J|A|C|S
ARTICLES
Published on Web 01/22/2005

Protein NMR Recall, Precision, and *F*-measure Scores (RPF Scores): Structure Quality Assessment Measures Based on Information Retrieval Statistics

Yuanpeng J. Huang,[†] Robert Powers,[‡] and Gaetano T. Montelione*,[†]

Contribution from the Center for Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers University, Northeast Structural Genomics Consortium, and Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5368, and Department of Chemistry, University of Nebraska−Lincoln, Lincoln, Nebraska 68588.

Received May 17, 2004; E-mail: guy@cabm.rutgers.edu

Abstract: One of the most important challenges in modern protein NMR is the development of fast and sensitive structure quality assessment measures that can be used to evaluate the "goodness-of-fit" of the 3D structure with NOESY data, to indicate the correctness of the fold and accuracy of the resulting structure. Quality assessment is especially critical for automated NOESY interpretation and structure determination approaches. This paper describes new NMR quality assessment scores, including Recall, Precision, and *F*-measure scores (referred to here are "NMR RPF" scores), which quickly provide global measures of the goodness-of-fit of the 3D structures with NOESY peak lists using methods from information retrieval statistics. The sensitivity of the *F*-measure is improved using a scaled Fold Discriminating Power (DP) score. These statistical RPF scores are quite rapid to compute since NOE assignments and complete relaxation matrix calculations are not required. A graphical method for site-specific assessment of structure quality based on the Precision statistic is also described. These statistical measures are demonstrated to be valuable for assessing protein NMR structure accuracy. Their relationships to other proposed NMR "R-factors" and structure quality assessment scores are also discussed.

Introduction

Traditionally, distance constraints interpreted from NOESY spectra are used as the predominant source of structural information for most high-resolution protein NMR structure determinations. Although other NMR information, including residual dipolar coupling [1−3] and scalar coupling [4−7] data play an increasingly important role in structure and dynamic analysis, the large numbers of distance constraints generated from NOESY data generally provide the primary information used for protein structure determination. Accordingly, one of the most important challenges in modern protein NMR is to develop a fast and sensitive structure quality assessment measure which can be used to evaluate the "goodness-of-fit" of the 3D structure with NOESY peak lists to indicate the correctness of the fold and accuracy of the structure. Quality assessment is especially critical for quality control of automated NOESY interpretation and structure determinations, and for guiding the process of using intermediate 3D structures in iterative NOESY cross-peak assignment.

Despite the fact that a relatively simple and standard R-factor has been available for X-ray crystallography for many years,[8,9] to date, there is no generally accepted "NMR R-factor". Protein NMR structures can be validated by comparison of back-calculated spectra, peak lists, and/or constraints, which represent different interpretation steps in the structure determination process. Traditionally, protein NMR structures are evaluated at the constraint-validation step by comparison of back-calculated distances, dihedral angles, and other structural features with a list of experimental constraints derived from the spectroscopic data.[10] Such measures are often biased by the fact that these constraint lists are human interpretations of the spectroscopic data. In some cases of automated NOESY data interpretation,[11,12] constraints that are inconsistent with intermediate or final molecular structures are excluded from the derived "constraint lists", further compromising the value of comparisons between

[†] Center for Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers University, Northeast Structural Genomics Consortium, and Robert Wood Johnson Medical School.

[‡] Department of Chemistry, University of Nebraska−Lincoln.

[1] Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111−1114.

[2] Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 9279−9283.

[3] Bax, A. *Protein Sci.* **2003**, *12*, 1−16.

[4] Biamonti, C.; Rios, C. B.; Lyons, B. A.; Montelione, G. T. *Advances in Biophys. Chem.* **1994**, *4*, 51−120.

[5] Cornilescu, G.; Hu, J.; Bax, A. *J. Am. Chem. Soc.* **1999**, *121*, 2949.

[6] Cordier, F.; Grzesiek, S. *J. Am. Chem. Soc.* **1999**, *121*, 1601.

[7] Bax, A.; Vuister, G. W.; Grzesiek, S.; Delaglio, F.; Wang, A. C.; Tschudin, R.; Zhu, G. *Methods Enzymol.* **1994**, *239*, 79−105.

[8] Stout, G. H.; Jensen, L. H. *X-ray Structure Determination; A Practical Guide*; Macmillan: New York, 1968.

[9] Wilson, A. J. C. *Acta Crystallogr.* **1950**, *3*, 397−398.

[10] Doreleijers, J. F.; Raves, M. L.; Rullmann, T.; Kaptein, R. *J. Biomol. NMR* **1999**, *14*, 123−132.

[11] Herrmann, T.; Guntert, P.; Wuthrich, K. *J. Mol. Biol.* **2002**, *319*, 209−227.

[12] Huang, Y. J.; Swapna, G. V.; Rajan, P. K.; Ke, H.; Xia, B.; Shukla, K.; Inouye, M.; Montelione, G. T. *J. Mol. Biol.* **2003**, *327*, 521−536.

10.1021/ja047109h CCC: $30.25 © 2005 American Chemical Society

constraint lists and the derived structures as a means of structure validation. Protein NMR structures generated from NOESY and other data can be independently validated against sets of residual dipolar coupling (RDC) data.[3,13] RDCs are not sensitive to translational variations that preserve consistent relative bond vector alignments. As a result, there is a wide distribution of 3D structures which provide equally good fits to residual dipolar coupling data, particularly when such data is available for only a single orientation tensor. Recently, a new approach for *Q*uantitative *E*valuation of each *E*xperimental *N*MR restraint (QUEEN) has been reported.[14] This method is based on information theory in combination with a description of the structure in distance space. The QUEEN method identifies the crucial (i.e., important and unique) NMR constraints defining the protein structure, but does not provide an overall assessment of the accuracy of the structure. Other methods of structure quality assessment, including analysis of packing contacts, dihedral angle distributions, and conformational energies [15−17] are valuable for protein structure validation, but do not provide an assessment of the accuracy of the structure against the experimental data from which the structure is derived.

One approach for NMR structure quality assessment against NOESY data uses an R-factor definition similar to that used in X-ray crystallography. The NOESY spectrum is compared with a simulated NOESY spectrum back calculated from the ensemble of 3D structures. However, direct adaptation of the crystallographic R-factor to NMR data is challenging for several reasons. While crystallographic data is organized on diffraction lattice indices, NOESY cross peaks correlate resonance frequencies that are often partially or completely overlapped. In the most direct analogy with a crystallographic R-factor, each proton resonance frequency pair is treated as a lattice point, and the NOE intensity at each point on this lattice is back-calculated from the structure(s) under evaluation. However, the NOE effect is generally only transmitted through space over a distance $< \sim 5$ Å, while the majority of interproton distances in protein structures are > 5 Å. The corresponding distance matrix is dominated by the numerous number of "true negative data"; i.e., interproton interactions which are not detected in the either the experimental or back-calculated NOESY spectrum. Such a quality assessment score will not be sensitive and meaningful if all these "true negative" data points are included. We refer to this as the *true negative domination problem*. An additional issue is the impact of differential nuclear relaxation rates, which manifest internal and intermolecular dynamics, and "relayed" dipole−dipole interactions (i.e., spin-diffusion effects) that modulate peak intensities in complicated ways.[18]

Rather than computing all possible interproton interactions, an alternative improved approach is to compare only the intensity differences for NOESY cross peaks observed in experimental and/or back-calculated NOESY peak lists.[19−21] The program R-FAC[21] provides a set of NMR R-factor scores using complete relaxation matrix formalism, including a global R-factor, different R-factors for the intraresidue, interresidue, sequential, medium range, and long-range NOEs. One particular R-factor calculated by R-FAC (monitoring primarily long-range NOEs, and referred to as R5) was reported to be most useful in measuring the quality of an NMR structure. However, cross-peak overlaps, effects of spin diffusion, internal and intermolecular dynamics, and differential heteronuclear polarization transfer efficiencies create difficulties in making accurate estimates of NOESY cross-peak intensities from 3D structures, even when using relaxation matrix calculations.[18] Accordingly, structure evaluation methods that focus on comparisons of relative NOESY cross-peak intensities may be severely biased when there are extensive cross-peak overlap and/or inaccuracies in computing cross-peak intensities from theoretical considerations. Although a complete and accurate analysis of the network of spin−spin interactions is feasible,[18] such calculations are time intensive and not generally suitable for use in guiding automated or manual structure refinement processes.

