J Biomol NMR (2013) 56:337–351
DOI 10.1007/s10858-013-9753-7

ARTICLE

PDBStat: a universal restraint converter and restraint analysis software package for protein NMR

Roberto Tejero · David Snyder · Binchen Mao · James M. Aramini · Gaetano T. Montelione

Received: 22 April 2013 / Accepted: 11 June 2013 / Published online: 30 July 2013

Abstract The heterogeneous array of software tools used in the process of protein NMR structure determination presents organizational challenges in the structure determination and validation processes, and creates a learning curve that limits the broader use of protein NMR in biology. These challenges, including accurate use of data in different data formats required by software carrying out similar tasks, continue to confound the efforts of novices and experts alike. These important issues need to be addressed robustly in order to standardize protein NMR structure determination and validation. PDBStat is a C/C++ computer program originally developed as a universal coordinate and protein NMR restraint converter. Its primary function is to provide a user-friendly tool for interconverting between protein coordinate and protein NMR restraint data formats. It also provides an integrated set of computational methods for protein NMR restraint analysis and structure quality assessment, relabeling of prochiral atoms with correct IUPAC names, as well as multiple methods for analysis of the consistency of atomic positions indicated by their convergence across a protein NMR ensemble. In this paper we provide a detailed description of the PDBStat software, and highlight some of its valuable computational capabilities. As an example, we demonstrate the use of the PDBStat restraint converter for restrained CS-Rosetta structure generation calculations, and compare the resulting protein NMR structure models with those generated from the same NMR restraint data using more traditional structure determination methods. These results demonstrate the value of a universal restraint converter in allowing the use of multiple structure generation methods with the same restraint data for consensus analysis of protein NMR structures and the underlying restraint data.

Keywords Protein NMR structure validation · BioMagResDatabase · XPLOR · CNS · CYANA · CS-Rosetta

Abbreviations

ACO	Dihedral angle constraint
CNSw	Protocol using the crystallography and NMR software (CNS) package for restrained structure refinement in explicit water solvent
DAOP	Dihedral angle order parameter
CS	Chemical shift
rCS-Rosetta	Restrained chemical shift-directed Rosetta
RDC	Residual dipolar coupling

Electronic supplementary material The online version of this article (doi:10.1007/s10858-013-9753-7) contains supplementary material, which is available to authorized users.

R. Tejero · B. Mao · J. M. Aramini · G. T. Montelione (✉)
Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, 679 Hoes Lane, Piscataway, NJ 08854-5638, USA
e-mail: guy@cabm.rutgers.edu

R. Tejero · B. Mao · J. M. Aramini · G. T. Montelione
Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, 679 Hoes Lane, Piscataway, NJ 08854-5638, USA

R. Tejero · B. Mao · J. M. Aramini · G. T. Montelione
Northeast Structural Genomics Consortium, 679 Hoes Lane, Piscataway, NJ 08854, USA

R. Tejero
Departamento de Química Física, Universidad de Valencia, Avenida Dr. Moliner 50, 46100 Burjassot, Valencia, Spain

D. Snyder
Department of Chemistry, William Paterson University, 300 Pompton Road, Wayne, NJ 07470, USA

SVD	Singular value decomposition
RMSD	Root mean squared deviation

Introduction

Protein structure determination by NMR methods involves integration of many different software tools. This heterogeneous environment presents a bottleneck in the structure determination process, and results in a steep learning curve that limits the broader use of NMR in molecular biophysics. The challenges of software integration are relevant to data collection, data analysis, and resonance assignments, as well as to the processes of 3D structure generation and structure quality assessment. Many computer programs and servers have been developed that integrate important parts of the process, including data collection, data processing (e.g. Delaglio et al. 1995), data analysis (e.g. Baran et al. 2006; Vranken et al. 2005; Zimmerman et al. 1997; Moseley and Montelione 1999; Moseley et al. 2001; Huang et al. 2006; Bahrami et al. 2009), and structure quality assessment (Bhattacharya et al. 2007; Doreleijers et al. 2012a, b). In particular, the challenge and importance of protein NMR structure validation has been the subject of several recent papers and reviews (Bhattacharya et al. 2007; Doreleijers et al. 2012a, b; Mao et al. 2011; Han et al. 2011; Rosato et al. 2012; Nabuurs et al. 2006; Hendrickx et al. 2013). However, the heterogeneous software environment of protein NMR spectroscopists continues to confound and slow down the efforts of novices and experts alike, and challenges efforts to standardize protein NMR structure determination and assessment.

PDBStat is a computer program originally developed as a universal coordinate and protein NMR restraint converter. Its primary function is to provide a user-friendly tool for interconverting between restraint data types. It also provides an integrated set of computational methods for protein NMR structure quality assessment. In order to streamline steps in the pipeline of NMR structure determination, and to provide information useful for protein NMR structure quality assessment, it also includes tools for standardized structural superimpositions, RMSD calculations, restraint summaries, restraint violation analyses, and various analyses validating models against experimental data.

Over the last several years, PDBStat has been used extensively by the Northeast Structural Genomics Consortium (NESG) as part of its platform for high throughout protein NMR structure determination (Liu et al. 2005; Huang et al. 2005; Baran et al. 2004). Within the NESG, PDBStat is used extensively to compare and interconvert input files for various third party software, allowing analysis of the same restraint data by multiple structure generation programs. PDBStat is also the restraint analysis software underlying the protein structure validation software (PSVS) server (Bhattacharya et al. 2007, 2008). However, the features and capabilities of PDBStat are much more extensive than the limited set of functions it provides for the PSVS server. In this paper we provide a detailed description of the PDBStat software, and highlight several of its valuable computational capabilities.

Description of the software

PDBStat is a stand-alone software program written largely in C. In the course of its evolution, PDBStat has also incorporated some Fortran and C++ subroutines. The program is freely available for non-commercial use at http://biopent.uv.es/~roberto. An on line version of the program can be accessed at "http://psvs.nesg.org/".

