Structure
Ways & Means

Recommendations of the wwPDB NMR Validation Task Force

Gaetano T. Montelione,[1,2,*] Michael Nilges,[4,5] Ad Bax,[6] Peter Güntert,[8,9] Torsten Herrmann,[10,11] Jane S. Richardson,[12] Charles D. Schwieters,[7] Wim F. Vranken,[13,14] Geerten W. Vuister,[15] David S. Wishart,[16,17] Helen M. Berman,[3] Gerard J. Kleywegt,[18] and John L. Markley[19]

[1]Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry
[2]Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School
[3]Department of Chemistry and Chemical Biology, Center for Integrative Proteomics Research
Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
[4]Département de Biologie Structurale et Chimie, Unité de Bioinformatique Structurale, Institut Pasteur, F-75015 Paris, France
[5]Unité Mixte de Recherche 3258, Centre National de la Recherche Scientifique, F-75015 Paris, France
[6]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases
[7]Division of Computational Bioscience, Center for Information Technology
National Institutes of Health, Bethesda, MD 20892-0520, USA
[8]Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance
[9]Frankfurt Institute of Advanced Studies
Goethe University Frankfurt am Main, Max-von-Laue-Strasse 9, 60438 Frankfurt am Main, Germany
[10]Centre de Résonance Magnétique Nucléaire à Très Hauts Champs, Ecole Normale Supérieure de Lyon, 5 rue de la Doua, 69100 Villeurbanne, France
[11]Institut des Sciences Analytiques, Unité Mixte de Recherche 5280, Centre National de la Recherche Scientifique, 5 rue de la Doua, 69100 Villeurbanne, France
[12]Department of Biochemistry, Duke University, Durham, NC 27710, USA
[13]Department of Structural Biology, Vlaams Instituut voor Biotechnologie, 1050 Brussels, Belgium
[14]Structural Biology Brussels, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[15]Department of Biochemistry, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester, LE1 9HN, UK
[16]Department of Computing Science
[17]Department of Biological Sciences
University of Alberta, Edmonton, AB T6G 2E8, Canada
[18]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[19]Biochemistry Department, University of Wisconsin, 433 Babcock Drive, Madison, WI 53706, USA
*Correspondence: guy@cabm.rutgers.edu
http://dx.doi.org/10.1016/j.str.2013.07.021

SUMMARY

As methods for analysis of biomolecular structure and dynamics using nuclear magnetic resonance spectroscopy (NMR) continue to advance, the resulting 3D structures, chemical shifts, and other NMR data are broadly impacting biology, chemistry, and medicine. Structure model assessment is a critical area of NMR methods development, and is an essential component of the process of making these structures accessible and useful to the wider scientific community. For these reasons, the Worldwide Protein Data Bank (wwPDB) has convened an NMR Validation Task Force (NMR-VTF) to work with wwPDB partners in developing metrics and policies for biomolecular NMR data harvesting, structure representation, and structure quality assessment. This paper summarizes the recommendations of the NMR-VTF, and lays the groundwork for future work in developing standards and metrics for biomolecular NMR structure quality assessment.

The Worldwide Protein Data Bank (wwPDB) (Berman et al., 2003, 2007) has convened several task forces to recommend metrics, standards, and software for biomolecular structure quality assessment. These include task forces providing recommendations for validating biomolecular structures determined by X-ray crystallography (Read et al., 2011), cryo-electron microscopy (Henderson et al., 2012), NMR, and small-angle X-ray or neutron scattering (SAXS/SANS) (Trewhella et al., 2013). The deliberations of these task forces are also important for defining critical research areas in the field of biomolecular structure analysis, and for guiding efforts of researchers developing their own structure validation platforms. Here we present the initial recommendations of the NMR Validation Task Force (NMR-VTF). These recommendations supplement those published by an earlier commission addressing related problems of NMR structure representation and interpretation (Markley et al., 1998).

NMR is now routinely used for determining 3D structures of small (<20 kDa) proteins to high accuracy, often using largely automated methods (Mao et al., 2011; Rosato et al., 2012; Serrano et al., 2012). In favorable cases, structures of proteins as large as 50 kDa or larger can be determined with good accuracy (Lange et al., 2012; Raman et al., 2010). It is critical to the further development of the field to establish metrics and standards for assessing the reliability and accuracy of NMR-derived structures. A wide variety of different data types and methods are used by different groups to generate NMR structures. For this reason, it is important to define unifying and standardized approaches for defining the precision and accuracy of NMR-derived biomolecular structures. This information is essential for appropriate use of these structures in biological research.

Several software packages have been developed that integrate various tools for NMR structure quality assessment (Bhattacharya et al., 2007; Doreleijers et al., 2012; Laskowski et al., 1996; Nabuurs et al., 2003; Spronk et al., 2003; Vriend, 1990; Tejero et al., 2013; Vuister et al., 2013). These analyses can provide useful information to experimentalists for improving the interpretation of NMR data (e.g., allowing more accurate identification of NOESY peaks) and to users of the data by indicating those parts of the structure that can be used to address specific questions about structure and function. Although such software suites are very useful, the field has not yet adopted uniformly accepted metrics and standards for NMR structure quality assessment.

The NMR-VTF recommends that certain well-developed methods and software packages can form a basis for a standardized platform for protein NMR structure assessment. In particular, the NMR-VTF recommends that existing tools developed for validation of X-ray diffraction-derived structures of biomacromolecules and their complexes (Read et al., 2011; Gore et al., 2012) are also appropriate for NMR structures. NMR-specific software tools and metrics that are already broadly adopted by the NMR community can be used together with these knowledge-based geometric validation methods to provide a first-stage NMR structure validation platform.