In this paper, we describe a novel, rapid and simple approach for calculating global structure quality scores that avoids the *true negative domination problem*, while preventing inaccuracies in simulating peak intensities from dominating the structure quality assessment. The field of information retrieval statistics has encountered a similar *true negative domination problem*. In particular, Recall, Precision and performance (*F*-measure) are statistical quality scores commonly used in information retrieval analysis that do not account for true negative data points.[22,23] The new NMR quality factors described in this paper, based on these statistical methods from information science, quickly provide a global measure of the goodness-of-fit of the 3D structures with NOESY peak lists and resonance assignment data. These statistical scores are quite rapid to compute, as NOE cross-peak assignments and complete relaxation matrix calculations are not required, and are shown here to have good correlations with structure accuracy. In addition, we show how site-specific information derived from these statistical scores can be used to identify problem regions of NOESY interpretation in the context of the 3D protein structure. These features make the quality scores useful for both evaluating intermediate structures used in the structure refinement process, and for quality assessment of the final protein NMR structures.

Materials and Methods

Recall and Precision Analysis for Quality Assessment of NMR Structures. We have developed NMR structure quality assessment scores from information retrieval statistics. Detailed formulations of these structure quality assessment scores are presented as Supporting Information. Here, we outline key definitions of these NMR structure quality scores. From the resonance assignment table and NOESY cross-peak lists, an *ambiguous NOE network* G_{ANOE} is built (Figure 1). Vertices (V) represent all protons from the resonance assignment table and edges (E_{ANOE}) connect the vertices and represent all potential associated NOEs from the NOESY peak lists within a *match tolerance*. In constructing G_{ANOE}, each NOESY cross peak (p) may be ambigu-

[13] Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc* **1998**, *120*, 6836−6837.

[14] Nabuurs, S. B.; Spronk, C. A.; Krieger, E.; Maassen, H.; Vriend, G.; Vuister, G. W. *J. Am. Chem. Soc.* **2003**, *125*, 12026−12034.

[15] Word, J. M.; Bateman, R. C. Jr.; Presley, B. K.; Lovell, S. C.; Richardson, D. C. *Protein Sci* **2000**, *9*, 2251−2259.

[16] Vriend, G. *WHAT IF: A molecular modeling and drug design program*, 1990.

[17] Laskowski, R. A.; Rullmannn, J. A.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. *J. Biomol. NMR* **1996**, *8*, 477−486.

[18] Borgias, B. A.; James, T. L. *Methods Enzymol* **1989**, 176, 169−183.

[19] Gonzalez, C.; Rullmann, J. A. C.; Bonvin, A. M. J. J.; Boelens, R.; Kaptein, R. *J. Magn. Reson* **1991**, *91*, 659−664.

[20] Zhu, L.; Dyson, H. J.; Wright, P. E. *J. Biomol. NMR* **1998**, *11*, 17−29.

[21] Gronwald, W.; Kirchhofer, R.; Gorler, A.; Kremer, W.; Ganslmeier, B.; Neidig, K. P.; Kalbitzer, H. R. *J. Biomol. NMR* **2000**, *17*, 137−151.

[22] Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Morgan Kaufmann: San Francisco, California, 2000.

[23] Hand, D. J.; Mannila, H.; Smyth, P. *Principles of Data Mining*; MIT Press: Cambridge, Massachusetts, 2001.

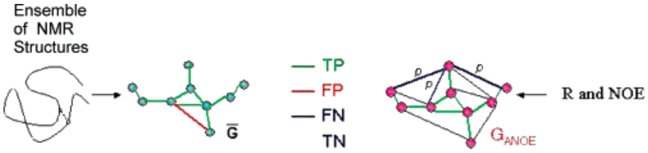

Figure 1. Comparison of distance network $\bar{G}$ generated from an ensemble of 3D query structures and G_{ANOE} generated from input NOE peaklist (NOE) and resonance assignment (R) data. Edges that are present in both $\bar{G}$ and G_{ANOE} are true positives (TP). Edges present in $\bar{G}$, but not in G_{ANOE} are false positives (FP). Edges that are not present in both $\bar{G}$ and G_{ANOE} are true negatives (TN). NOE cross peaks (p) are counted (only once) as false negatives (FN) if corresponding linking edges in G_{ANOE} are not present in $\bar{G}$.

ously linked to more than one proton pair, as indicated by chemical shift degeneracies and match tolerances. The solution network, G_{NOE}, corresponding to the true 3D structure, is a subgraph of G_{ANOE}.

Given complete NOESY peak lists and resonance assignments, for each NOESY cross peak, at least one of its linked proton pairs belongs to G_{NOE}. From an ensemble of query 3D structures, an ensemble-average distance network $\bar{G}$ is then calculated from the sum of inverse sixth powers of individual degenerate proton−proton distances, assuming uniform effects of nuclear relaxation processes (Figure 1). Protons (vertices) are connected (edges) if their corresponding midrange interproton distance in the ensemble of model structures is ≤ d_{NOE_max}, where d_{NOE_max} is the maximum distance detected in the NOESY spectrum. In this approach, the problem of finding a global measure of the goodness-of-fit of the query structures with the NOESY spectra is reduced to comparing the differences of the two graphs $\bar{G}$ (derived from the structures(s)) and G_{ANOE} (derived from the NOESY peak list data).

To provide a statistical measure of the agreement between $\bar{G}$ and G_{ANOE}, we have adopted the *F*-measure metric from information retrieval statistics,[22,23] in which the performance of a search algorithm is assessed by its ability to correctly distinguish “documents” relevant to a particular query from those that are not relevant to the query. The four possible outcomes of a retrieval search are summarized in Table 1. “Relevant” documents retrieved by the algorithm are classified as true positives (TP), while “not-relevant” documents retrieved by the algorithm are false positives (FP). “Relevant” documents not retrieved by an algorithm are false negatives (FN) and “not-relevant” documents that are also not retrieved by an algorithm are true negatives (TN). Recall is defined as the fraction of relevant documents that are retrieved by the algorithm and Precision is defined as the fraction of retrieved documents that are in fact relevant. The *F*-measure characterizes the combined performance of Recall and Precision.

In the context of NOESY-based structure analysis, proton pair interactions (h1, h2) are analogous to “documents”. Observed NOESY cross peaks are defined as true relevant documents, assuming the peak lists (set NOE) have no noise. Potential NOESY peaks not observed in the data are analogous to not-relevant documents, assuming the input data are complete. As illustrated in Figure 1, particular proton pair interactions present in (or “retrieved by”) the atomic coordinates of a model structure may either be represented in the graphical representation of the NOESY peak list data G_{ANOE} (TP), or not represented in G_{ANOE} (FP). Proton pair interactions “not retrieved” by the structure and also not represented in G_{ANOE} are defined as TNs. Proton pair interactions not retrieved by the structure but represented in G_{ANOE} have to be considered carefully with respect to the ambiguous relationship between peaks and their multiple possible assignments. Since G_{ANOE} is an ambiguous network, a FN score is assigned to the peak only if none of the several possible interactions are observed in $\bar{G}$. In this context, Recall (eq 1) measures the fraction of NOE cross peaks that are retrieved by the query structures, while Precision (eq 2) measures the fraction of retrieved proton pair interactions in the query structure that are relevant (in G_{ANOE}), weighted by interproton distance. The upper-bound observed distance, d_{NOE_max}, used in these measures is 5 Å, but can also be calibrated from the NOESY data. Accordingly, the performance score (*F*-measure) of the final ensemble of structures $F(\bar{G})$ is assessed by the following set of statistics:

$$\mathrm{Recall}\ (\bar{G}) = \frac{|\{p|(h1,\ h2,\ p) \in G_{\mathrm{ANOE}},\ (h1,\ h2,\ d) \in \bar{G}\}|}{|\{p|(h1,\ h2,\ p) \in G_{\mathrm{ANOE}}\}|} \quad (1)$$

$$\mathrm{Precision}_w(\bar{G}) = \frac{\displaystyle\sum_{\substack{(h1,h2,d)\in \bar{G},\\ (h1,h2,p)\in G_{\mathrm{ANOE}}}} d(h1,\ h2)^{-6}}{\displaystyle\sum_{(h1,h2,d)\in \bar{G}} d(h1,h2)^{-6}} \quad (2)$$

$$F(\bar{G}) = \frac{2 \times \mathrm{Recall}(\bar{G}) \times \mathrm{Precision}_w(\bar{G})}{\mathrm{Recall}(\bar{G}) + \mathrm{Precision}_w(\bar{G})} \quad (3)$$