Universal protein coordinate and protein NMR restraint converter

One of the challenges of working in the heterogeneous software environment that has evolved in the protein NMR community, is that the naming conventions and formats for atomic coordinates (and the corresponding naming conventions and formats for NMR restraints) are different for various important software tools. PDBStat addresses this key issue by providing a universal coordinate and restraint converter, which can convert between any of the following coordinate and restraint formats: XPLOR-NIH (Schwieters et al. 2003), CNS (Brunger et al. 1998), DYANA (Güntert et al. 1997), CYANA (Herrmann et al. 2002), CHARMM, Rosetta (Rohl et al. 2004) (versions 2.3.0 and 3.x), as well as from the older formats of DIANA (Güntert et al. 1991), DISMAN (Braun and Go 1985), DISGEO (Williamson et al. 1985; Havel and Wüthrich 1985), and CONGEN (Bassolino-Klimas et al. 1996; Tejero et al. 1996). The most common use of PDBStat is for converting between XPLOR/CNS, CYANA, and Rosetta coordinates and restraints. PDBStat can read atomic coordinate and restraint files in any of these formats and convert them to standard PDB format coordinate files, with correct prochiral hydrogen labels, and a corresponding restraint file (with pseudoatom restraints where appropriate) consistent with these PDB coordinates. These files are the preferred format for submission of coordinates and restraint files to the PDB.

PDBStat can also be used to convert restraint lists and atomic coordinate files used for one program (e.g. CNS) into properly formatted files for use with another program (e.g. restrained CS-Rosetta). PDBStat uses a central representation, the IUPAC definitions for atom labels (Markley et al. 1998). Hence, operationally, coordinate and

restraint lists are converted first to IUPAC (e.g. CNS to IUPAC), and then from IUPAC to the desired format (e.g. IUPAC to restrained CS-Rosetta).

Protein atomic coordinate analysis

Relabeling of prochiral methylene and isopropyl methyl groups

IUPAC conventions (Markley et al. 1998) define the naming of prochiral sites in proteins, including the labeling of hydrogen atoms of methylene groups, and the labeling of isopropyl methyl groups. They also provide conventions for the labeling of sidechain amide protons of Asn and Gln residues. It is well known that some of the software packages commonly used in protein structure determination use atom labeling schemes that do not follow these IUPAC conventions. PDBStat provides automated analysis of prochiral sites and sidechain amide protons using a wide range of atomic coordinate formats, and relabels these atoms based on the local stereochemistry. This algorithm relies on structure geometry calculations done in the process of translating the coordinates, rather than on conversion tables. Hence the accuracy of the resulting prochiral labels does not depend on the accuracy of the labeling in the original coordinate file.

The PDBStat algorithm for prochiral atom naming is summarized in Fig. 1 for a methylene CβH2 group. In this case, the vectors from Cα to Cβ ($\overrightarrow{\boldsymbol{vcacb}}$), Cβ to Cγ ($\overrightarrow{\boldsymbol{vcbcg}}$), Cβ to one Hβ ($\overrightarrow{\boldsymbol{vcbhb1}}$) and Cβ to the second Hβ ($\overrightarrow{\boldsymbol{vcbhb2}}$) are first computed. Then the normal vector ($\overrightarrow{N}$) to the plane (shown in Fig. 1) defined by $\overrightarrow{\boldsymbol{vcacb}}$ and $\overrightarrow{\boldsymbol{vcbcg}}$ is computed ($\overrightarrow{\boldsymbol{N}} = \overrightarrow{\boldsymbol{vcacb}} \times \overrightarrow{\boldsymbol{vcbcg}}$). Finally the dot products with this normal, $\overrightarrow{\boldsymbol{vbhb1}} \cdot \overrightarrow{\boldsymbol{N}}$ and $\overrightarrow{\boldsymbol{vbhb2}} \cdot \overrightarrow{\boldsymbol{N}}$, are computed. The two dot products have *different signs*, since the Hβ's are separated by the plane. The Hβ atom (β1 or β2) with $\overrightarrow{\boldsymbol{vbhb}} \cdot \overrightarrow{\boldsymbol{N}} > 0$ is HB2, and the Hβ atom with $\overrightarrow{\boldsymbol{vbhb}} \cdot \overrightarrow{\boldsymbol{N}} < 0$ is HB3. The same procedure is applied for other prochiral methylene sites, as well as for the prochiral isopropyl methyl sites of Leu and Val sidechains. An alternative procedure for relabeling prochiral methylene and isopropyl methyl groups would be to follow IUPAC definition of the dihedral angles. For example for the Hβ2/Hβ3 case, using dih(N, Cα, Cβ, Hβ2)—dih(N, Cα, Cβ, Cγ) = ∼ + 120° and dih(N, Cα, Cβ, Hβ3)—dih(N, Cα, Cβ, Cγ) = ∼ −120°. PDBStat also provides proper stereospecific relabeling of the side-chain amide protons of Asn and Gln residues, and correct naming of the Cγ atoms of Thr and Ile residues.

Identifying "well defined" and "not well defined" regions of the protein structure

The representation of protein NMR structures generally involves providing an ensemble of conformers. Each member of the ensemble is generally a best-fit solution to the experimental data. The variation in conformations across the ensemble provides an estimate of the precision of the representation in various regions of the structure. Typically, some parts of the structure are "well defined", and other regions are "not well defined". However, the variations observed across NMR ensembles are often much greater than the variations that provide observable electron density in a crystal structure; i.e. the variations of coordinates in "not well defined" regions of the NMR structure are often of a magnitude that the corresponding structure would not be observable at all as electron density in an X-ray crystal structure.