However, further research and development is required to address additional important issues of NMR structure validation that are needed for a comprehensive platform. These include validation of 3D structures against chemical shift, residual dipolar coupling, chemical shift anisotropy, paramagnetic relaxation enhancement, small-angle X-ray scattering, and NOESY peak list data, as well as assessing the impact of internal dynamics and ensemble-averaging effects on the interpretation of these NMR data. It is recognized that more extensive metrics than those presented here will be required to fully capture the full breadth of all these effects. Further work by the NMR-VTF in consultation with the community is required in order to standardize these additional structure quality metrics so that they can be broadly adopted by the wwPDB on behalf of the biological NMR community. It is very important that the broader scientific community has a voice in the evolution of standards and conventions for validating biomolecular NMR structures. For this reason, we have established an "NMR Community Discussion Site," at http://nmr-community.wwpdb.org.

Principal Recommendations

The NMR-VTF recommends that the depositors of biomolecular NMR structures be encouraged to provide atomic coordinates for all atoms in residues for which backbone and/or side-chain resonance assignments have been determined. This would include large internal loops, interdomain linkers, N- and C-terminal regions, and purification tags, even where the structure in these regions is "ill defined" (i.e., not well converged or not well defined in terms of a unique conformation) by the ensemble. The concept is that every residue for which some experimental data are available should be represented in the atomic coordinates. However, as explained below, ill-defined regions should be specifically identified by the wwPDB through the use of software agreed upon by the community, and these regions should be handled distinctly in the structure quality assessment process. Notwithstanding these recommendations, coordinates for regions of the structure that the depositor feels are not reliable may be excluded from the deposition at the discretion of the depositor.

The NMR-VTF further recommends that the wwPDB implement a standardized structure validation pipeline in three phases.

> Phase 1. Validation metrics that can be implemented easily and immediately by the wwPDB by using existing software.
> Phase 2. Validation metrics for which software/methods are available but that need more assessment before standards and conventions can be defined for the wwPDB.
> Phase 3. Validation methods requiring further research over the coming years.

The NMR-VTF has focused its initial efforts on validation of protein structures determined primarily from NMR data. Further work by the NMR-VTF is needed in order to establish recommendations for validation of nucleic acids, carbohydrates, and other biological structures determined primarily from NMR data.

Phase 1. Validation Metrics that Can Be Implemented Easily by the wwPDB Using Existing Software

Existing software packages can be used to generate validation reports for all submitted protein NMR structures. Software that is freely accessible to the scientific community, and in general use by the biomolecular NMR community, should be used for generating these validation reports. These wwPDB NMR Structure Validation Reports should include four components: (1) a report validating the completeness and global referencing of chemical shift data, independent of 3D structure; (2) analysis of "well-defined" versus "ill-defined" regions; (3) a knowledge-based model validation report; and (4) a restraint-based model-versus-data validation report, comparing each member of the ensemble of NMR models to the available NMR restraints.

These validation reports should also be generated for all NMR structures already in the PDB and distributed by wwPDB member sites. It is recognized that as many of these structures do not have chemical shift data and/or complete restraint data, these validation reports for many of the archived NMR structures will be incomplete. It is also recognized that some structures will have poor validation analyses, often reflecting the early vintage of some of the NMR structures in the archive. These validation reports on the archived NMR structures are nonetheless valuable to the scientific community. Care should also be taken to first identify and exclude from this analysis "averaged atomic coordinates," which do not correspond to physically reasonable models. In the initial phase, these reports will be provided primarily for protein structures, until similar recommendations have been developed for nucleic acids and other biomolecular NMR structures.

In certain cases, spectra may indicate the presence of two or more structures in slow exchange on the chemical shift timescale (e.g., a mixture of reduced and oxidized forms of a protein, or conformations distinguished by slow proline *cis/trans* isomerization). In the event that two or more nontrivially distinct NMR structures are generated for the biomolecule from distinct restraint data, the two (or more) atomic coordinate sets should be deposited and validated separately. An example of two structures that should be deposited and validated separately would be one protein for which different coordinates have been determined for surface loops containing different proline (*cis* versus *trans*) peptide-bond conformations.

Chemical Shift Data Validation Report. All new NMR structures that are deposited in the PDB are now required to include the chemical shift data used to determine the structure. The NMR-VTF recommends that a Phase 1 NMR Structure Validation Report include an analysis of the completeness of these chemical shift data for all assigned atoms, global reference corrections, and a list of chemical shift outliers. Completeness of assignments refers to the percentage of resonances (e.g., ^{1}H, ^{13}C, and ^{15}N) for which assignments are reported, relative to the number of potentially assignable atoms in the full-length protein construct, excluding highly exchangeable protons (e.g., N-terminal and Lys amino and Arg guanido groups, hydroxyl hydrogens of Ser, Thr, and Tyr, and carboxyl hydrogens of Asp and Glu, or the equivalent hydrogen atoms of nucleic acids) and nonprotonated nitrogens and carbons (e.g., Pro N and aromatic Cγ). Backbone carbonyl carbons of peptide bonds shall generally be included in the number of assignable atoms. For the purpose of calculating assignment completeness percentages, the three hydrogens of a methyl group are counted as one atom. If a single chemical shift has been assigned for a diastereotopic pair, this same shift should be reported for both hydrogen or methyl groups, unless it has been experimentally established that it originates from only one of the two diastereotopic partners.