In this analysis, a distance (d^{-6}) weighting of the precision metric, $\mathrm{precison}_w(\bar{G})$, is used to reduce the otherwise dominant influence of the many weak NOEs arising from interproton distances close to the upper-bound detection limit, d_{NOE_max}. This weighting also makes the quality scores less sensitive to the value chosen for d_{NOE_max}.

Discriminating Power (DP-score). While the *F*-measure statistic is useful for distinguishing accurate from inaccurate structures, we have found it useful to also report a normalized *F*-measure statistic that accounts for lower-bound and upper-bound values of the *F*-measure that are indicated by the NMR data quality. The lower-bound of $F(\bar{G})$ is estimated by the performance $F(G_{free})$, where G_{free} is a distance network graph computed from interproton distances in a freely rotating polypeptide chain model first described by Flory and co-workers[24,25] (details are presented in Supporting Information). The upper-bound of $F(\bar{G})$ is estimated by $F(G_{ideal})$, where G_{ideal} is the graph of a hypothetical ideal structure that is perfectly consistent with G_{ANOE}. Specifically, G_{ideal} is defined so that recall$(G_{ideal}) = 1$ and precision$(G_{ideal}) =$ precision(G_{local}), where G_{local} is a network of all conformation-independent two- and three-bond connected proton pairs. With these definitions, $F(G_{ideal})$ represents the *best possible performance considering the quality of the input NOESY peak lists and resonance assignments*. $F(G_{ideal})$, and particularly the Precision of G_{ideal}, thus provides a measure of the combined quality of the resonance assignment and NOESY peak lists for one or more spectra. $F(G_{ideal})$ and $F(G_{free})$ describe the two bounds of the performance $F(\bar{G})$; i.e., $F(G_{ideal}) \geq F(\bar{G}) \geq F(G_{free})$. With these definitions, the fold *Discriminating Power* (DP) for $\bar{G}$ is then estimated as:

$$\mathrm{DP}(\bar{G}) = \frac{F(\bar{G}) - F(G_{\mathrm{free}})}{F(G_{\mathrm{ideal}}) - F(G_{\mathrm{free}})} \quad (4)$$

where, $\mathrm{DP}(G_{ideal}) = 1$ and $\mathrm{DP}(G_{free}) = 0$.

The *F*-measure score provides an assessment of the overall fit between the query model structure(s) and the experimental data, assuming that the input data are near complete; the Discriminating Power score, $\mathrm{DP}(\bar{G})$, measures how the query structure is distinguished from the freely rotating chain model.

NMR Datasets. We have validated the sensitivities of NMR RPF scores on experimental NMR data sets of: human basic fibroblast growth factor (FGF-2, 154 a.a.),[26,27] the inhibitor-free catalytic fragment of human fibroblast collagenase (MMP-1, 169 a.a.),[28,29] and human

[24] Flory, P. J. *Statistical Mechanics of Chain Molecules*; Interscience Publishers: New York, 1969.

[25] Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry*; W. H. Freeman: San Francisco, 1980.

[26] Moy, F. J.; Seddon, A. P.; Campbell, E. B.; Bohlen, P.; Powers, R. *J. Biomol. NMR* **1995**, *6*, 245−254.

[27] Moy, F. J.; Seddon, A. P.; Bohlen, P.; Powers, R. *Biochemistry* **1996**, *35*, 13552−13561.

[28] Moy, F. J.; Pisano, M. R.; Chanda, P. K.; Urbano, C.; Killar, L. M.; Sung, M.-L.; Powers, R. *J. Biomol. NMR* **1997**, *10*, 9−19.

Table 1. Recall and Precision Analysis for Information Retrieval and Its Application for Quality Assessment of NMR Structures, Assuming Input Data Are Complete with No Noise

	truth: relevant	truth: not-relevant
algorithm: relevant (retrieved)	TP	FP
algorithm: not-relevant (not retrieved)	FN	TN

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad F\text{-measure} = \frac{2 \times \mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}$$

	peak is observed $\{p \mid (h1,h2,p) \in G_{\mathrm{ANOE}}\}$	peak is not observed $(h1, h2, p) \notin G_{\mathrm{ANOE}}$
interaction retrieved by query structures $(h1,h2,d) \in \bar{G}$	TP	FP
interaction is not retrieved by query structures $(h1,h2,d) \notin \bar{G}$	FN	TN

interleukin-13 (IL-13, 113 a.a.).[30] For each protein, 3D ^{13}C− and ^{15}N−NOESY peak lists (set NOE) and resonance assignments (set R) were used to generate the ambiguous NOE network G_{ANOE}. Atomic coordinates for these three proteins (the Expert I group), determined using the same NOESY peak lists and resonance assignments, were obtained from the Protein Data Bank (PDB): FGF-2 (PDB-ID: 1BLD; a β-fold); MMP-1 (PDB-ID: 1AYK; an α/β fold); IL-13 (PDB-ID: 1IK0; an α fold). For each structure evaluated, a second independently determined 3D structure was also evaluated (the Expert II group), including the 1.9-Å X-ray crystal structure of FGF-2 (PDB-ID: 1BAS),[31] the 1.56-Å X-ray crystal structure of MMP-1 (PDB-ID: 1HFC),[32] and a second solution NMR structure of IL-13 (PDB-ID: 1GA3).[33] In this paper, we also report the NMR RPF scores for quality control in determining the 3D structure of the 100-residue *Escherichia coli* YggU protein, a target of the Northeast Structural Genomics Consortium (*http://www.nesg.org*).

Solution NMR structures and resonance assignments for FGF-2,[26,27] MMP-1,[28,29] and IL-13[30] were described in detail previously, based on manual analysis methods. In this work, we also used the previously unpublished NOESY peak lists. NMR spectra were recorded on a Bruker DRX or AMX 600 spectrometer equipped with a triple-resonance gradient probe. Spectra were processed using the NMRPipe software package [34] and manually peak-picked and analyzed with the software package PIPP.[35] ^{13}C/^{15}N and ^{15}N-enriched samples of FGF-2, IL-13, and MMP-1 were prepared in 90% H_2O/10% D_2O and "100%" D_2O at a 1 mM concentration. FGF-2 NMR spectra were collected at 25 °C in a buffer containing 50 mM potassium phosphate, 2 mM NaN_3, 10 mM deuterated DTT at pH 5.5. IL-13 NMR spectra were collected at 25 °C in a buffer containing 40 mM sodium phosphate, 2 mM NaN_3, 40 mM NaCl at pH 6.0. MMP-1 NMR spectra were collected at 35 °C in a buffer containing 10 mM deuterated Tris-Base, 100 mM NaCl, 5 mM $CaCl_2$, 0.1 mM $ZnCl_2$, 2 mM NaN_3, 10 mM deuterated DTT at pH 6.5. The assignments of the ^{1}H, ^{15}N, ^{13}CO, and ^{13}C resonances were based primarily on the following experiments: CBCA(CO)NH, CBCANH, C(CO)NH, HC(CO)NH, HBHA(CO)NH, HNCO, HCACO, HNHA, HNCA, HCCH−COSY and HCCH−TOCSY.[36] The ^{15}N-edited NOESY and ^{13}C-edited NOESY experiments were collected with 100 ms and 120 ms mixing times, respectively. The structures were calculated using the hybrid distance geometry-dynamical simulated annealing method of Nilges et al. [37] using the program XPLOR.[38,39]