The superimposed bundle of NMR conformers is therefore a valuable means of identifying which regions of the structure are precisely defined by the NMR experiments, and which are not well defined. While the superimposition process itself is relatively straightforward (Kabsch 1976, 1978), there are challenges in defining which atoms to use in the rotation/translation operators used to superimpose the structures. Methods and standard criteria for labeling residues or atoms as "well defined" or "not well defined" are an active area of research and discussion in the computational NMR community [see for example refs (Snyder and Montelione 2005; Hyberts et al. 1992; Kirchner and Güntert 2011)].

PDBStat provides two means of annotating "well defined" atomic coordinates: (1) the dihedral angle order parameter (Hyberts et al. 1992) and (2) the "core atom set" defined by analysis of the distance variance matrix (Snyder and Montelione 2005). Both of these methods are

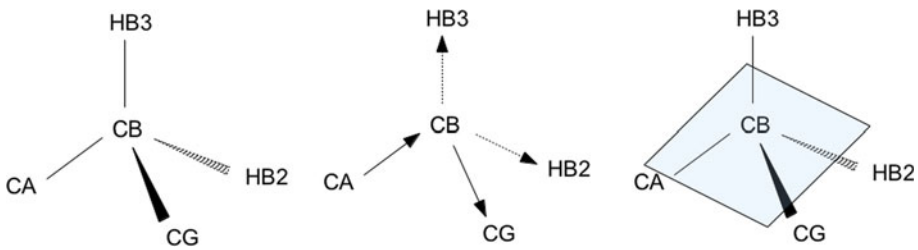

Fig. 1 Schematic depicting the algorithm used to define the stereospecific labeling of prochiral protons of the $C\beta H_2$ methylene group

independent of making an initial superimposition. Having defined a core atom set by one or another of these methods, these atoms can be used to compute appropriate superimpositions using standard methods (e.g. Kabsch 1976).

Dihedral angle order parameters

One of the most commonly used and generally accepted methods for distinguishing 'well-defined' from "not well defined" residue backbones is the dihedral angle order parameter (DAOP) (Hyberts et al. 1992). Using Eq. 1, PDBStat can calculate the DAOP and locate "well-defined residues" across an ensemble of N conformers.

$$S(\phi_i) = \frac{1}{N}\sqrt{\left(\sum_{j=1}^{N}\sin\phi_{i,j}\right)^2 + \left(\sum_{j=1}^{N}\cos\phi_{i,j}\right)^2} \quad (1)$$

A cutoff value $S(\phi_i) = 0.90$ (corresponding to a standard deviation of $\pm 24°$) (Hyberts et al. 1992) has been used to define a "well defined dihedral angle". PDBStat uses the default convention that if the sum of backbone DAOPs $S(\phi) + S(\psi) \geq 1.8$, then the entire residue is taken to be "well defined".

Variance matrix algorithm

The DAOP method has the advantage that it is fast and simple, and widely used by the protein NMR community. However, it has some significant shortcomings. The DAOP cannot distinguish local from long-range order; e.g. it is not possible to identify two well-defined "domains" or secondary structure elements which are themselves well-defined from the data, but connected by a flexible linker (Snyder and Montelione 2005). Secondly, this approach is backbone oriented, and does not provide a distinction between residues with "well-defined" and "not well defined" sidechains, or sidechains that are only partially "well-defined". PDBStat can also be used to define sidechain dihedral angle order parameters, which can be interpreted to provide information on the consistency of sidechain conformations across the ensemble of models. PDBStat also provides an implementation of "FindCore" variance matrix algorithm (Snyder and Montelione 2005) to identify well-defined atoms by partitioning atoms into core and non-core sets based on the variance in distances to all of the other atoms in the structure. The resulting "core atom sets" can be used to superimpose conformers, and for structure quality assessment. These well defined atom sets can also be used by PDBStat to identify "well-defined" backbone regions. The default criteria used by PDBStat for interpreting well-defined residue ranges from well-defined atom sets is to identify the residues as "well defined" if two or more of the N, Cα, and C′ atoms are in the well-defined core atom set.

Optimally superimposing coordinates (RMSD calculations)

In order to calculate root mean squared deviations (RMSDs) in atomic positions, PDBStat rotates and translates each conformer so as to minimize the RMS deviation with one reference conformer from the bundle, referred to as "refmol". These superimpositions are done using the core atom set(s) defined by either the DAOP or FindCore algorithms, described above, using the method of eigenvalue decomposition by multipliers of Kabsch (1976, 1978). The resulting superimposed coordinates are used to compute a mathematical average coordinate set for the ensemble, and the root-mean-squared deviations in atomic positions (RMSDs) are computed relative to these average coordinates. (Note that the "average coordinates" are not physically meaningful except for computing RMSDs). Tests have demonstrated that, when using well-defined atom sets, almost the same "RMSDs to average coordinates" are obtained regardless of which conformer is selected as "refmol". RMSDs can be reported for the "well defined" core atom set (using the command *rmsd best*), or for various subsets of atoms; e.g. RMSDs can be computed for alpha carbon (Cα), backbone (N, Cα, C′), all heavy (N, C, O, S), or all atoms including hydrogens.

Selecting a representative NMR model from the ensemble

NMR structures are generally reported as ensembles of models. The ensemble representation provides information about the consistency of the interpretation of the experimental data in different regions of the structure. However, biologists and other users of NMR structures are often confused by the ensemble representation. For this reason, it is advisable for the NMR experimentalist to designate a "representative" model from the ensemble. While no standard conventions have been agreed upon by the community for selecting the "representative structure", one useful convention is to select the single model that is most similar to all the other models. Specifically, we have adopted the convention that the representative structure is selected from the ensemble by considering only the well-ordered atoms, and then identifying the medoid (Struyf et al. 1996); i.e. the model in the ensemble that minimizes the RMSDs between it and all (other) models of the ensemble (Snyder and Montelione 2005). This selection can be done using the *representative structure* (rep) command in PDBStat. Using the same algorithm described above, the refmol which results in the lowest RMSD (i.e., the one most like all the other models) is selected as the representative structure. The model selected by this algorithm should be designated as Model #1—representative structure, in the PDB deposition.