Standardized chemical shift completeness and global referencing reports can be generated by using the same tools that are used by the BioMagResBank (BMRB) (Ulrich et al., 2008), including the assignment validation software suite (AVS) (Moseley et al., 2004), linear analysis of chemical shifts (LACS) (Wang et al., 2005; Wang and Markley, 2009), and SPARTA+ (Shen and Bax, 2010). Additional tools that could be useful for validation of the integrity and accuracy of the chemical shift data, and to identify unusual or outlier chemical shifts, include the probabilistic approach for protein NMR assignment validation (PANAV) (Wang et al., 2010), CheckShift (Ginzinger et al., 2007), SHIFTX2 (Han et al., 2011), and validation of archived chemical shifts through atomic coordinates (VASCO) (Rieping and Vranken, 2010). The output of an appropriate subset of these tools could be combined into a single consistent chemical shift data validation report. As discussed below, it is premature to include in this phase 1 report a validation of 3D structure models on the basis of chemical shift data.

Well-Defined versus Ill-Defined Atoms or Residue Ranges. NMR structure validation methods are generally applicable only to the well-defined regions of the macromolecular structure. Atoms that are not well defined in their atomic positions by the experimental NMR data should not be included as part of the global NMR structure validation. However, such ill-defined regions of the structure may still be useful for expert applications, and models for these regions generally will also be included in the atomic coordinate file. Users of protein NMR structure models need to be made aware of which atoms in the PDB coordinate file are well defined in the NMR structure. For these reasons, it is important that NMR structure coordinates are flagged in a way that identifies well-defined and ill-defined residues and atoms.

Solution NMR structures typically are represented as "ensembles" of coordinate sets. Each member of the ensemble represents a single model that is consistent with the experimental data. The distribution of models across the ensemble provides insight into how well defined the structure is in different regions. Well-defined regions are those that are precisely (although not necessarily accurately) modeled across the ensemble. The ill-defined parts of the structure may correspond to regions of a molecule undergoing conformational dynamics, or may simply reflect incompleteness of the restraining data.

The ensemble representation of molecular models is sometimes confusing to biologists attempting to use an NMR structure. Although each model in the ensemble is considered to be a valid representation of the structure, the uncertainty in these atomic coordinates is commonly assessed by statistical analysis across the ensemble. However, for various reasons, the ensemble representation does not provide a statistically sound estimate of the precision of the atomic coordinates given uncertainties of the experimental data (Andrec et al., 2007; Clore et al., 1993; Snyder et al., 2005; Snyder and Montelione, 2005; Spronk et al., 2003), nor does it provide a true estimate of conformational dynamics. Nonetheless, the ensemble representation is the current convention of the field for distinguishing those regions of a structure that are well defined by the experimental data from those that are ill defined. This distinction is critical for appropriate quality assessment of NMR structures. Accordingly, it is important that the ensemble information is conveyed in a simple way to users of NMR structures.

Chemical shifts, residual dipolar couplings, relaxation rates, and other NMR parameters that provide structural information may be associated with ill-defined regions. Ill-defined regions may also include transient structural information that is functionally important. For these reasons, it is recommended that atomic coordinates be provided for all residues for which chemical shift data are available, even though these coordinates may be imprecisely defined by the experimental data.

It is recognized that such ill-defined regions, which may be flexibly disordered, are often biologically and/or biophysically important, particularly in determining the biochemical functions of macromolecules (Dyson and Wright, 2005; Dunker et al., 2008). NMR can provide unique information about amplitudes and timescales of dynamic fluctuations, which often contribute to biomolecular function. These are important considerations for the NMR-VTF in phases 2 and 3, as outlined below. Although the convention of designating residues or atoms as "ill defined" is helpful for users of biomolecular structures, the terminology "ill defined" should in no way be interpreted as devaluing the significance of these regions of the structure. Considering that flexibly disordered regions of biomolecular structures are important structural and functional features, and that methods for interpreting ensemble-averaged information in these regions of the structure are still under development, the NMR-VTF recommends that the standard validation report include plots of backbone and side-chain circular variance versus residue number for all residues for which atomic coordinates are provided.

Depositor Specification of Well-Defined and Ill-Defined Regions. The NMR-VTF recommends that the wwPDB allow depositors to specify regions of the biomolecular structure that are well defined across the ensemble of NMR structures and those that are ill defined. Tags for such designators have already been developed by the wwPDB and are ready to be implemented as part of PDB depositions.

Automated Analysis of Well-Defined and Ill-Defined Regions. Although the initial designation of ill-defined regions can be provided by depositors, for the purposes of uniform structure quality assessment, the wwPDB should adopt an automated method for defining those parts of the NMR structure that are well defined and ill defined. Several algorithms and software packages are available to make these assessments automatically. These include methods based on (1) the locations of elements of secondary structure, (2) backbone dihedral angle circular variance (Hyberts et al., 1992), and (3) variance matrix analysis (Brünger et al., 1993; Kelley et al., 1996, 1997; Kirchner and Güntert, 2011; Snyder and Montelione, 2005), including methods that use maximum-likelihood superimposition based on principal components analysis (Theobald and Wuttke, 2006, 2008). Definitions based on locations of elements of secondary structure exclude irregular structures in proteins that may, in fact, be well defined. Methods based on backbone dihedral angle circular variance (Hyberts et al., 1992) are very popular in the protein NMR community, but do not provide information about long-range order; that is, they cannot assess how well defined subdomains of the structure are with respect to one another.

