reson* 1997; **125**: 171-177.

61 Kuszewski J, Gronenborn AM, Clore GM. Improving the packing and accuracy of nmr structures with a pseudopotential for the radius of gyration. *J Am Chem Soc* 1999; **121**: 2337-2338.

62 Zhu X, Komiya H, Chirino A, Faham S, Fox GM, Arakawa T, Hsu BT, Rees DC. Three-dimensional structures of acidic and basic fibroblast growth factors. *Science* 1991; **251**: 90-93.

63 Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* 1996; **8**: 477-486.

64 Spurlino JC, Smallwood AM, Carlton DD, Banks TM, Vavra KJ, Johnson JS, Cook ER, Falvo J, Wahl RC, Pulvino TA, Wendoloski JJ, Smith DL. 1.56 Å structure of mature truncated human fibroblast collagenase. *Proteins* 1994; **19**: 98-109.

65 Eisenmesser EZ, Horita DA, Altieri AS, Byrd RA. Solution structure of interleukin-13 and insights into receptor engagement. *J Mol Biol* 2001; **310**: 231-241.

66 Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 1998; **11**: 739-747.

67 Clore GM, Starich MR, Bewley GA, Cai M, Kuszewski J. Impact of residual dipolar couplings on the accuracy of NMR structures determined from a minimal number of NOE restraints. *J Am Chem Soc* 1999; **121**: 6513-6514.

68 Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M. Refinement of protein structures in explicit solvent. *Proteins* 2003; **50**: 496-506.

69 Jee J, Güntert P. Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *J Struct Funct Genomics* 2003; **4**: 179-189.

70 Ferentz AE, Wagner G. NMR spectroscopy: a multifaceted approach to macromolecular structure. *Q Rev Biophys* 2000; **33**: 29-65.

71 Russell SJ, Norvig P. *Artificial intelligence : a modern approach*. Englewood Cliffs, NJ: Prentice-Hall; 1995.

PROTEINS
Revised v2 06/28/05

Revised

Supplemental Material for:

A topology-constrained distance network algorithm for protein structure determination from NOESY data

Yuanpeng Janet Huang[1], Roberto Tejero[1,э],

Robert Powers[2] and Gaetano T. Montelione[1, 3*]

[1]Center for Advanced Biotechnology and Medicine and
Department of Molecular Biology and Biochemistry
Rutgers University, Piscataway, NJ 08854-5638

[2]Department of Chemistry
University of Nebraska-Lincoln
Lincoln, NE 68588

[3]Department of Biochemistry, Robert Wood Johnson Medical School,
Univ. of Medicine and Dentistry of New Jersey, Piscataway, NJ 08854-5638

[э]Present address: Departamento de Química Física., Universidad de Valencia, Dr. Moliner, 50646100-Burjassot (Valencia), SPAIN

* To whom correspondence should be addressed:
Prof. Gaetano T. Montelione
CABM, Rutgers University
679 Hoes Lane
Piscataway, NJ 08854-5638
Phone: 732-235-5321   Fax: 732-235-5633
e-mail: guy@cabm.rutgers.edu

Supplemental Descriptions of AutoStructure Algorithms

Analysis of spectral aliasing and construction of ambiguous network G^0_{ANOE} from input data sets R and NOE

AutoStructure supports input NOESY peaks lists with extensive aliasing, without the need to "unfold" these spectral data prior to analysis. For every peak p = (δ_1, δ_2, I) $\in$ NOE, an aliasing order (i.e. maximum fold of aliasing) m_i is calculated for each frequency dimension *i* by

$$m_i = \max\left(\left|\delta_{obs(i)} - \delta_{up(i)}\right|, \left|\delta_{obs(i)} - \delta_{down(i)}\right|\right) / sw_i \qquad \text{Eqn. 1}$$

where $\delta_{obs(i)}$ is the observed chemical shift, $\delta_{up(i)}$ and $\delta_{down(i)}$ are the most upfield and downfield chemical shifts in the resonance assignment table for the *i*th dimension, and sw_i is its acquired spectral sweep width. The aliasing order m_i is zero for unaliased chemical shifts. A temporary list of possible corresponding "unfolded" chemical shifts (δ_i) in each frequency dimension is then generated by equation (2).

$$\delta_i = \delta_{obs(i)} \pm \left(n \times sw_i\right) \qquad \text{Eqn. 2}$$

where $n \in 0..m_i$. The possible values *n* used in Eqn. 2 for each dimension can be restricted if the sign (i.e. positive or negative) of the intensity indicates an even or odd aliasing order in that frequency dimension [1]. Each possible value δ_i is then matched with atoms from set R within error tolerance Δerr^i (Table 1).