Restraint analysis

Restraint statistics and restraint violation analyses

A key component of a protein NMR structure validation report is an analysis of how well each of the protein models reported in the NMR structure ensemble satisfies the experimental restraints. PDBStat provides extensive tools for assessment of restraint satisfaction and violations. The restraint analyses supported by PDBStat include (1) distance restraints (NOE, disulfide bond, and hydrogen bond restraints), (2) dihedral angle restraints, and (3) residual dipolar coupling restraints. Statistics are reported on the numbers and distributions of different types of restraints; e.g. intra and inter-residue, medium and long range restraints, hydrogen bond restraints, and inter-chain restraints of dimeric structures. In addition, the program reports statistics on the extent of restraint violations in these various categories. Using the universal structure and restraint converter of PDBStat, protein NMR structures and restraint lists generated for use by a wide range of programs (e.g. CYANA, CNS, XPLOR, or Rosetta) can be analyzed and reported using identical restraint violation statistics. This is a unique feature of the PDBStat software.

NOE distance restraint violations may be assessed using various interpretations of the relationship between interproton distance and NOE peak intensity. In generating restraints from NOE peak intensities, or in assessing restraints against atomic coordinates, PDBStat assumes the following "r^{-6} summation" relationship (Nilges 1995)

$$r = \left(\sum r_{ij}^{-6}\right)^{-\frac{1}{6}} \quad (2)$$

This interpretation assumes that the NOEs arising from each of the several interproton distances r_{ij} contributing to a single NOESY crosspeak contribute independently to the NOESY cross peak volume. This same "r^{-6} summation" interpretation of NOESY peak volumes (or intensities if volumes are not available) in terms of distance restraints is used for methyl groups (3 protons), chiral methylene protons lacking stereospecific assignments (2 protons), degenerate methylene protons (2 protons), chiral isopropyl methyl groups lacking stereospecific assignments (6 protons), degenerate isopropyl methyl groups (6 protons), and for degenerate aromatic ring protons of tyrosine or phenylalanine.

NOE completeness metric

NOE Completeness (Doreleijers et al. 1999) is defined as the ratio of NOE-derived interatom contacts indicated in the restraint list to the number of NOE-derived interatom contacts that are possible considering the 3D atomic coordinates. It is a metric reflecting the completeness of the NOE-derived restraint list. PDBStat has two methods for evaluating the NOE completeness, differing in the way the set of expected NOE-derived contacts is evaluated. In the first method, following the description in the original paper (Doreleijers et al. 1999), the set of expected contacts is generated based on a list of "observable atoms". These atom definitions are stored in an independent file, called Observable.nmr. In the second method, the "observable atoms" are defined automatically by PDBStat, rather than being provided by the user. In this case, the "observable atoms" are simply the set of hydrogen atoms for which chemical shift data is available, and the maximum NOE completeness is determined by the completeness of the proton resonance assignments. In either case, all interproton distances between "observable" hydrogen atoms are calculated from the NMR structure models, and all interproton distances below a cutoff are considered to be a potential NOE contact. The cutoff distance for evaluating the number of expected contacts can be selected by the user; the default value is 4.0 Å. An r^{-6} summed average is used for evaluating the interproton distances to degenerate atoms (e.g. methyl hydrogens), and a normal average is used to average each interproton distance across the ensemble of NMR models.

Analysis of residual dipolar coupling (RDC) restraints

PDBStat also provides an evaluation of the axial and rhombic components of the molecular alignment tensor, D_a and R, respectively, and calculation of RDCs based on protein atomic coordinates. A singular value decomposition (SVD) (Losonczi et al. 1999) is used to calculate D_a and R, providing results that are similar to those provided by standard programs [e.g. PALES (Zweckstetter and Bax 2000) and REDCAT (Valafar and Prestegard 2004)] used for RDC analysis. Statistics are reported summarizing how well the RDC values computed from the NMR conformers compare with the experimentally-determined RDC values, including the RDC Q factor (Cornilescu et al. 1998). Using the universal restraint format converter, this analysis can be done easily using structures and data that have been generated with different programs.

Parsing restraint files downloaded from the PDB

When depositing data in the PDB, restraints are collected together in a single file (extension .mr) that is archived together with the protein atomic coordinates. It is often interesting to re-analyze these data extracted from the PDB. PDBStat has a data parser to read the .mr restraint file, extract these restraint data, and to provide a statistical analysis of the restraints and the restraint violations.

PDBStat currently supports this feature for CNS/XPLOR and CYANA/DYANA distance restraint formats, which are used for the vast majority of the protein NMR restraint files archived in the PDB. However, the universal format converter of PDBStat will allow other restraint types to be handled as required.

Applications

In the following sections, we describe some valuable and/or unique applications of the PDBStat software.

Identifying conformationally-restricting restraints

PDBStat provides a comprehensive distance restraint summary analysis using distance restraint lists for the most commonly used structure generation programs (e.g. CNS/XPLOR, DYANA/CYANA, and Rosetta). As illustrated in Table 1, these include summaries of NOE-derived, hydrogen bond, disulfide, and ambiguous restraints, classified as intraresidue, sequential, medium-range, long-range, intrachain, and interchain restraints. An important feature of distance restraint analysis involves distinguishing conformationally-restricting restraints from those that are too loose to restrict the conformational space of the intervening dihedral angles. Such tools are available in some, but not all, computer software developed for analyzing protein structures from NMR data [e.g., redundant restraint analysis can be done using the CYANA program (Herrmann et al. 2002). The PDBStat program provides a "Clean NOE" utility that can be applied to restraint lists generated for use with several different structure generation programs. The Clean NOE utility functions to: (1) locate and remove duplicate restraints present in single or multiple restraint lists, (2) locate cases where the same atom pairs are restrained with different upper-bound distances, and removes the looser of these distance bounds, and (3) identify and remove restraints that do not restrict the conformational space of the intervening dihedral angles, which are typically intraresidue or sequential restraints. Rather than computing distances based on the conformations of intervening dihedral angles, a precompiled set of intraresidue and sequential restraining-distance bounds for various amino acid residue types has been generated using standard bond lengths and angles. These distance bounds account for different peptide libraries used by different structure calculation programs. This library of restraining-distance bounds is used to build rules and remove restraints that do not fulfill these rules. This approach has the advantage that the user can change some of the restraining-distance bound values in these precompiled lists (without recompiling the program), to allow for special cases or include new rules. The program then outputs an edited restraint list, excluding these non-functional restraints, along with a report of which restraints have been removed in this process. Table 1 shows the results of this restraint processing for NOE-based restraint lists generated for monomeric and homodimeric protein structures.