The NMR-VTF recommends that the wwPDB adopt one of the several software packages for discriminating between well-defined and ill-defined regions of the protein structure. The method adopted should include the ability to distinguish multiple well-defined regions or "domains" that are not well defined with respect to one another. Examples of these would include two domains of well-defined atoms connected by an ill-defined linker, or a well-defined domain and independent well-defined helix, connected by an ill-defined linker. In these cases, each of the corresponding subdomains can generally be identified by distance variance matrix methods, and should be assessed separately. The software package recommended for this analysis is CYRANGE (Kirchner and Güntert, 2011); other similar software tools have also been described in the literature (Brünger et al., 1993; Kelley et al., 1996, 1997; Snyder and Montelione, 2005) and may also be suitable for this purpose.

Representative NMR Structure. It is recognized that the user community requires the designation of one NMR model from the calculated ensemble, or derived from the ensemble, that is a single representative of the solution structure. The NMR-VTF recommends that the PDB identify the medoid model (Struyf et al., 1997; Snyder et al., 2005; Tejero et al., 2013) that is most similar to all the other conformers (i.e., the model in the ensemble with the smallest average root-mean-square deviation (rmsd) between it and all [other] models of the ensemble), and designate it as the single representative NMR structure. The medoid model should be identified using only the well-defined residue range(s). It can be computed using the algorithm described by Tejero et al. (2013). For NMR structures containing multiple domains that are ill defined with respect to one another, the representative model should be chosen using this approach for the largest domain. If the domains are of identical size, the representative multidomain structure should be selected as the one containing the domain resulting in the smallest rmsd. In addition, the depositor may identify a depositor-designated representative structure as part of the deposition process, based on alternative criteria to be provided at the time of deposition. The PDB might annotate the "medoid representative structure" and "depositor-designated representative structure" in order to facilitate their use.

The "representative model" should also be annotated to indicate which residues and/or atoms are well defined and which are ill defined in the NMR ensemble, either on the basis of the depositor-defined or the automatically generated designations as outlined in the previous sections. Specifically, the information about atoms or residues being well or ill defined should go into the PDB file of the structure and be distributed by the wwPDB so as to be readily available to users and external software. These annotations may be used by visualization programs to color code or exclude ill-defined regions when displaying the representative model(s). Such annotations will be valuable to users of NMR structure coordinates, particularly users who are not familiar with interpreting traditional ensemble representations.

Knowledge-Based Protein Structure Validation. It is the consensus of the NMR-VTF that knowledge-based model validation of protein NMR structures, including covalent geometry, dihedral angle conformations, and core packing, should utilize the same methods, software, and standards as those recommended for the model validation of protein structures determined by X-ray crystallography (Read et al., 2011). In particular, MolProbity software (Chen et al., 2010) should be used for analysis of overpacking (e.g., all-atom steric clashes) and RosettaHoles software (Sheffler and Baker, 2009) for analysis of underpacking. Ramachandran backbone dihedral analysis should utilize recently updated parameters (Arendall et al., 2005; Read et al., 2011).

Knowledge-based validation should be carried out on either the automated or depositor-specified well-defined regions of the structure outlined above. Global *Z* scores or percentiles, which may be plotted as bar graphs as proposed for X-ray crystal structures (Read et al., 2011; Gore et al., 2012), should be reported only for well-defined regions of the structure, and should be graded using the same set of structures used in grading X-ray crystal structures (Gore et al., 2012). In particular, users should be able to compare structures determined by X-ray crystallography and by NMR using metrics and scales that are common to X-ray and NMR structures.

Specifically, the NMR-VTF recommends that knowledge-based validation scores be reported on two scales: (1) relative to the entire protein crystal structure archive of the PDB (i.e., the same reference structures used for assessment of X-ray crystal structures), and (2) relative to the NMR structure archive of the PDB. However, implementers are encouraged to consider alternate basis sets of structures to use in determining such assessment statistics. Scores should be reported as first quartile, mean, and third quartile.

In addition, knowledge-based validations should be reported for each residue of the structure. For such local model structure validation, it is recommended to consider residues in both the well-defined and ill-defined regions of the structure. By analogy with X-ray structures, for which local structural information is graded by the wwPDB by comparison with crystal structures refined to similar diffraction resolutions, for NMR structures this local structural information should be graded based on the entire database of NMR structures. Although ill-defined regions of the structure may or may not have energetically reasonable conformations, depositors should be encouraged to model these regions with plausible conformations. However, the final decision regarding how to model regions of the structure that are underconstrained by the experimental data should be left to the experimentalists who have determined the NMR structure.

Validation of the Consistency between Experimental Restraints and Structural Models. NMR structures also should be validated against distance and dihedral restraint data that are submitted as part of a PDB deposition. In phase 1, the NMR-VTF recommends a simple model-versus-data validation of the structure against only the submitted experimental restraints. These should include (1) distance restraints, (2) hydrogen-bond restraints, (3) dihedral angle restraints, and (4) any additional distance restraints provided with the PDB deposition. These restraint data should be compared with the coordinates of each model to determine restraint violations by each model. Nuclear Overhauser effect (NOE)-based distance-restraint violations should be interpreted with the assumption of r^{-6} summation for ambiguous restraints (Nilges, 1995). The numbers of intraresidue ($i = j$), sequential ($|i - j| = 1$), medium-range ($1 < |i - j| < 5$), long-range ($|i - j| \geq 5$), and interchain restraints should be summarized, together with the number of restraints in each category (NOE-based, hydrogen-bond, dihedral angle, etc.). The number of scalar coupling, residual dipolar coupling, chemical shift anisotropy, paramagnetic relaxation enhancement, and other restraint data should also be summarized. The numbers of restraint violations, in each class, should be reported in bins (e.g., 0–0.2 Å, 0.2–0.5 Å, >0.5 Å), along with the values of the largest restraint violations in each restraint class. If appropriate, such NMR specific metrics could be graded by comparison against the corresponding values observed in all NMR structures in the PDB for which such restraint data are available. These data should be summarized for all the models in the ensemble in a concise format, as well as for the individual models.