RPF analyses were also carried out using unpublished NMR data for 100-residue *Escherichia coli* protein YggU, a target of the Northeast Structural Genomics Consortium (http://www.nesg.org) with unknown biochemical function. Atomic coordinates for YggU are deposited in the Protein Data Bank (PDB-ID 1YH5), and the structure determination will be presented in detail elsewhere (Aramini & Montelione, in preparation). NMR spectra were recorded on Varian INOVA 500, 600 and 750 MHz spectrometers. Spectra were processed using the NMRPipe software package and manually peak-picked and analyzed with SPARKY.[40] ^{13}C/^{15}N and ^{15}N-labeled samples of YggU were prepared in 95% H_2O/5% D_2O at a 1 mM concentration. NMR spectra were collected at 20 °C in a buffer containing 20 mM MES, 50 mM NaCl, 5 mM DTT at pH 6.5. The assignments of the ^{1}H, ^{15}N, ^{13}CO, and ^{13}C resonances were based on the following experiments: 2D ^{1}H−^{15}N HSQC, 3D HNCO, HN(CO)CACB, HNCACB, HN(CO)CA, HNCA, HA(CA)NH, HA(CACO)NH, 3D (H)CC(CO)NH−TOCSY, H(CCCO)NH−TOCSY, HCCH−COSY, RD $\underline{\mathrm{HC}}$CH−COSY, and 2D $\underline{\mathrm{HBCB}}$(CGCD)HD and H−TOCSY−$\underline{\mathrm{HC}}$H−COSY RD experiments.[41] The ^{15}N-edited NOESY and ^{13}C-edited NOESY experiments were collected with 80 ms and 70 ms mixing times, respectively. NOESY peak lists were interpreted using a fully automated approach[12] and the structures were calculated using the program XPLOR.[38,39]

Generation of Different Incorrect-Fold Structures: 6−12 Å rmsd Range. To test the sensitivity of RPF scores for identifying 3D structures with incorrect folds, we generated sets of different incorrect structures using the homology-modeling tool HOMA.[42] An incorrect β-fold (incorrect fold I) of FGF-2 was generated by modeling with a different beta barrel protein template, cyclophilin isomerase (PDB−ID: 1CLH), in which two of the β-strands form FGF-like interactions, but the rest of the protein structure is significantly different from the correct FGF-2 structure. An incorrect α-fold (incorrect fold II) for FGF-2 was modeled from the 3D structure of myoglobin (PDB-ID: 101M), and an incorrect α/β-fold (incorrect fold III) for FGF-2 was modeled using the coordinates of MMP-1 (PDB-ID: 1AYK). Similarly, an incorrect β-fold (incorrect fold I) of MMP-1 was modeled based on the structure of a beta barrel, *E. coli* cyclophilin isomerase (PDB-ID: 1CLH), an incorrect α-fold (incorrect fold II) of MMP-1 was modeled

[29] Moy, F. J.; Chanda, P. K.; Cosmi, S.; Pisano, M. R.; Urbano, C.; Wilhelm, J.; Powers, R. *Biochemistry* **1998**, *37*, 1495−1504.

[30] Moy, F. J.; Diblasio, E.; Wilhelm, J.; Powers, R. *J. Mol. Biol.* **2001**, *310*, 219−230.

[31] Zhu, X.; Komiya, H.; Chirino, A.; Faham, S.; Fox, G. M.; Arakawa, T.; Hsu, B. T.; Rees, D. C. *Science* **1991**, *251*, 90−93.

[32] Spurlino, J. C.; Smallwood, A. M.; Carlton, D. D.; Banks, T. M.; Vavra, K. J.; Johnson, J. S.; Cook, E. R.; Falvo, J.; Wahl, R. C.; Pulvino, T. A.; et al. *Proteins* **1994**, *19*, 98−109.

[33] Eisenmesser, E. Z.; Horita, D. A.; Altieri, A. S.; Byrd, R. A. *J. Mol. Biol.* **2001**, *310*, 231−241.

[34] Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277−293.

[35] Garrett, D. S.; Powers, R.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson.* **1991**, *95*, 214−230.

[36] Clore, G. M.; Gronenborn, A. M. *Methods Enzymol.* **1994**, *239*, 349−362.

[37] Nilges, M.; Gronenborn, A. M.; Bruenger, A. T.; Clore, G. M. *Protein Eng.* **1988**, *2*, 27−38.

[38] Clore, G. M.; Appella, E.; Yamada, M.; Matsushima, K.; Gronenborn, A. M. *Biochemistry* **1990**, *29*, 1689−1696.

[39] Brunger, A. T. *X-PLOR, Version 3.1: A System for X-ray Crystallography and NMR*; Yale University Press: New Haven, 1992.

[40] Goddard, T. D.; Kneller, D. G. SPARKY 3. University of California: San Francisco, 1999.

[41] Aramini, J. M.; Mills, J. L.; Xiao, R.; Acton, T. B.; Wu, M. J.; Szyperski, T.; Montelione, G. T. *J. Biomol. NMR* **2003**, *27*, 285−286.

[42] Li, H.; Tejero, R.; Monleon, D.; Bassolino-Klimas, D.; Abate-Shen, C.; Bruccoleri, R. E.; Montelione, G. T. *Protein Sci.* **1997**, *6*, 956−970.

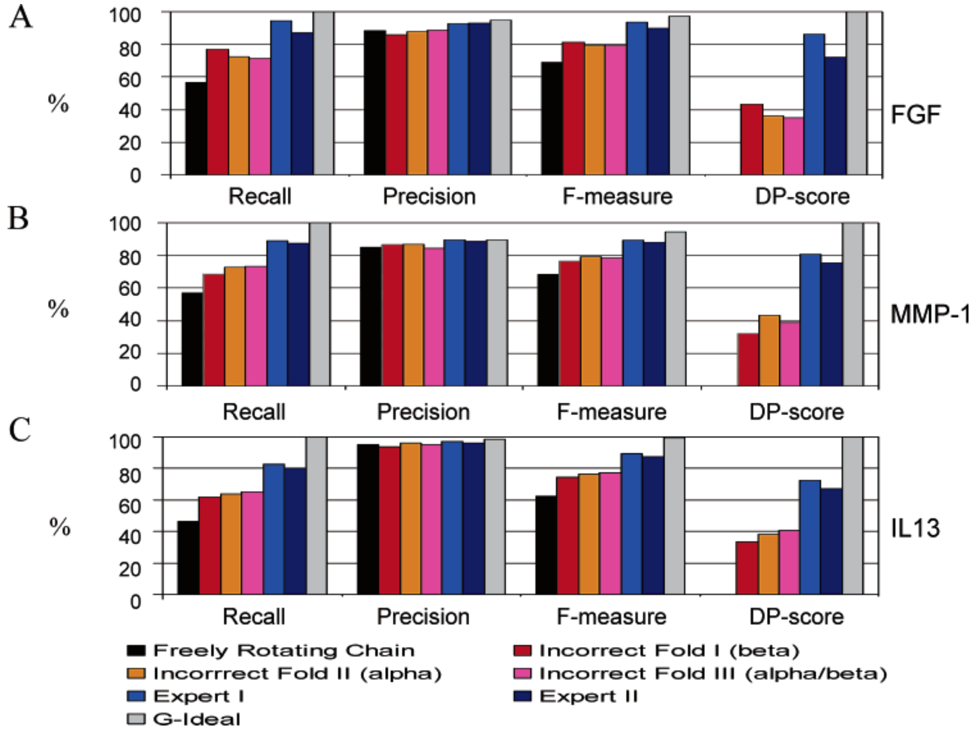

Figure 2. Sensitivity analysis of the quality scores Recall, Precision, *F*-measure and DP for three protein NMR data sets: (A) FGF-2, (B) MMP-1 and (C) IL-13. For each data set and each quality score, the first columns (black) present values computed for a freely rotating polypeptide chain model, as described in text. The second-fourth columns (red, orange, pink) present quality scores of a set of coordinates modeled from different incorrect folds (i.e., β, α, and $\alpha+\beta$ folds). The fifth columns (blue) present quality scores for structures determined by manual NMR structure analysis using the same chemical shift list R and NOE data, and structure generation with XPLOR. The sixth columns (dark blue) present quality scores for structures determined by X-ray crystallography or from independent manual NMR structure determinations. The seventh columns (grey) present quality scores for theoretical "ideal" structures, as defined in the text. The average DP scores of incorrect folds are ~0.38 while the average DP scores of high quality protein structures are ~0.75.