Relabeling of prochiral methylene and isopropyl methyl groups and sidechain amide atoms

In the process of converting coordinate file formats, PDBStat can also label prochiral methylene protons, prochiral isopropyl methyl atoms (both C and H), and sidechain amide protons with their correct IUPAC names. The program will also correctly relabel Thr OG1 and CG2, which are often mislabeled in older PDB coordinate files. This functionality is illustrated in PDBStat output shown in Table 2.

Table 1 Summary of statistics for restraints for NESG target protein CcR55 (PDB id 2jqn), a monomeric structure, and NESG target protein HR3057H (PDB id 2kw6), a homodimeric structure

Summary of restraints				
	PDB id 2jqn		PDB id 2kw6	
	Original	Clean	Original	Clean
Total number of restraints	1,676	1,200	1,901	1,831
Intra-residue restraints (i = j)	628	221	628	560
Sequential restraints |i–j| = 1	428	360	443	441
Backbone-backbone	162	119	84	82
Backbone-side chain	32	23	50	50
Side chain-side chain	234	218	309	309
Medium range restraints 1< |i–j| < 5	244	244	447	447
Backbone-backbone	64	64	104	104
Backbone-side chain	53	53	164	164
Side chain-side chain	127	127	179	179
Long range restraints |i–j| ≥ 5	376	376	383	383
Total hydrogen bond restraints	66	66	0	0
Long range H-bond restraints |i–j| ≥ 5	38	38	0	0
Disulfide restraints	0	0	0	0
Intrachain restraints	1,742	1,266	1,671	1,601
Interchain restraints	0	0	230	230
Ambiguous restraints	0	0	0	0

"Original" refers to the restraint sets as they would have been deposited in PDB prior to the introduction of PDBStat into our standard deposition process, and "Clean" summarizes the restraint lists regenerated using the Clean NOE utility of PDBStat

Table 2 PDBStat output demonstrating the relabeling of prochiral methylene protons, isopropyl methyl atoms of Leu and Val, sidechain amide protons of Asn and Gln, and Thr gamma atoms

```
PdbStat>  --> Fixing StereoNames of model   1
PdbStat>   The Leucine Carbons (2 CD's)
PdbStat>        ( CD2 ) renamed ( CD1 ) in  29 (LEU)
PdbStat>        ( CD1 ) renamed ( CD2 ) in  29 (LEU)
PdbStat>   The Valine Carbons (2 CG's)
PdbStat>        ( CG2 ) renamed ( CG1 ) in  27 (VAL)
PdbStat>        ( CG1 ) renamed ( CG2 ) in  27 (VAL)
PdbStat>   The Isoleucine Carbons (2 CG's)
PdbStat>        ( CG2 ) renamed ( CG1 ) in   4 (ILE)
PdbStat>        ( CG1 ) renamed ( CG2 ) in   4 (ILE)
PdbStat>   The threonine OG1 and CG2 check
PdbStat>        ( OG2 ) renamed ( OG1 ) in  45 (THR)
PdbStat>        ( CG1 ) renamed ( CG2 ) in  45 (THR)
PdbStat>   General case of Two Beta Protons (HB's)
PdbStat>        (3HB  ) renamed (2HB  ) in   1 (MET)
PdbStat>        (2HB  ) renamed (3HB  ) in   1 (MET)
PdbStat>   General case of Two Gamma Protons (HG's)
PdbStat>        (3HG  ) renamed (2HG  ) in   1 (MET)
PdbStat>        (2HG  ) renamed (3HG  ) in   1 (MET)
PdbStat>   General case of Two Delta Protons (HD's)
PdbStat>        (3HD  ) renamed (2HD  ) in   6 (LYS)
PdbStat>        (2HD  ) renamed (3HD  ) in   6 (LYS)
PdbStat>   General case of Two Epsilon Protons (HE's)
PdbStat>        (3HE  ) renamed (2HE  ) in   6 (LYS)
PdbStat>        (2HE  ) renamed (3HE  ) in   6 (LYS)
PdbStat>   The Glycine Protons (2 HA's)
PdbStat>        (3HA  ) renamed (2HA  ) in  52 (GLY)
PdbStat>        (2HA  ) renamed (3HA  ) in  52 (GLY)
PdbStat>   Two Beta Protons (HB's) of Cystine
PdbStat>   Two Beta Protons (HB's) of Serine
PdbStat>        (3HB  ) renamed (2HB  ) in  75 (SER)
PdbStat>        (2HB  ) renamed (3HB  ) in  75 (SER)
PdbStat>   ASN side chain amides
PdbStat>        (2HD2 ) renamed (1HD2 ) in  55 (ASN)
PdbStat>        (1HD2 ) renamed (2HD2 ) in  55 (ASN)
PdbStat>   GLN side chain amides
PdbStat>        (2HE2 ) renamed (1HE2 ) in 108 (GLN)
PdbStat>        (1HE2 ) renamed (2HE2 ) in 108 (GLN)
```