Calculation of potential contact scores pct(r1, r2)

AutoStructure utilized dynamically-generated residue-residue contact map information to rule out incorrect NOESY cross peak assignments and to rule-in structurally-consistent assignments. This step of is conceptually similar to the process of "network anchoring" used by the program CANDID[9], though somewhat more sophisticated in using knowledge of expected short distances within and between secondary structure elements. All NOE-linked proton pairs in G_{ANOE} are assigned a linking score (default value = 1) which provides an assessment of the densities of potential interproton connections for that protein pair across G_{ANOE}. For example, NOE peaks that are identified as having a symmetric peak in the NOESY data set, and thus validated, are assigned a linking score of 2. In this contact map analysis, methylene protons with different chemical shift values but attached to the same heavy atom are grouped together. We define the maximum number of NOE cross peaks linking any two heavy atom groups (max_{heavy}) as 2 × total number of observed proton chemical shifts associated with these two groups. For example, between two methylene groups with four distinct ^{1}H chemical shifts, $max_{heavy} = 8$. If the number of the linking peaks found in G_{ANOE} is less than $max_{heavy}/2$, all these peaks are treated at this stage as providing a sparsely supported linkage (linking score = 0) and are excluded from the following potential contact support analysis. For each potential residue contact pair (r1, r2), a potential contact supporting score pct(r1, r2) providing an assessment of the confidence in the proposed contact, is calculated:

$$
\begin{cases}
\text{pct}(r1, r2) = 0 \text{ if one of the residue pair is charged and another is hydrophobic} \\
\text{pct}(r1, r2) = 0 \text{ if the number of interactions between different (or non-degenerate)} \\
\qquad \text{protons of residue r1 and different (or non-degenerate) protons of residue} \\
\qquad \text{r2} \leq 2 \\
\text{otherwise,} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Eqn. 3}
\end{cases}
$$

pct(r1, r2) = $\sum$ linking scores of all NOE-linked proton pairs from G_{ANOE}(r1, r2)

where the second condition is a means of avoiding incorrect interpretations due to missing resonance assignments in R.

Apparently unique (frq(p) = 1) connections (h1, h2, p) from residue pairs (r1, r2) are added into HG_{NOE} only if the potential contact support score pct(r1,r2) $\geq$ $\text{pct}_{\text{cutoff}}$ (*potential contact support score cutoff*, Table 1), or pct(r1,r2) $\neq$ 0 and $\exists$(r3, r4), pct(r3,r4) $\geq$ $\text{pct}_{\text{cutoff-n}}$ (*potential contact support score cutoff for neighboring residue contacts*, Table 1), where (r3, r4) is a neighboring residue pair of the (r1, r2) pair in the contact map. For backbone protons, protons of small residues such as Gly and Ala, and peripheral protons like H^{ε} of Met which have lower densities of interresidue ^{1}H-^{1}H contacts, the cutoff threshold for $\text{pct}_{\text{cutoff}}$ is set lower than for other interactions. If the (r1, r2) residue pair is involved in helical-helical packing, supporting neighbor residue contacts includes residue pairs (r1, r2+i) and (r1+i, r2) (i$\in$\{1, 4, -3\}). For other types of contacts, supporting neighbor contacts include residue pairs (r1±i, r2± j) (1$\leq$ i+j $\leq$ 2, i $\in$ \{0,1,2\}, j $\in$ \{0,1,2\}).

Distance constraint generation

Interproton distances (d) between H^N, H^α, and H^β atoms are calibrated from the NOESY cross peak intensities (I), assuming the simple isolated two-spin pair approximation

$$I = k \times d^{-6}$$

Eqn. 4

where constant k depends on the scaling of the NOESY spectrum (see below), and is converted into four distance constraint ranges, *viz* ≤ 2.5 Å, ≤ 3.0 Å, ≤ 4.0 Å, and ≤ 5.0 Å.[2] Prior to binning, all the calibrated interproton distances are increased by 10%. All other NOESY-derived distance constraints involving side chain atoms are assigned to upper-bound values of 5.0 Å without calibration. We use loose upper bound constraints not only to compensate for NOESY cross peaks whose intensities are affected by either spin diffusion or partial overlapping, but also to allow molecules to overcome local minima in the search for a global minimum with respect to both the NMR-derived constraints and conformational energy.[3] For interacting groups of degenerate methylene or methyl protons, intensities of the corresponding NOESY cross peaks are divided by a factor of 2 or 3, respectively, and then calibrated using the isolated two-spin pair approximation (Eqn. 4). Similarly, in cases where multiple NOE interactions are assigned to a single NOESY cross peak, the NOESY cross peak intensities are first divided by the total number of multiple constraints and then calibrated using the isolated two-spin pair approximation calibration. In all cases, the sum of van der Waals radii (1.8Å) is used as the lower-bound distance limit. Upper- and lower- bound distance limits are then generated in a format suitable for input to structure generation programs with standard pseudo-atom corrections[4], as needed.