from the structure of sperm whale myoglobin (PDB-ID: 101M) and an incorrect α/β-fold (incorrect fold III) was modeled from the structure of *E. coli* hypothetical protein Ygdk (PDB-ID: 1NI7). For IL-13, an incorrect β-fold (incorrect fold I) was modeled from the 3D structure of the *C. elegans* major sperm protein (PDB-ID: 1M1S), an incorrect α-fold (incorrect fold II) was modeled from the structure of yeast transcription elongation factor S-II (PDB-ID: 1EO0), and an incorrect α/β-fold (incorrect fold III) was modeled from a structure of *E. coli* ribosomal binding factor A (PDB-ID: 1KKG). The rmsd's of all-heavy-atoms between these incorrect folds and the corresponding three experimental NMR structures (defined here as the correct structures) range from 6 to 12 Å.

Generation of Partially Incorrect and Distorted Protein Structures: < 6 Å rmsd Range. To further test the sensitivity of the RPF scores in assessing 3D structures with smaller distortions from the correct structure, we also generated a set of partially incorrect (or slightly distorted) coordinates sets with all-heavy atom rmsd's within 6 Å of the correct NMR structures. This set was generated using either intermediate coordinate sets obtained in the process of automated NMR structure analysis [12] or by modifying the three NMR-reference structures using programs SWISS-Model[43] or MOLMOL.[44] When programs SWISS-Model or MOLMOL were used, incorrect structures were generated by varying dihedral angles of only one or two residues that moved secondary structures apart. Additional incorrect structures were generated by rotation, translation, and incorrect repacking of one or two secondary structure elements. Most of the structural features in these "partially correct" (2−6 Å rmsd's for all-heavy atoms relative to the "correct" NMR structures) and "distorted" (<2 Å all-heavy-atom rmsd) 3D structures are identical with the original reference structures.

Calculation of the NMR RPF Scores. In analyzing RPF scores for ensembles of NMR structures, the first 10 structures in each PDB coordinate file were used. Calculations were carried out on Linux-based Pentium and Athalon processors. Execution times for RPF analysis of the largest proteins in our sample are under two minutes for one 3D structure model on a 1060 MHz Athlon Processor.

Results and Discussion

Discriminating Correct Folds from Incorrect Folds. NMR RPF scores were computed for three experimental NMR datasets: human basic fibroblast growth factor (FGF-2, 154 a.a.),[26,27] the inhibitor-free catalytic fragment of human fibroblast collagenase (MMP-1, 169 a.a.),[28,29] and human interleukin-13 (IL-13, 113 a.a.).[30] Figure 2 illustrates the RPF scores of different model structures compared to the respective input NOE peak lists (NOE) and resonance assignment table (R) for correct and incorrect structures of FGF-2, MMP-1, and IL-13. For each of these three NMR data sets, the first score (black bars) is the quality factor computed based on average interproton distances in a freely rotating polypeptide chain. The second through fourth scores (red, orange, and pink) are quality factors for sets of coordinates with different incorrect folds, (i.e. β, α, and α/β folds), generated for each protein using "homology modeling" methods[42] described in the Materials and Methods section. The fifth score set (blue) measures the quality of structures determined by manual NMR analysis with XPLOR using the same data, and the sixth score set (dark blue) measures the quality of structures determined by X-ray crystallography or from an independently determined NMR structure. The last score (grey) represents the best possible quality score (G_{ideal}), providing an assessment of the quality of the input NOE and *R* data sets.

[43] Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. *Nucleic Acids Res.* **2003**, *31*, 3381−3385.

[44] Koradi, R.; Billeter, M.; Wuthrich, K. *J. Mol. Graph.* **1996**, *14*, 51−55, 29−32.

Recall rates for these data sets clearly distinguish correct from incorrect folds (Figure 2). The Recall rate for the freely rotating chain model (black bars) indicates that short distances arising from intra and sequential proton pairs account for more than 50% of the observed NOESY cross peaks. Within the remaining $<$ 50% of NOESY cross peaks are the subset of data that determine the overall fold. Significantly, structures generated with incorrect folds can satisfy as much as ~50% of these conformationally important NOE cross peaks. Conversely, high-quality structures account for $>$ 80% of the observed NOESY cross peaks. The residual fraction of NOESY peak list data that are not accounted for by the correct structures (<20%) results from several sources, including missing resonance assignments, spectral noise and other spectral artifacts, local structure distortions, inaccuracies in the $d_{\mathrm{NOE\_max}}$ estimate, and the presence of spin-diffusion peaks that are inconsistent with the local structure and distance cutoffs. Despite these inaccuracies, these Recall rates provide clear distinctions between correct and incorrect folds.

Good quality structures should have high Precision rates; i.e., few short interproton distances that do not have corresponding data in the NOESY peak list(s). In addition to inaccuracies in the atomic coordinates, factors contributing to FP information and reducing the Precision score include surface amide proton saturation transfer, solvent exchange broadening, differential nuclear relaxation, and conformational exchange broadening, when these effects are severe enough to cause NOEs arising from short distances to be missing from the spectra. As expected, Precision rates for the freely rotating chain and incorrect-fold models are lower than Precision rates for the set of correctly folded structures (Figure 2). However, Precision rates are not as discriminative as the Recall rates. These data demonstrate that even incorrect folds can be consistent with a large fraction of the NOESY data, especially when there is significant resonance degeneracy. The relatively high Precision rates observed for incorrect folds is also attributed to the domination by less structurally informative but short-distance (strong) intraresidue and sequential NOE interactions over fold-critical but longer-distance (weak) long-range NOE interactions.

The performance score, $F(\bar{G})$, can also be dominated by less-informative NOEs arising from these short-range interactions. However, as can be seen for the three proteins analyzed in Figure 2, DP scores, reflecting the combined information of the Recall and Precision scores and normalized to account for the less-informative local NOE interactions and the data quality, are much more effective in discriminating between correct and incorrect folds. The average DP scores of incorrect folds are ~0.38, while the average DP scores of high-quality protein structures are ~0.75. These results demonstrate the value of DP scores in distinguishing correct from incorrect folds determined by NMR.

Structures determined and scored using the same data have better goodness-of-fit scores than the structures determined by independent groups using different protein samples and data sets. For example, since the X-ray structures of FGF-2 have no reported coordinates for residues 1-27 and 153−155,[31] distances greater than $d_{\mathrm{NOE\_max}}$ are assigned to proton pairs from these residues during the Recall and Precision score calculations. However, some of these proton pairs have NOE interactions that are present in the NOESY peaks lists. The observation that quality scores for the X-ray crystal structure of FGF-2 are somewhat lower than those of NMR structures is attributed to the fact that the NMR structures in fact have defined structures for residues 1−27 and 153−155, which fit to data in the NOESY peak lists. Similarly, X-ray crystal structures of MMP-1 have no coordinates reported for residues 1-6 and 164-169, which do include some NOE data. In addition, the X-ray crystal structure of MMP-1 is complexed with a hydroxamate inhibitor and there are local conformational differences in the active-site region due to interactions with the inhibitor. Since the NMR peak list data are for inhibitor-free MMP-1, the overall *F*-measure and DP quality scores for this NMR structure are slightly higher than those calculated for the ligand-bound X-ray crystal structure. The two IL-13 NMR structures are also slightly different (backbone rmsd difference ≈ 1.0 Å), and IL-13 structures from Expert Group I have slightly better fit to the NOESY data than structures from Expert Group II (Figure 2). As peak list data is not available for the Group II structures it is not possible to determine if indeed the Expert Group I IL-13 structures are also a better fit to the Expert Group II data, or if indeed the underlying data are different.