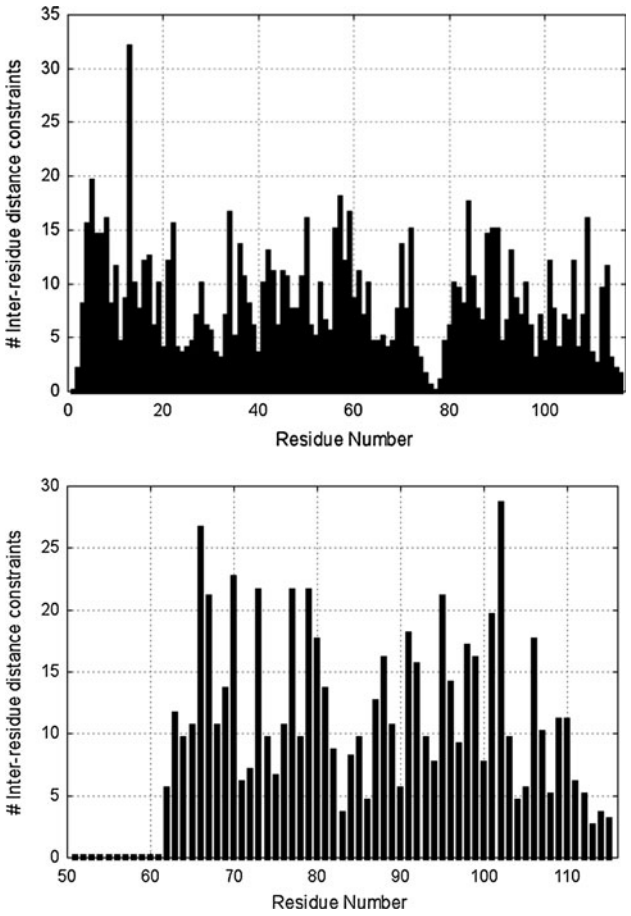

Fig. 2 Histogram plots generated by PDBStat of conformationally-restricting restraints per residue, analyzed from a restraint list. Results are shown for a monomeric protein (NESG id CcR55; PDB id 2jqn) at the *top*, and for a homodimeric protein (NESG id HR3057H; PDB id 2kw6) at the *bottom*

Analysis of restraint density

PDBStat also provides an analysis of the restraint density along the protein sequence. In this analysis, an interresidue distance restraint between residues i and j is assigned 0.5 units to residue i and 0.5 units to residue j. The resulting histogram plots of interresidue restraints per residue, providing a survey of restraint density along the protein sequence (Fig. 2), is output as a PNG format file using the gnuplot software (Williams and Kelley 2011), suitable for inclusion as a figure in a manuscript or associated supplementary material.

Analysis of contact maps derived from restraint lists or derived from 3D models

Contact maps are an important tool for assessing restraint data sets, 3D structures of protein models, and the agreement between restraint data sets and 3D protein model coordinates. PDBStat has utilities to generate contact maps from distance restraint lists, as illustrated in the left panels of Fig. 3, and contact maps from the 3D protein coordinates (using a default distance cutoff of 5 Å, which can be modified by the user), illustrated in the right panels of Fig. 3. This analysis can be done for monomers (top panels of Fig. 3), homodimers (bottom panels of Fig. 3), or heterodimers (results not shown). Contact maps may be generated for residue-residue contacts (as illustrated in Fig. 3) or for atom–atom contacts (result not shown). These residue-residue contact maps are produced directly by PDBStat either as text files or as postscript images. Comparison of contacts maps derived from restraint lists and 3D structures are useful both for validating structures (Huang et al. 2012, 2005) and for iterative analysis of NOESY data to provide more complete restraint lists (Huang et al. 2005, 2006, 2012; Herrmann et al. 2002). For example, the RPF software (Huang et al. 2005, 2012) for assessing the agreement between a protein model and NOESY peak list data is based on the concept of comparing all possible atomic contact maps derived from the NMR resonance assignment and NOESY data with contact maps derived from the 3D model.