The scaling factor k of equation (4) is estimated based on the observation that the spatial distribution of hydrogen atoms in different globular proteins is closely similar [5]. Specifically, $k \cong \langle I \rangle / \langle d^{-6} \rangle$ where the average intensity $\langle I \rangle$ is computed from all non-diagonal NOESY cross peaks of G_{ANOE}, and the average distance value $\langle d^{-6} \rangle$ for NOEs among the backbone and H^{β} protons is assumed to be similar for all globular proteins . In order to estimate $\langle d^{-6} \rangle$, we selected 20 sets of high-resolution crystal protein coordinates from Protein Data Bank and computed all interproton backbone and H^{β} protons distances $d < 5$ Å. The relevant average value, $\langle d^{-6} \rangle \approx 0.0018$ Å^{-6}, corresponds to an average distance between backbone H^{N}, H^{α}, and H^{β} protons giving rise to an NOE interaction of ~ 3.8 Å. The value of the parameter $\langle d^{-6} \rangle$ can be adjusted for specific NOESY data sets, as required[16].

Dihedral angle constraint generations

Dihedral angle constraints are generated using the conformational grid search program HYPER [6], which is incorporated as part of the AutoStructure process. HYPER calculates the set of ϕ and ψ dihedral angles and stereospecific assignments of β methylene protons that are consistent with a combined analysis of vicinal scalar coupling constants, and local intra-residue and sequential NOE data calibrated using the isolated two-spin pair approximation. Loose dihedral angle constraints from the identified segments of secondary structures are also used in HYPER as a prior information: –95° <

$\phi < -35°$ and $-70° < \psi < 0°$ for residues adopting helical conformations, $-180° < \phi < -75°$ and $65° < \psi < 175°$ for residues adopting β-sheet conformations. For the three test data sets, loose dihedral angle (±40° for ϕ, ±50° for ψ) of high confident (score =10) dihedral angles derived from C^{α}/C^{β} chemical shifts [7] are also used as constraint input. Some of these input dihedral angle constraints, which were violated in the intermediate structures, were removed for final structure calculations.

Identification of hydrogen bonds

Backbone-backbone (bb/bb) hydrogen bond constraints $O(i)$ -$H^N(i+4)$ for α-helix residues are added only if the characteristic proton pair interactions $H^{\alpha}H^N(i, i+4)$ are presented in HG_{NOE} , or the hydrogen bond is consistently detected in intermediate structures, as described below. Characteristic bb/bb hydrogen bonds that are consistent with β-sheet NOE patterns and for which H^N donors are indicated by slow amide ^1H exchange data, are also identified and used in the structure calculation [2]. When slow amide ^1H exchange data are not available, postulated hydrogen bond distance constraints derived on the basis of well-characterized characteristic β-sheet NOE patterns [2] are used in the structure calculation. During iterative fold analysis, additional bb/bb hydrogen bond constraints are identified when *all* of the following conditions are satisfied: (i) H^N donors are included in the slow amide ^1H exchange list, (ii) O-H distance < 2.4 Å and angle H-N-O < 35° in at least 20% of the ensembles [8], (iii) the hydrogen bond residue pair is at least two residues apart, (iv) they have nearby assigned NOE interactions, and (v) both the donor and acceptor are not involved in other possible hydrogen-bonded interactions. Although bifurcated hydrogen bonds do occur in protein structures, in order to avoid potential errors, the current version of the AutoStructure software does not provide for bifurcated hydrogen-bond constraints. The upper / lower-bound constraints of these identified hydrogen bonds for O_i to $H^N{}_j$ and N_j are set to 2.3 / 1.5 Å and 3.3 / 2.4 Å, respectively.