Comparing the NMR RPF Scores with Structure Accuracy. Assuming the reference structures are accurate interpretations of the corresponding resonance assignment and NOESY data, we have used the rmsd's of all-heavy-atom coordinates between reference and intentionally distorted/incorrect structures as a measure of the accuracy of these incorrect structures. For the FGF-2 data set, all heavy atoms of only residues 29−152 were used for these rmsd calculations. For the MMP-1 data set, only heavy atoms from regions 7−137 and 145−163 were included, and for IL-13 all heavy atoms of all residues were used for rmsd calculations. Figure 3 shows scatter plots of NMR RPF scores for these incorrect structures versus these rmsd measures of structure accuracy. For the set of incorrect/distorted structures generated for each of the three protein NMR data sets considered (represented by data points colored the same in Figure 3), the *F*-measure and DP scores decrease monotonically as structure accuracy is reduced (i.e., as rmsd gets larger). The DP score (Figure 3) is particularly sensitive to this measure of structure accuracy. Structures with small all-heavy-atom rmsd values (high accuracy) have high DP scores, while structures with large rmsd values (lower accuracy) have significantly lower DP scores. The Pearson's correlation coefficients for Recall, Precision, *F*-measure and DP scores versus rmsd are −0.795, −0.459, −0.866, and −0.882, respectively. The fact that these correlations for F and DP scores are significantly better than for Recall or Precision alone demonstrates the value of combining both the Recall and Precision in a performance statistic. All structures within 1.5 Å rmsd of the corresponding "correct" structure have *F*-measure performance scores >0.8 and DP scores >0.70.

Structures in 2−6 Å rmsd range are clearly distinguishable from more accurate structures by lower *F*-measure and DP-scores, although in this accuracy range both of these statistics have larger variations than in the higher or lower accuracy ranges. These variations reflect the fact that most structural features in these partially incorrect structures are identical with those of the reference structures. One of the structures of MMP-1 (6 Å rmsd, green point) in this accuracy range has the lowest DP score, attributed to a particularly low Precision score arising

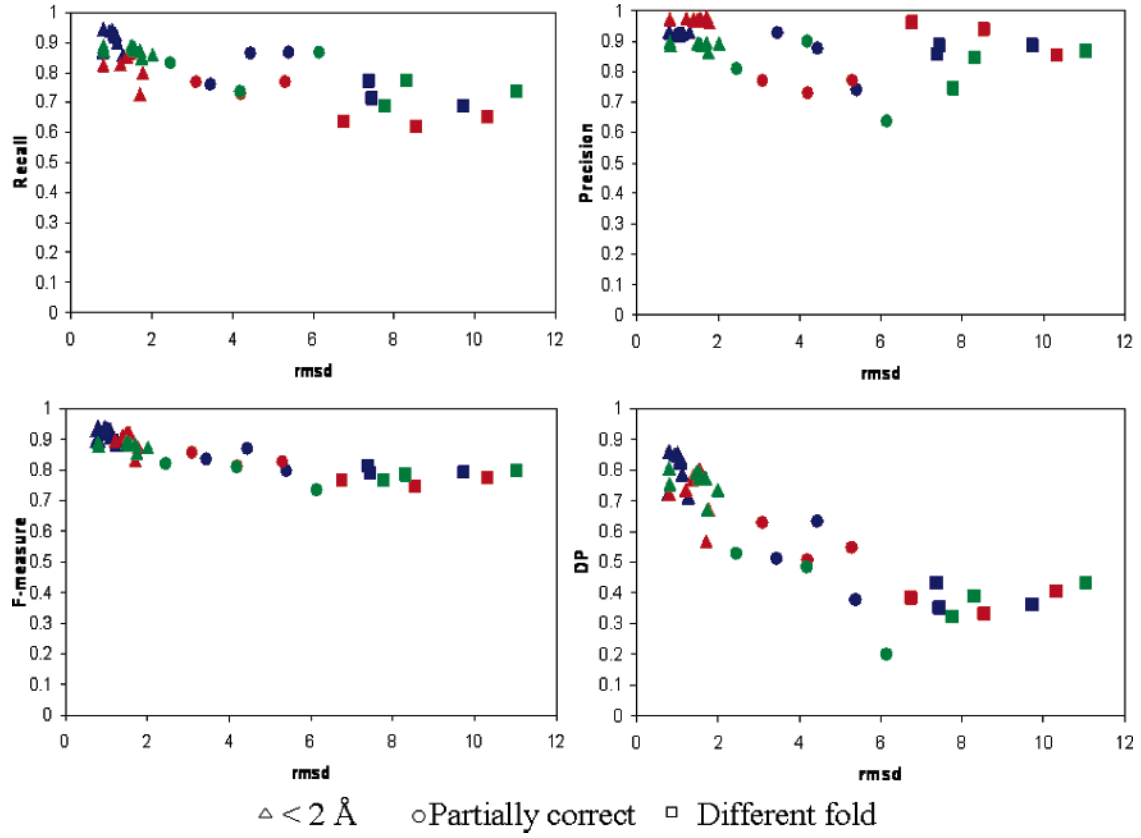

Figure 3. Scatter plots of Recall, Precision, *F*-measure, and DP scores verses all-heavy-atom rmsd values for different structures compared with manually analyzed NMR structures deposited in the Protein Data Bank. Quality scores for structures modeled to be slightly distorted (i.e., with rmsd within 2 Å of the correct structure) are indicated by △. Scores for structures modeled to be partially correct (i.e., with 2−6 Å rmsd to the correct structure) are indicated by ○. Scores for structures "homology modeled" to have completely incorrect folds are indicated by □. Quality scores for FGF-2 (blue), MMP-1 (green) and IL-13 (red) NMR data are consistently higher for the more accurate models.

from many bad contacts in certain regions of the structure, although this structure has a similar high Recall rate as the structures within < 2 Å rmsd accuracy region. This example again demonstrates that it is important to combine both Recall and Precision scores for structure quality assessments. "Slightly distorted structures" in the <2 Å rmsd accuracy range also show good correlations between accuracy and these NMR RPF statistic scores.

Overall, these data demonstrate that the combined analysis of Recall and Precision scores, and particularly the use of a normalized DP score, provides means for distinguishing correct from distorted and partially incorrect structures, particularly for inaccuracies of >2 Å rmsd for all-heavy-atom coordinates.

Sensitivity to Match Tolerances. Match tolerances are key parameters used to calculate the G_{ANOE} graph. We have carried out sensitivity analyses to assess the impact of match tolerance parameters on RPF scores. These results are presented in Supplementary Figure S1. Our studies show that for good quality structures (DP > 0.7), RPF scores are relatively insensitive to match tolerances typically used in structure analysis; i.e., they are relatively insensitive to match tolerances over the range 0.05 to 0.1 ppm in directly or indirectly detected proton dimensions, and from 0.2 to 1.0 ppm in indirectly detected C/N dimensions. For example, for match tolerances ranging from 0.05 to 0.1 ppm in the indirect proton dimension, the *F*-measure scores increase by <∼2% for high-quality protein structures; over the same range of match tolerances in the directly detected proton dimension, *F*-measure scores increase by <∼1%. Using match tolerances in indirect C/N dimensions ranging from 0.2 to 1.0 ppm, the *F*-measure scores increase by only ∼0.4%. For highly inaccurate structures (e.g., with DP score ∼0.4), *F*-measure scores are only a little more sensitive to match tolerance; e.g., they increase by only ∼5% for match tolerance variations ranging from 0.05 to 0.1 ppm. In this work, match tolerances of 0.05 ppm for H and 0.5 ppm for C/N were used to generate the G_{ANOE} graphs.