Standardized NMR Structure Validation Report. A standard wwPDB NMR structure validation report should be developed. The committee recommends that initially only a core set of standardized validation metrics be adopted. The report should include a summary of the completeness of the chemical shift data, including a summary of unusual chemical shift values, along with a validation of the NMR structure models using knowledge-based and restraint violation statistics. The report should include a version number, along with raw scores generated by the underlying knowledge-based structure validation software. It should also include machine-readable output. These would be expanded over time, as the NMR-VTF assesses and recommends more sophisticated model-versus-data metrics.

Useful models of such reports are provided by the protein structure validation software suite (PSVS) (Bhattacharya et al., 2007), CING software package (Doreleijers et al., 2012), and PDBStat software (Tejero et al., 2013). A recently published survey of NMR structure validation software (Vuister et al., 2013) also provides useful guidance for the development of NMR structure quality assessment reports. An example of an NMR Structure Validation Report for Phase 1, including chemical shift completeness statistics, restraint violation summaries and statistics, and knowledge-based structure validation statistics, taken from a recent paper (Aramini et al., 2012) is presented in Table 1. This example is provided only as a guide to the kind of concise summary that the wwPDB might include in their validation reports. Additional information, such as chemical shift validation statistics (Moseley et al., 2004), could also be provided. Appropriate criteria will need to be developed for structures refined from NOESY-derived distance restraints that do not specify upper or lower bounds (Nilges, 1995). The X-ray crystal structure validation reports described by Gore et al. (2012) also provide useful examples to guide the design of a concise wwPDB NMR validation report. In addition, more extensive NMR structure validation data and graphical assessment tools, similar to those provided for X-ray crystal structure validation reports (Read et al., 2011), should be provided.

Phase 2. Methods and Software Exist but Require Additional Assessment before Adopting Standard Validation Conventions

A critical task for the NMR-VTF is to continue to assess model-versus-data validation metrics that can be used to validate the degree to which 3D NMR structures fit the underlying experimental data; that is, "NMR R factors." These model-versus-data metrics could include assessment of scalar coupling, residual dipolar coupling (RDC), chemical shift anisotropy (CSA), unassigned NOESY peak list, paramagnetic resonance enhancement (PRE), paramagnetic pseudocontact shift, solid-state dipolar coupling, and SAXS or SANS data. Several tools for validating structures against these data are available, including methods for validation of protein structures against RDC data (Bryson et al., 2008; Clore et al., 1993; Valafar and Prestegard, 2004), CSA data (Cornilescu et al., 1998), NOESY peak lists (Bagaria et al., 2012; Huang et al., 2005, 2012; Nilges, 1995), and SAXS data (Grishaev et al., 2005).

Although these methods are very powerful and generally robust, they have not yet been uniformly adopted across the biomolecular NMR community. Metrics based on these data require clear definitions and further assessment, as well as a process for harvesting these data by the wwPDB in an appropriate format for validation. For these reasons, the NMR-VTF does not recommend including these model-versus-data metrics in standard wwPDB validation reports in phase 1.

During phase 2, the NMR-VTF will assess and then recommend the software packages most suited to model-versus-data validation. In order to provide an expanded NMR Structure Validation Report in Phase 2, with additional model-versus-data assessments, depositors of biomolecular NMR structures are encouraged to archive (where available) in the BioMagResBank (Ulrich et al., 2008) NOESY peak lists, RDC, PRE, and SAXS or SANS experimental data, as well as unprocessed free induction decay data, for biomolecular structures deposited in the PDB.

Table 1. Example of a Table Providing a Summary of Structural Statistics Developed and Based on The Recommendations of the NMR-VTF for Bacterial Protein Alr2454

	Alr2454[a]
Completeness of resonance assignments (%)[b]	
Backbone	99.4
Side chain	98.3
Aromatic	96.6
Stereospecific methyl	100
Conformationally restricting restraints[c]	
Distance restraints	
Total	2,478
Intraresidue ($i = j$)	688
Sequential ($\|i - j\| = 1$)	619
Medium range ($1 < \|i - j\| < 5$)	462
Long range ($\|i - j\| \geq 5$)	709
Dihedral angle restraints	162
Hydrogen-bond restraints	0
Disulfide restraints	0
No. of restraints per residue	25.5
No. of long-range restraints per residue	6.8
Residual restraint violations[c]	
Average no. of distance violations per structure	
0.1–0.2 Å	8.75
0.2–0.5 Å	1.85 (0.35 max)
>0.5 Å	0
Average no. of dihedral angle violations per structure	
1–10°	8.75 (3.8 max)
>10°	0
Model quality[c]	
Rmsd backbone atoms (Å)[d]	0.6
Rmsd heavy atoms (Å)[d]	0.9
Rmsd bond lengths (Å)	0.018
Rmsd bond angles (°)	1.1
MolProbity Ramachandran statistics[c],[d]	
Most favored regions (%)	96.8
Allowed regions (%)	3.1
Disallowed regions (%)	0.1
Global quality scores (raw/Z score)[c]	
Verify3D	0.40/−0.96
Prosall	0.66/0.04
PROCHECK (ϕ-ψ)[d]	−0.15/−0.28
PROCHECK (all)[d]	−0.03/−0.18
MolProbity clash score	12.51/−0.62
Model contents	
Ordered residue ranges[d]	1–100
Total no. of residues	108
BMRB accession number	17965
PDB ID code	2LJW[a]

[a] Structural statistics computed for the ensemble of 20 deposited structures.