Parallel structure calculation using XPLOR/CNS or DYANA

AutoStructure can generate input constraint files for either XPLOR/CNS, or DYANA for protein structure calculations. In each cycle, a large number (typically 64) of XPLOR/CNS or DYANA calculations are submitted to a Linux cluster and the best representative conformations (typically 10) with lowest energies or smallest target functions are selected for analysis in the next step. Our Linux cluster used is based on loosely-coupled dual 1600 MHz Athlon computers, managed by a distributed queuing system. Both PBS and DQS queuing systems are currently supported. A part of AutoStructure called CreateProc is responsible for managing the calculation, assigning different seed numbers for course grain parallel calculations, setting the calculation program and protocol to use, and activating the additional input data (e.g., residual dipolar coupling, chemical shifts, etc.) as needed.

The protocols for XPLOR [9] structure calculations are based on the standard schemes distributed with the program and modified for calculations using the Linux cluster. These protocols are loosely related to the original hybrid distance geometry-dynamical simulated annealing method of Nilges et al [10,11] using the XPLOR [12] program. In our case initial structures can be taken from previous models, generated from an extended conformation, or generated randomly. In the calculations described in this paper, we used randomly-generated initial coordinates. Next, a high temperature (Cartesian space) dynamics simulation is run to allow for good sampling of conformational space. During this stage, weights on covalent structure elements (bonds, angles, etc) are slowly increased from small initial values where van der Waals forces are applied only to C^{α} atoms, allowing other atoms to pass through each other in order to satisfy the experimentally derived data. Once the structure generation is complete, a high-temperature annealing stage takes place to explore conformational space. After the high temperature stage, special care is taken in the slow cooling stages by incorporating pseudopotentials for $^{3}J(H^{N}$-$H^{\alpha})$ coupling constants [13], secondary $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shift restraints,[14] and a conformational database potential [15,16]. The target function that is minimized during restrained minimization and simulated annealing comprises only quadratic harmonic terms for covalent geometry, $^{3}J(H^{N}$-$H^{\alpha})$ coupling constants and secondary $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shift restraints, square-well quadratic potentials for the experimental distance, torsion angle restraints, and a quartic van der Waals term for non-bonded contacts. All peptide bonds were constrained to be planar and trans unless otherwise indicated by experimental data. There were no hydrogen-bonding, electrostatic, or 6-12 Lennard-Jones empirical potential energy terms in the target function. The force constant for the conformational database was kept relatively low throughout the simulation to allow the experimental distance and torsional angle restraints to predominantly influence the resulting structures. The force constants for the NOE and dihedral restraints were 30 times and 10 times stronger, respectively, then the force constants used for the conformational database. Upon completion of the simulation, an energy evaluation (no minimization) including all potential energy function terms is carried out to obtain an estimation for the total energy of the calculated structures. Similar protocols for CNS structure calculations with AutoStructure are described elsewhere.[17]

For DYANA structure calculations, standard protocols were used. Each structure calculation used the fast DYANA torsion angle dynamics algorithm with the standard simulated annealing schedule with 4000 torsion angle dynamics, where no special simulated annealing strategies were taken.

Refinement of structures generated by AutoStructure using XPLOR with restrained molecular dynamics in explicit solvent

We also energy- refined the structures of FGF-2, MMP-1 and IL-13 generated by AutoStructure/XPLOR using the program CNS[18] by calculating a short restrained molecular dynamics in explicit solvent, using the protocol described by Linge et al.[19]. The NMR conformers were immersed in a 8 Ǻ shell of ‘TIP3P water’ molecules. The solvated protein was first heated in 200 MD steps from 100 to 500K, followed by a short refinement run of 1,000 MD steps at 500K. Finally, the system was cooled in 2,000 MD steps from 500 to 25 K followed by a very short energy minimization. The PARALLHDG5.3 force field described by Linge et al.[19] was used.

Table S1 shows that the stereochemical qualities of the energy refined structures are improved. In fact, the refined structures of FGF-2 and IL-13 have slightly better stereochemical qualities[20] than the structures determined by the manual analysis. RPF scores[21] are also slightly improved which indicates that these refined structures equally/or better fit with the input data set. The MOLPROBITY clash scores [22] for final structures calculated with AutoStructure/XPLOR are, however, slightly higher than the manually-determined structures after refinement (data not shown), suggesting that structures generated with carefully-manually assigned constraints are somewhat more accurate than those generated by the fully automated method.