Quality Control in an Experimental Protein NMR Structure Determination Trajectory. NMR RPF scores can be used as quality control for de novo protein NMR structure determinations, especially with iterative refinement approaches, using either manual or automated analysis. To illustrate the impact of using NMR RPF scores for quality control in a protein structure determination trajectory, sets of NMR RPF scores (Figure 4) were calculated during the course of 10 cycles of automated iterative analysis of the NMR structure of the *E. coli* YggU protein. At cycle 1, the initial fold stage, the ensemble of 10 YggU structures has a low DP-score (i.e., ∼0.46) that indicates a low quality structure, but which is significantly better than a corresponding incorrect fold (<0.4; Figure 2). Over the course of iterative analysis, increases in the Recall, *F*-measure and DP scores indicate that the qualities of the intermediate structures are improved. By the final cycle of iterative NOESY data analysis and structure refinement, the *F*-measure is >0.9 and DP-score is >0.7. The improvement of these scores through the trajectory correlates with improved accuracy and convergence, as measured by the all-heavy-atom-rmsd of each

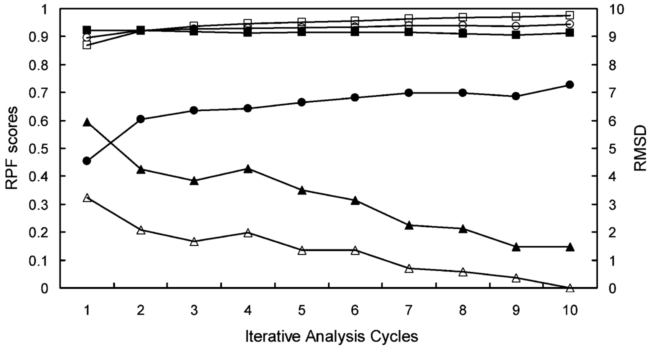

Figure 4. Illustration of the use of RPF scores for quality control of a protein structure determination trajectory using automated NOESY analysis software. Sets of NMR RPF scores (Recall, Precision, *F*-measure, and DP-score, indicated by the symbols □, ■, ○, and ● respectively) were calculated during the course of 10 cycles of automated iterative NOESY spectral analysis and structure determination of *E. coli* protein YggU. Symbols (△) provide a measure of the accuracy of intermediate structures generated in the automated analysis trajectory, computed as the all-heavy-atom rmsd for the mean coordinates of the ensemble of structures determined in each cycle compared with the mean coordinates of the ensemble of structures submitted to the PDB (PDB-ID: 1YH5). Symbols (▲) represent the convergence of coordinates within each intermediate ensemble, computed as all-heavy-atom rmsd to mean atomic coordinates within the computed ensemble of structures. Through the trajectory, there is a good correlation between structure accuracy, convergence within the ensemble, and the Recall, *F*-measure, and DP scores.

intermediate ensemble with the mean atomic coordinates deposited in the Protein Data Bank (1YH5) and the mean all-heavy-atom coordinates within each ensemble, respectively. The final RPF scores for YggU are similar to those obtained with structures that have <~2 Å rmsd for all heavy atoms relative to the refined experimental structure.

During the iterative refinement process, as long as the structure ensemble does not have many bad proton−proton packing interactions, the Precision rate should be high and stay relatively constant. High Precision is a necessary, but not sufficient, criteria for a good quality structure. Figure 4 shows that the Precision rates decrease slightly (~1%) during the iterative refinement processes. This arises as a result of the increased compactness of the structure over the course of the refinement; additional weak NOE cross peaks predicted from the more compact final structures are often missing from the input NOE peak lists. The small decrease in Precision over the course of the refinement is diagnostic of the quality and completeness of the input NMR data.

Graphical Representation of the Distribution of False Positive Errors. False positive structural features are represented by edges that are present in $\bar{G}$ and not in G_{ANOE} (Figure 1). Precision rates measure the fraction of NOE interactions that are predicted from the query structure but missing from the input NMR data. Thus, the higher the number of false-positive structural features, the lower the Precision rate. A particularly valuable feature of these assessment statistics is the ability to visualize false-positive structural features on a per atom basis for the entire protein structure. For this purpose, we have developed a graphical representation tool to display the distribution of false positive errors on the molecular structure. For each false positive edge of residue pair (i,j), a false positive number, FPN, is computed as:

$$\mathrm{FPN} = d^{-6}/2 \qquad (5)$$

where d = the corresponding midrange interproton FP distance in the query structures

The FPN for all false-positive structural features between residues i and j are then summed, and these values are represented graphically on the 3D structure for each residue of the protein. For the four data sets studied here, amide protons tend to give a uniformly distributed number of false positives, which are not useful for identifying inaccurate regions (data not shown). This is attributable to intensity attenuations of surface amide protons due to solvent exchange and conformational exchange broadening. Therefore, FP interactions involving amide protons are not counted for color-coding. After excluding all amide FP interactions, the remaining false positives are generally clustered around spectral regions with low signal-to-noise ratios and/or structure regions with incorrect local structures and/or side-chain packings.

Examples of the graphical representation of false positive distributions computed for the IL-13 NMR data set for three different structures are shown in Figure 5. The reference structure of IL-13 (Figure 5A) has a low number of false positives, and good overall structure quality scores (Recall = 0.825, Precision = 0.971, *F* = 0.892, and DP = 0.723). A decoy structure (Figure 5B), in which the N-terminal helix of IL-13 is pulled away from the core, with all-heavy-atom rmsd compared with the reference structures of 3.1 Å, also has a low number of false positives. The inaccuracy of this decoy structure is not indicated by its Precision score (0.969), but is indicated by comparing other structure quality factors (Recall = 0.769, F = 0.857 and DP = 0.629) with the corresponding values for the reference structure; in particular, this incorrect underpacked structure can be distinguished from the correct structure by its low Recall value and resulting low DP score. A second decoy structure (Figure 5C), in which the N-terminal helix of IL13 has been reorientated and incorrectly repacked, and for which the conformation around the β-sheet region is not correct (all-heavy-atom rmsd to the correct structure of 4.2 Å), has a high number of false positives localized in the inaccurate regions of the structure. In this case, inaccuracy is also indicated by lower global structure quality assessment scores; Recall = 0.729, Precision = 0.917, F = 0.812, and DP = 0.508. This graphical analysis demonstrates that the structural distributions of FPN values are quite useful for identifying inaccurate regions of the structure which do not fit well with the experimental chemical shift and NOESY peak list data. These mappings may also be useful for identifying structural regions which may be inaccurately determined because of conformational exchange broadening and other dynamic effects.

Comparison of Recall and Precision Rates. The FGF-2 and MMP-1 data sets exhibit similar Recall and Precision scores (Figure 2). However, the Precision rate for the IL-13 data set is higher than its Recall rate (Figure 2), while the Recall rate for the final YggU structure is somewhat higher than the corresponding Precision rate. At least part of the explanation for these observations is that there are more “noise peaks” in the IL13 NOESY peak lists, which generally results in a reduced Recall rate. Conversely, higher Recall rate compared with the Precision rate, like the YggU data set, suggests that some weak NOE cross peaks have not been included in the NOESY peak lists because the corresponding signal-to-noise ratios are low. This information is invaluable for evaluating data collection and