[b] Computed using AVS software (Moseley et al., 2004) from the expected number of resonances, excluding highly exchangeable protons (N-terminal, Lys, amino and Arg guanido groups, hydroxyls of Ser, Thr, and Tyr), carboxyls of Asp and Glu, nonprotonated aromatic carbons, and the C-terminal His_6 tag.

[c] Calculated using PSVS version 1.4 (Bhattacharya et al., 2007). Average distance violations were calculated using the sum over r^{-6}.

[d] Based on ordered residue ranges [$S(\phi) + S(\psi) > 1.8$].

The NMR-VTF also recognizes the value of biomolecular structures that have been deposited in the PDB and subsequently found to include some inaccuracies as valuable test data sets useful for the development of structure validation methods. Coordinates that have been designated by depositors as ‘‘obsolete’’ that are archived in the PDB are also valuable for testing and developing structure validation tools. The NMR-VTF recommends that a set of such ‘‘inaccurate NMR structure coordinates’’ is collected and provided to the community for methods development.

Phase 3. Areas Requiring Additional Research

The NMR-VTF has identified the validation of NMR structures of polynucleic acids, including DNA and RNA, and polysaccharides as critical areas that require additional research. Although it is likely that some of the same tools used for validating NMR structures of proteins and X-ray crystal structures of nucleic acids will be appropriate for NMR structures of nucleic acids and polysaccharides, the NMR-VTF has agreed to make standardization of metrics for nucleic acid NMR structure validation a future priority of the committee.

A key metric requiring further research is the validation of structures in terms of the experimental information content of the data on which they are based. This ‘‘information content measure’’ would be analogous to the ‘‘resolution’’ measure so central for X-ray crystal structures. For NMR, there could potentially be both a global and a local version of such a measure. It was generally agreed by the NMR-VTF that the metric of ‘‘restraints per residue,’’ although in the spirit of such an information content measure, is not satisfactory, because different restraints have different information content. In particular, the restraints per residue metric do not correlate well with structural accuracy. This is an important area of research.

Chemical shift data can also be used for validation of 3D biomolecular structures (Han et al., 2011; Rieping and Vranken, 2010; Shen and Bax, 2010). This is a significant motivation for capturing chemical shift data for all protein and nucleic acid structures deposited in the PDB. However, chemical shifts are dominated by local effects and hence need to be combined with other data sensitive to longer-range structural features as part of a comprehensive model-versus-data quality assessment. Although advances have been made in this high-impact area of computational NMR, additional research is needed before standardized methods for validating structures directly against chemical shifts can be recommended for inclusion in the wwPDB NMR validation pipeline.

The NMR-VTF also recognized that biomolecular NMR data generally are an ensemble average, with Boltzmann-weighted contributions from the various conformers present in the sample. Accordingly, the NMR data may not be best fit by a single conformer. For this reason, it is critical to develop tools that can be used to assess to what degree the lack of precision in defining atomic coordinates is due to such underlying internal dynamics as can be assessed experimentally by nuclear relaxation, chemical shift, dipolar coupling, and/or residual dipolar coupling data. Methods for generating ensembles of conformers that best satisfy the experimental data (e.g., Clore and Schwieters, 2004; Lindorff-Larsen et al., 2005), particularly in highly dynamic regions of a structure, and validation of these ensembles of conformers against the ensemble-averaged data, are also an important area for future research.

Conclusions

There is no a priori reason to believe that biomolecular structures determined by NMR in solution or the solid state are fundamentally different from those determined by X-ray crystallography, even though intermolecular packing effects in the crystal lattice may stabilize local conformations that are not predominant in solution. For this reason, the knowledge-based validation of NMR structures should be done using the same metrics and standards, and scaled against the same or comparable structural data sets, as has been recommended for X-ray crystal structures (Read et al., 2011). As there is no generally accepted information content measure in NMR similar to a resolution, these knowledge-based statistics should be reported relative to (1) all crystal structures in the archive and (2) all NMR structures in the archive.

Model-versus-data validation of NMR structures is critical for the maturation of the field of biomolecular NMR. However, the recommendation of consensus statistics for model-versus-data validation (i.e., NMR R factors) is complicated by the fact that NMR structures are often derived from a large number of different kinds of NMR data types. Quality assessment is simplified in these initial recommendations by focusing on restraint violation analyses. However, the restraints are interpreted data, which may not capture all of the information present in NOESY spectra and other NMR data sets. Although methods are available to assess models against all these kinds of experimental data, more work is needed to define standards and metrics before incorporating these metrics into a wwPDB NMR validation pipeline. Hence, additional work will be needed to develop standards and methods for a comprehensive model-versus-data assessment.

Considering these caveats, software is available today to generate a useful and extensive Phase 1 wwPDB NMR Structure Validation Report. This report will include chemical shift data validation (completeness and outliers), assessment of "well-defined" and "ill-defined" regions of the structure, knowledge-based validation of well-defined regions, and a comprehensive validation of the structure against restraint data. Such reports will provide valuable information on the precision and accuracy of NMR structures useful for guiding biological research.