Table S1. Analysis of structure quality scores before and after CNS refinement

Protein	FGF-2	MMP-1	IL-13
ProCheck[20] G-factors(phi-psi only /all dihedrals)			
AS-before refinement	-1.09/-1.03	-0.74/-0.93	-0.17/-0.56
AS-after refinement	-0.85/-0.74	-0.42/-0.54	0.04/-0.24
Manual analysis	-0.99/-0.78	0.04/-0.24	-0.06/0.02
RPF[21] Analysis (R/P/F/DP)			
AS-before refinement	94.2/92.1/93.2/85.4	88.6/89.2/88.9/79.5	87.3/96.9/91.9/79.8
AS-after refinement	94.3/92.3/93.3/85.8	89.2/89.3/89.2/80.7	87.0/97.3/91.8/79.5
Manual analysis	94.5/92.6/93.5/86.2	88.9/89.6/89.2/80.7	82.5/97.1/89.2/72.3
ProCheck[20] Ramachandran Statistics			
AS-before refinement	75.0/23.9/1.0/0.0	81.7/18.0/0.3/0.0	90.4/9.3/0.1/0.1
AS-after refinement	85.7/14.1/0.3/0.0	86.6/13.1/0.3/0.0	92.8/7.2/0.0/0.0
Manual analysis	77.5/21.5/1.0/0.0	90.1/9.8/0.31/0.0	91.1/8.0/0.9/0.0

References

1. Oschkinat H, Griesinger C, Kraulis PJ, Sorensen OW, Ernst RR, Gronenborn AM, Clore GM. Three-dimensional NMR spectroscopy of a protein in solution. Nature 1988;332:374-376.
2. Wuthrich K. NMR of proteins and nucleic acids. New York: Wiley; 1986.
3. Bassolino-Klimas D, Tejero R, Krystek SR, Metzler WJ, Montelione GT, Bruccoleri RE. Simulated annealing with restrained molecular dynamics using a flexible restraint potential: theory and evaluation with simulated NMR constraints. Protein Sci 1996;5:593-603.
4. Wuthrich K, Billeter M, Braun W. 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 1983;169:949-961.
5. Mumenthaler C, Guntert P, Braun W, Wuthrich K. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. J Biomol NMR 1997;10:351-362.
6. Tejero R, Monleon D, Celda B, Powers R, Montelione GT. 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 1999;15:251-264.
7. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 1999;13:289-302.
8. Billeter M, Qian Y, Otting G, Muller M, Gehring WJ, Wuthrich K. Determination of the three-dimensional structure of the Antennapedia homeodomain from Drosophila in solution by 1H nuclear magnetic resonance spectroscopy. J Mol Biol 1990;214:183-197.
9. Schwieters CD, Kuszewski JJ, Tjandra N, Marius Clore G. The Xplor-NIH NMR molecular structure determination package. J Magn Reson 2003;160:65-73.
10. Nilges M, Gronenborn AM, Bruenger AT, Clore GM. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. Protein Eng 1988;2:27-38.
11. Clore GM, Appella E, Yamada M, Matsushima K, Gronenborn AM. Three-dimensional structure of interleukin 8 in solution. Biochemistry 1990;29:1689-1696.
12. Brunger AT. X-PLOR, Version 3.1 : a system for X-ray crystallography and NMR. New Haven: Yale University Press; 1992.
13. Garrett DS, Kuszewski J, Hancock TJ, Lodi PJ, Vuister GW, Gronenborn AM, Clore GM. The impact of direct refinement against three-bond HN-C.alpha.H coupling constants on protein structure determination by NMR. J Magn Reson, Ser B 1994;104:99-103.

14. Kuszewski J, Qin J, Gronenborn AM, Clore GM. The impact of direct refinement against $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts on protein structure determination by NMR. J Magn Reson, Ser B 1995;106:92-96.
15. Kuszewski J, Gronenborn AM, Clore GM. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci 1996;5:1067-1080.
16. Kuszewski J, Gronenborn AM, Clore GM. Improvements and extensions in the conformational database potential for the refinement of NMR and x-ray structures of proteins and nucleic acids. J Magn Reson 1997;125:171-177.
17. Zheng D, Huang YJ, Moseley HN, Xiao R, Aramini J, Swapna GV, Montelione GT. Automated protein fold determination using a minimal NMR constraint strategy. Protein Sci 2003;12:1232-1246.
18. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54 (Pt 5):905-921.
19. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M. Refinement of protein structures in explicit solvent. Proteins 2003;50:496-506.
20. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996;8:477-486.
21. Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 2005;127:1665-1674.
22. Word JM, Bateman RC, Jr., Presley BK, Lovell SC, Richardson DC. Exploring steric constraints on protein mutations using MAGE/PROBE. Protein Sci 2000;9:2251-2259.