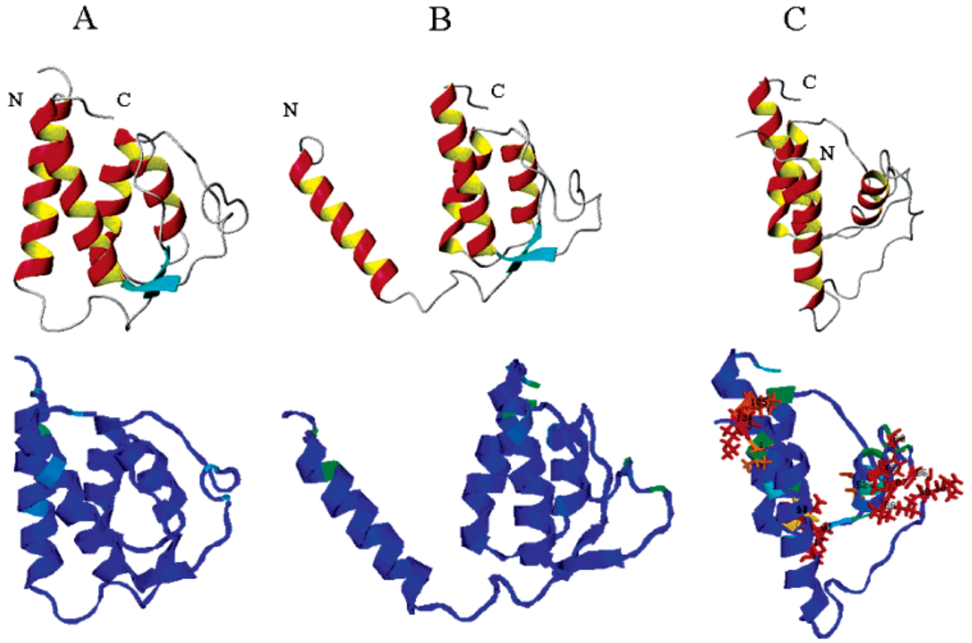

Figure 5. Graphic representations of false positive (FP) distributions on IL-13 structures. False positives correspond to short average distances ($<d_{\mathrm{NOE\_max}}$) in the ensemble of protein NMR structures that are not supported by data in the NOESY peak list. The top panel shows ribbon representations of the corresponding query structures; the bottom panel shows the false positive distributions in these structures. (A) The reference structure of IL-13 (PDB-ID: 1IK0), which has a low number of false positives. (B) A decoy structure, in which the N-terminal helix of IL-13 is pulled away from the core. (C) A second decoy structure, in which the N-terminal helix of IL-13 is repacked in an incorrect orientation and the conformation around the β-sheet region is not correct. False positive numbers (FPNs) are computed as described in the text. Residues of the query structures are color-coded using a spectrum of colors ranging from red, if the summed FPN ≥ the FPN for a residue having six 2.5 Å proton−proton FP interactions; to yellow, if the summed FPN = the FPN for a residue having eight 3.0 Å proton−proton FP interactions; to blue if its FPN ≤ the number for a residue having 10 missing 4.0 Å proton−proton FP interactions. Rasmol[46] is used to display these distributions of false positive errors on the query structure.

analysis strategies for improving the accuracy of a particular protein NMR structure. For example, although low Precision rates can arise from several sources, including inaccurate side-chain packing and attenuating effects of conformational exchange, comparisons of Recall and Precision rates during the course of a structure refinement can help to improve the peak picking process and/or identify errors in the experimental input data, allowing refinement of the input data used in the structure generation process.

Relation between RPF Scores and other Structural Quality Assessment Scores. Structural quality assessment scores, such as packing contacts, dihedral angle distributions, and conformational energies, are valuable tools for protein structure validation,[15−17] comparing observed conformational distributions and packing with values observed in nature and/or expected on first principles. RPF scores measure global goodness-of-fits of NOE peak lists with NMR structures. In general, the goal should be to generate protein structures that score well in these several different and complementary views of structure quality. For example, high RPF scores and high Procheck[45] scores indicate that the structures both fit the data well and have good stereochemical qualities. High RPF scores and slightly lower Procheck scores indicate that the structures fit the data well, but that the data may not be sufficient to define correct local structure, and additional data and/or refinement processes may be required. Importantly, good stereochemical and/or packing scores alone do not necessarily demonstrate that the corresponding structure fits well to the experimental NOE data. Similarly, while the “traditional” NMR quality scores such as distance constraint violations, constraints-per-residue, and convergence across the conformational ensemble (rmsd) are important measures of structure quality, they do not necessarily correlate with goodness-of-fit RPF scores. While rmsd and constraint-per-residue assessments are minimal criteria for good quality structures, neither provides a reliable assessment of the accuracy of the structure, or how well it fits to the experimental data; highly inaccurate structures may exhibit good convergence (low rmsd) with a network of incorrect constraints. Furthermore, while it is critical to compare structures against the constraint lists from which they are generated, these constraint lists are interpretations of NOESY peak lists, while RPF scores directly measure the quality of structures against the NOESY peak list data. For example, Precision has similarities with *NOE Completeness score*;[10] the Precision score measures the completeness of back-calculated peak lists *relative to NOESY peak list data*, while the NOE Completeness score computes the completeness of the back-calculated distance constraints *relative to a derived (and potentially incorrect) constraint list*. While the NOESY peak lists themselves are “derived” information, they are closer to the raw NMR spectral data than constraint lists, which involve much higher levels of interpretation and (sometimes) data omission.

Conclusions

“NMR R-factors” provide a quality measure of the agreement between the experimental and back-calculated NOESY peak lists. Although critical to the development of the field, such analyses have not been routinely used in NMR structure calculations because conventional methods of back calculating

[45] Laskowski, R. A.; Moss, D. S.; Thornton, J. M. *J. Mol. Biol.* **1993**, *231*, 1049−1067.
[46] Sayle, R. A.; Milner-White, E. J. *Trends Biochem. Sci.* **1995**, *20*, 374.

NOESY peak lists are computationally intensive and require significant expertise. Recall, Precision and *F*-measure are types of “NMR R-factor” measurements. Unlike other R-factor assessment scores,[19−21] RPF scores place emphasis on the *presence or absence of distance relationships* as opposed to the exact distance values, and do not require accurate complete relaxation matrix calculations. The *F*-measure score provides an assessment of the overall fit between a query model structure and the experimental data, and the Discriminating Power score, DP, measures how the query structure is distinguished from a freely rotating chain model, accounting for the data quality. Low *F* scores indicate that the query structure does not fit well with the data. A high-quality NMR structure is expected to be well fit to the NMR data (i.e., high *F*-measure score) and have enough long-range contacts to distinguish it from a freely rotating chain model (i.e., high DP scores). High *F* scores and low DP scores indicate that the NMR data does not have enough long-range information to distinguish the structures from a freely rotating chain model. In particular, results presented in this paper demonstrate the value of DP scores in distinguishing correct from incorrect folds determined by NMR. The data also demonstrate that the combined analysis of Recall and Precision scores, and particularly the use of a normalized DP score, provides means for distinguishing correct from distorted and partially incorrect structures, particularly for inaccuracies of > 2 Å rmsd for all-heavy-atom coordinates. We also present a graphical representation tool for analyzing the distribution of false positive errors, which is useful in identifying potentially inaccurate regions of the protein structures, and in providing information useful for NOESY peak list refinement and structure quality assessment.

The RPF scores described here are rapid and easy to compute, as NOESY assignments and complete relaxation matrix calculations are not required. They are therefore well suited for routine use in quality control of NMR structure determinations at different stages of analysis, using either manual or automated analysis methods. They are relatively insensitive to small variations of nuclear relaxation rates throughout the protein structure as they do not use NOESY peak intensity quantitatively, although they are affected by severe nuclear relaxation effects that cause peaks corresponding to short distances to be absent from the NOESY spectra. These NMR RPF scores are particularly valuable for assessing the correctness of a protein fold in the initial stages of automated structure analysis, and in guiding the use of these intermediate structures in making additional NOESY cross-peak assignments. In final refinement stages, the RPF scores can be used together with the false positive distribution analysis to identify inaccurate regions of the protein structures for further refinement, and to compare alternative structures generated from the same NMR data.

Acknowledgment. We thank J. Aramini for collecting and processing NMR data for the *E.coli* YggU protein, and Y. Darian, and R. Tejero for helpful discussions and comments on the manuscript. This work was supported by a grant from the Protein Structure Initiative of the National Institutes of Health (P50 GM62413). The RPF software is available upon request from the authors.

Supporting Information Available: Theory section describes the detail formulations of the NMR RPF scores. Figure S1 illustrates the sensitivity of the *F*-measure and DP scores to changes of match tolerances. This material is available free of charge via the Internet at http://pubs.acs.org.

JA047109H