ACKNOWLEDGMENTS

We thank the following people for participation in the discussions of the wwPDB NMR-VTF and for their comments on these recommendations: J. Aramini, J. Block, R.A. Byrd, A. Gutmanas, P.M.S. Hendrickx, Y.P. Huang, N. Kobayashi, C. Lawson, H. Nakamura, R.J. Read, A. Rosato, D. Snyder, R. Tejero, E.L. Ulrich, and J. Westbrook. Support for this work was provided by members of the Worldwide PDB: RCSB PDB (NSF DBI 0829586); PDBe (Wellcome Trust 075968 and 088944; BBSRC BB/E007511/1); PDBj (NBDC-JST); BMRB (NLM NIH P41 LM05799); and The Pasteur Institute and NIGMS Protein Structure Initiative grant U54 GM094597 (to G.T.M.). Funding was also provided by NIH Intramural Research Programs of the NIDDK and CIT (to A.B. and C.S.), BBSRC grants BB/J007471/1 and BB/J007897/1 (to G.J.K. and G.W.V., respectively), NIH grant R01-GM073930 (to J.S.R.), and Brussels Institute for Research and Innovation (Innoviris) grant BB2B 2010-1-12 (to W.F.V.).

Received: February 19, 2013
Revised: July 19, 2013
Accepted: July 29, 2013
Published: September 3, 2013

REFERENCES

Andrec, M., Snyder, D.A., Zhou, Z., Young, J., Montelione, G.T., and Levy, R.M. (2007). A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. Proteins *69*, 449–465.

Aramini, J.M., Petrey, D., Lee, D.Y., Janjua, H., Xiao, R., Acton, T.B., Everett, J.K., and Montelione, G.T. (2012). Solution NMR structure of Alr2454 from Nostoc sp. PCC 7120, the first structural representative of Pfam domain family PF11267. J. Struct. Funct. Genomics *13*, 171–176.

Arendall, W.B., III, Tempel, W., Richardson, J.S., Zhou, W., Wang, S., Davis, I.W., Liu, Z.J., Rose, J.P., Carson, W.M., Luo, M., et al. (2005). A test of enhancing model accuracy in high-throughput crystallography. J. Struct. Funct. Genomics *6*, 1–11.

Bagaria, A., Jaravine, V., Huang, Y.J., Montelione, G.T., and Güntert, P. (2012). Protein structure validation by generalized linear model root-mean-square deviation prediction. Protein Sci. *21*, 229–238.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980.

Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. *35*, D301–D303.

Bhattacharya, A., Tejero, R., and Montelione, G.T. (2007). Evaluating protein structures determined by structural genomics consortia. Proteins *66*, 778–795.

Brünger, A.T., Clore, G.M., Gronenborn, A.M., Saffrich, R., and Nilges, M. (1993). Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. Science *261*, 328–331.

Bryson, M., Tian, F., Prestegard, J.H., and Valafar, H. (2008). REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data. J. Magn. Reson. *191*, 322–334.

Chen, V.B., Arendall, W.B., III, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *66*, 12–21.

Clore, G.M., and Schwieters, C.D. (2004). How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? J. Am. Chem. Soc. *126*, 2923–2938.

Clore, G.M., Robien, M.A., and Gronenborn, A.M. (1993). Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. J. Mol. Biol. *231*, 82–102.

Cornilescu, G., Marquardt, J.L., Ottiger, M., and Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J. Am. Chem. Soc. *120*, 6836–6837.

Doreleijers, J.F., Sousa da Silva, A.W., Krieger, E., Nabuurs, S.B., Spronk, C.A., Stevens, T.J., Vranken, W.F., Vriend, G., and Vuister, G.W. (2012). CING: an integrated residue-based structure validation program suite. J. Biomol. NMR *54*, 267–283.

Dunker, A.K., Silman, I., Uversky, V.N., and Sussman, J.L. (2008). Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. *18*, 756–764.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Ginzinger, S.W., Gerick, F., Coles, M., and Heun, V. (2007). CheckShift: automatic correction of inconsistent chemical shift referencing. J. Biomol. NMR *39*, 223–227.

Gore, S., Velankar, S., and Kleywegt, G.J. (2012). Implementing an X-ray validation pipeline for the Protein Data Bank. Acta Crystallogr. D Biol. Crystallogr. *68*, 478–483.

Grishaev, A., Wu, J., Trewhella, J., and Bax, A. (2005). Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. J. Am. Chem. Soc. *127*, 16621–16628.

Han, B., Liu, Y., Ginzinger, S.W., and Wishart, D.S. (2011). SHIFTX2: significantly improved protein chemical shift prediction. J. Biomol. NMR *50*, 43–57.

Henderson, R., Sali, A., Baker, M.L., Carragher, B., Devkota, B., Downing, K.H., Egelman, E.H., Feng, Z., Frank, J., Grigorieff, N., et al. (2012). Outcome of the first Electron Microscopy Validation Task Force meeting. Structure *20*, 205–214.

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

Huang, Y.J., Rosato, A., Singh, G., and Montelione, G.T. (2012). RPF: a quality assessment tool for protein NMR structures. Nucleic Acids Res. *40*, W542–W546.

Hyberts, S.G., Goldberg, M.S., Havel, T.F., and Wagner, G. (1992). The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. Protein Sci. *1*, 736–751.