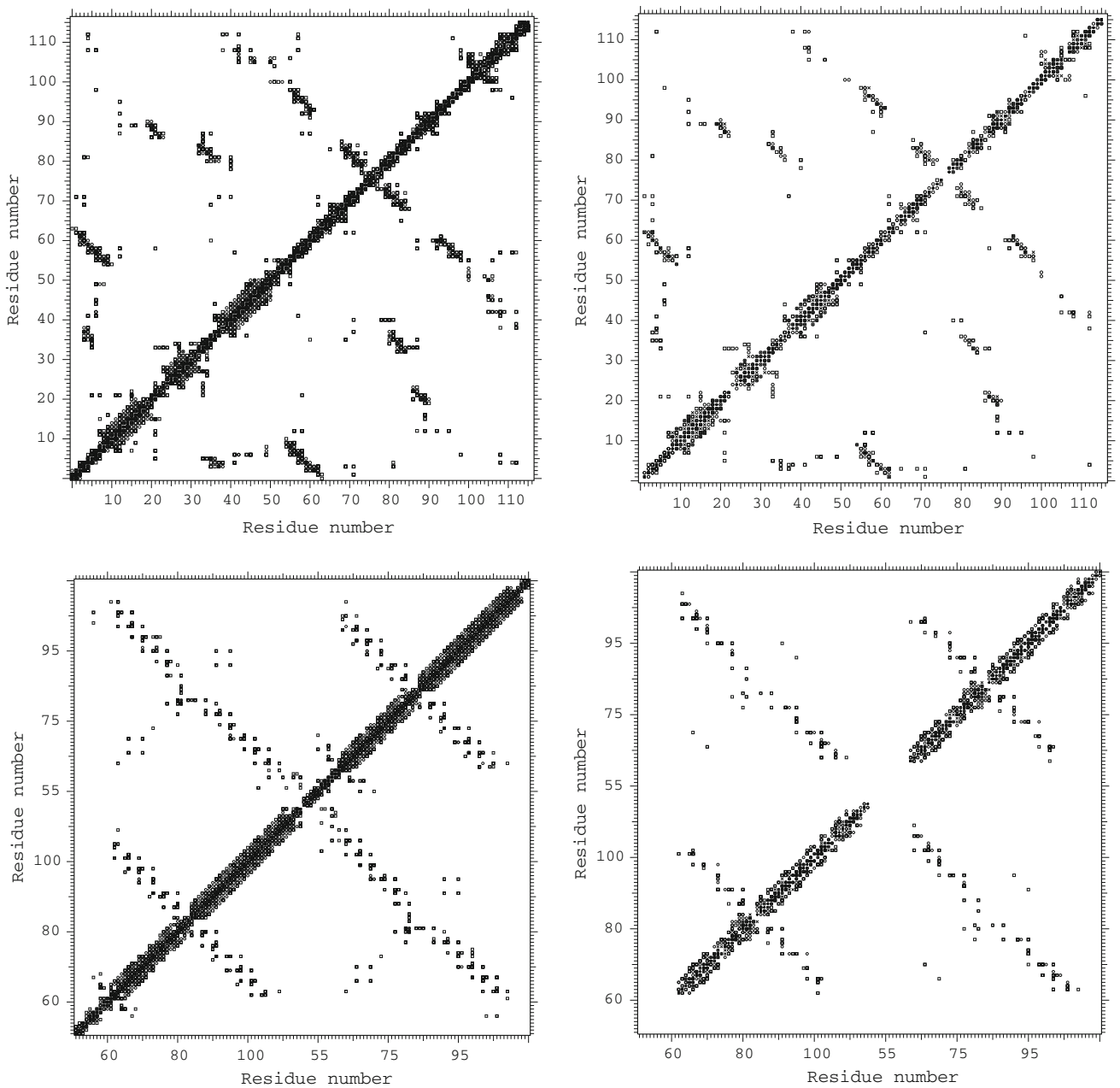

Fig. 3 Contact maps generated by PDBStat for conformationally-restricting restraints in an input restraint list, *left*, and for short interproton distances derived from atomic coordinates, *right*. Results are shown for a monomeric protein (NESG id CcR55; PDB id 2jqn) at *top*, and for a homodimeric protein (NESG id HR3057H; PDB id 2kw6) at *bottom*. Comparisons of such plots provide a visual analysis of how well a 3D structure model fits to the experimental restraint list

Analysis of backbone and sidechain dihedral angle order parameter (DAOP)

As outlined in the Description of Software section above, PDBStat provides an analysis of both backbone and sidechain DAOPs. These analyses are illustrated in Fig. 4 for two NMR-derived conformational ensembles archived in the PDB; one for a momomeric protein structure (left panel), and the other for one protomer of a homodimeric protein structure (right panel). These analyses are provided as both text files and as PNG images.

Analysis of "well defined" and "not well defined" regions of the protein structure

The DAOP analysis may be used to provide information regarding which regions of the polypeptide backbone are well-defined (defined by convention in PDBStat as

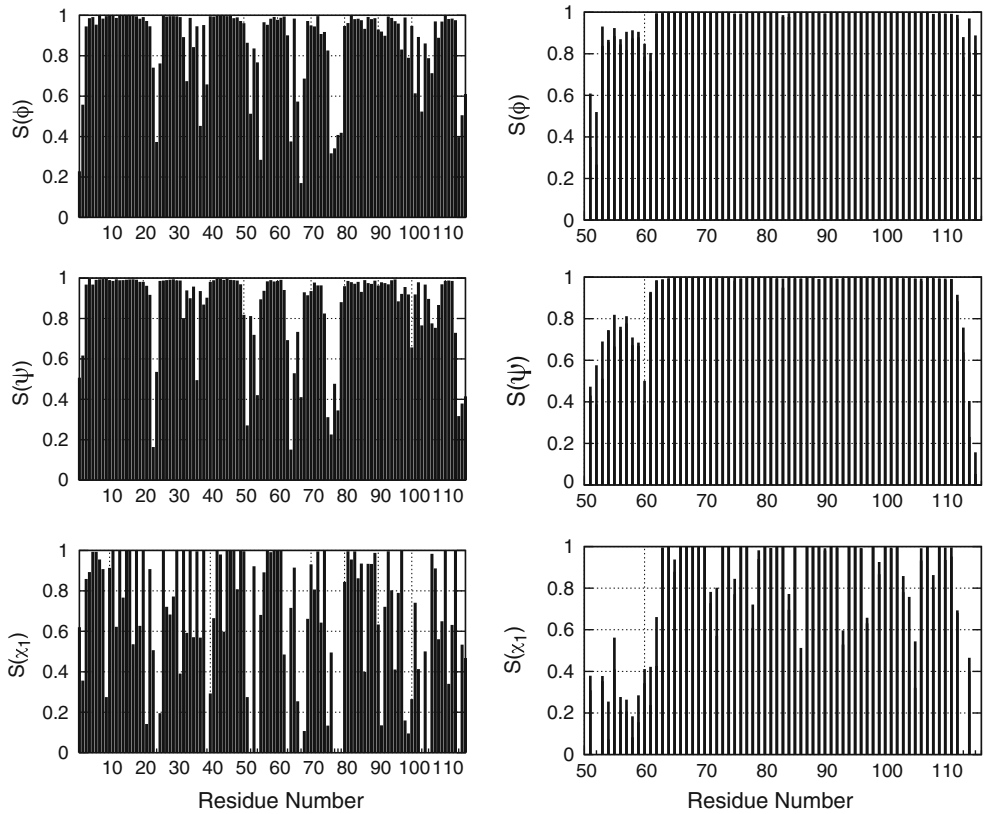

Fig. 4 *Histogram* plots of dihedral angle order parameters (DAOP) for ϕ, ψ and χ_1 versus amino acid sequence obtained from PDBStat. Results are shown for a monomeric protein (NESG id CcR55; PDB id 2jqn) at *left*, and for one protomer of a homodimeric protein (NESG id HR3057H; PDB id 2kw6), at *right*

$S(\phi) + S(\psi) \geq 1.8$), and which are not well-defined, as outlined above in the Description of Software section. Figure 4 illustrates the DAOP analysis for backbone ϕ and ψ, as well as sidechain χ_1, dihedral angles in monomeric and homodimeric protein structures. PDBStat also provides an implementation of the FindCore algorithm (Snyder and Montelione 2005), identifying sets of atoms that are well-defined with respect to one another using a variance matrix analysis.