Kelley, L.A., Gardner, S.P., and Sutcliffe, M.J. (1996). An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. Protein Eng. *9*, 1063–1065.

Kelley, L.A., Gardner, S.P., and Sutcliffe, M.J. (1997). An automated approach for defining core atoms and domains in an ensemble of NMR-derived protein structures. Protein Eng. *10*, 737–741.

Kirchner, D.K., and Güntert, P. (2011). Objective identification of residue ranges for the superposition of protein structures. BMC Bioinformatics *12*, 170.

Lange, O.F., Rossi, P., Sgourakis, N.G., Song, Y., Lee, H.W., Aramini, J.M., Ertekin, A., Xiao, R., Acton, T.B., Montelione, G.T., and Baker, D. (2012). Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc. Natl. Acad. Sci. USA *109*, 10873–10878.

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

Lindorff-Larsen, K., Best, R.B., Depristo, M.A., Dobson, C.M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. Nature *433*, 128–132.

Mao, B., Guan, R., and Montelione, G.T. (2011). Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure *19*, 757–766.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E., and Wüthrich, K. (1998). Recommendations for the presentation of NMR structures of proteins and nucleic acids. Pure Appl. Chem. *70*, 117–142.

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

Nabuurs, S.B., Spronk, C.A., Krieger, E., Maassen, H., Vriend, G., and Vuister, G.W. (2003). Quantitative evaluation of experimental NMR restraints. J. Am. Chem. Soc. *125*, 12026–12034.

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

Raman, S., Lange, O.F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T.A., Eletsky, A., Szyperski, T., et al. (2010). NMR structure determination for larger proteins using backbone-only data. Science *327*, 1014–1018.

Read, R.J., Adams, P.D., Arendall, W.B., III, Brunger, A.T., Emsley, P., Joosten, R.P., Kleywegt, G.J., Krissinel, E.B., Lütteke, T., Otwinowski, Z., et al. (2011). A new generation of crystallographic validation tools for the Protein Data Bank. Structure *19*, 1395–1412.

Rieping, W., and Vranken, W.F. (2010). Validation of archived chemical shifts through atomic coordinates. Proteins *78*, 2482–2489.

Rosato, A., Aramini, J.M., Arrowsmith, C., Bagaria, A., Baker, D., Cavalli, A., Doreleijers, J.F., Eletsky, A., Giachetti, A., Guerry, P., et al. (2012). Blind testing of routine, fully automated determination of protein structures from NMR data. Structure *20*, 227–236.

Serrano, P., Pedrini, B., Mohanty, B., Geralt, M., Herrmann, T., and Wüthrich, K. (2012). The J-UNIO protocol for automated protein structure determination by NMR in solution. J. Biomol. NMR *53*, 341–354.

Sheffler, W., and Baker, D. (2009). RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Sci. *18*, 229–239.

Shen, Y., and Bax, A. (2010). SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J. Biomol. NMR *48*, 13–22.

Snyder, D.A., and Montelione, G.T. (2005). Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. Proteins *59*, 673–686.

Snyder, D.A., Bhattacharya, A., Huang, Y.J., and Montelione, G.T. (2005). Assessing precision and accuracy of protein structures derived from NMR data. Proteins *59*, 655–661.

Spronk, C.A., Nabuurs, S.B., Bonvin, A.M., Krieger, E., Vuister, G.W., and Vriend, G. (2003). The precision of NMR structure ensembles revisited. J. Biomol. NMR *25*, 225–234.

Struyf, A., Hubert, M., and Rousseeuw, P. (1997). Clustering in an object-oriented environment. J. Stat. Softw. *1*, 1–30.

Tejero, R., Snyder, D., Mao, B., Aramini, J.M., and Montelione, G.T. (2013). PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. J. Biomol. NMR. Published online July 30, 2013. http://dx.doi.org/10.1007/s10858-013-9753-7.

Theobald, D.L., and Wuttke, D.S. (2006). THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics *22*, 2171–2172.

Theobald, D.L., and Wuttke, D.S. (2008). Accurate structural correlations from maximum likelihood superpositions. PLoS Comput. Biol. *4*, e43.

Trewhella, J., Hendrickson, W.A., Kleywegt, G.J., Sali, A., Sato, M., Schwede, T., Svergun, D.I., Tainer, J.A., Westbrook, J., and Berman, H.M. (2013). Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. Structure *21*, 875–881.

Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., et al. (2008). BioMagResBank. Nucleic Acids Res. *36*, D402–D408.

Valafar, H., and Prestegard, J.H. (2004). REDCAT: a residual dipolar coupling analysis tool. J. Magn. Reson. *167*, 228–241.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. *8*, 52–56, 29.

Vuister, G.W., Fogh, R.H., Hendrickx, P.M., Doreleijers, J.F., and Gutmanas, A. (2013). An overview of tools for the validation of protein NMR structures. J. Biomol. NMR. Published online July 23, 2013. http://dx.doi.org/10.1007/s10858-013-9750-x.

Wang, L., and Markley, J.L. (2009). Empirical correlation between protein backbone ^{15}N and ^{13}C secondary chemical shifts and its application to nitrogen chemical shift re-referencing. J. Biomol. NMR *44*, 95–99.

Wang, L., Eghbalnia, H.R., Bahrami, A., and Markley, J.L. (2005). Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J. Biomol. NMR *32*, 13–22.

Wang, B., Wang, Y., and Wishart, D.S. (2010). A probabilistic approach for validating protein NMR chemical shift assignments. J. Biomol. NMR *47*, 85–99.