A comparison of these two methods, DAOP and FindCore, using a representative NESG NMR structure ensemble, is illustrated in Fig. 5. In the left panel, the backbone atoms (N, Cα, C′) of the 20 conformers of the ensemble archived in the PDB are superimposed on one representative conformer. In this case, the representative conformer was selected as the medoid structure, with lowest backbone RMSD to all the other conformers in the ensemble. The atoms used in the superimposition are those that were identified as "well-defined" backbone atoms by the FindCore implementation in PDBStat. Backbone atoms colored in yellow are those which both methods identify as "not well defined", while atoms shown in dark blue are those identified by both methods as "well defined". Atoms colored light blue are "well defined" based on FindCore, but "not well defined" based on DAOP, while atoms colored green are "well defined" based on DAOP, but not "well defined" based on FindCore. The backbone atomic coordinates of the corresponding X-ray crystal structure are shown in red. Interestingly, some segments within the consensus "not well defined" polypeptide regions (yellow) have DAOP above the threshold (green); i.e. residues in these green regions have relatively consistent backbone dihedral angles. On the other hand, the regions identified by FindCore, but not DAOP analysis, as "well defined", shown in light blue, have atomic variances that are similar to those in the consensus well-defined regions (dark blue). The left panel of Fig. 5 thus demonstrates the complementary value of FindCore and DAOP analysis in identifying "well defined" regions of the protein structure that might best be used in structure quality assessment, for computing superimpositions, or in interpreting the precision of various regions of the NMR structure for structure–function studies.

The right panel of Fig. 5 illustrates "well-defined" and "not well defined" regions of protein sidechains. The panel shows a region of the protein structure where the backbone is well defined (blue), while the corresponding side chains, or parts of these side chains, associated with these well

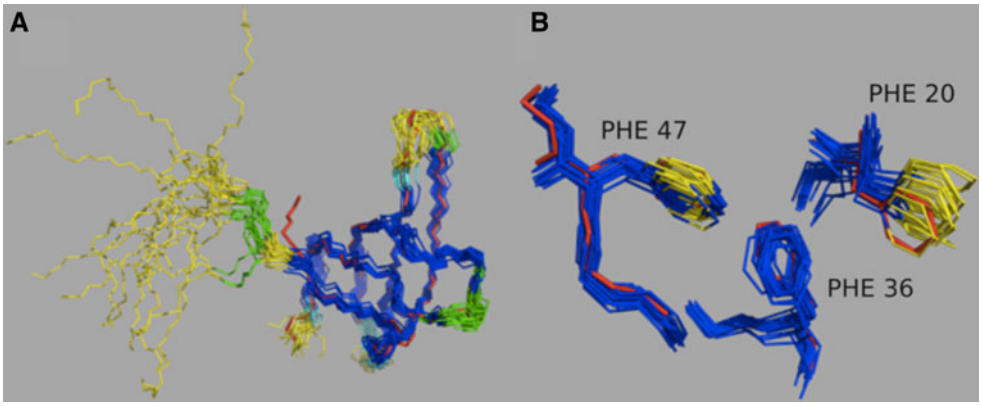

Fig. 5 FindCore provides atom-specific designations for well-defined and not-well-defined regions of NESG protein SgR42 (PDB id 2jz2). **a** Ensemble superimposition showing residues defined as "well defined" by the FindCore DAOP analysis and those identified as "well defined" by the FindCore variance matrix analysis. Residues identified as "well defined" by both methods are shown in *dark blue*, those identified as "well defined" by variance matrix but not by DAOP in *light blue*, and those identified as "well defined" by DAOP but not by variance matrix in *green*. Residues identified as "not well defined" by both methods are shown in *yellow*. The backbone atoms of the corresponding X-ray crystal structure (PDB id 3c4s) are shown in *red*. **b** Expansion showing atom-specific "well defined" (*dark blue*) and "not well defined" (*yellow*) designations for the sidechains of residues Phe20, Phe36, and Phe47 in protein SgR42

defined backbone atoms are themselves "not well defined" (yellow), based on the FindCore analysis. Interestingly, rotation about the ring axis of Phe47 results in less-well-defined positions for the Cδ and Cε atoms relative to the rest of this side chain. This result demonstrates the special value of *atom specific* designators of structural precision over the standard convention of defining only residue ranges of the well-defined regions of the protein NMR structure.

Generating protein structures using restrained CS-Rosetta

Recently the Rosetta program has been further developed to allow the use of a wide range of interatomic distance restraints and residual dipolar coupling (RDC) data (Raman et al. 2010; Lange et al. 2012). These enhancements, directed to the challenges of solving larger protein structures, also allow the general use of Rosetta or CS-Rosetta together with NMR-derived distance restraints in a manner similar to conventional distance-restrained structure generation calculations with CNS, XPLOR, CYANA or other more traditional protein structure generation program. We refer to these as restrained CS-Rosetta (rCS-Rosetta) NMR structure generation calculations.

Using the universal restraint converter of PDBStat, restraint lists originally prepared for CNSw calculations were generated and used as distance restraints in rCS-Rosetta calculations. This approach was benchmarked in this study using two small NESG target proteins, ZR18 (91 residues) and PfR193A (114 residues), for which both solution NMR (PDB ids 1pqx and 2kl6, respectively) and X-ray crystal structures (PDB id 2ffm and 3idu, respectively) have been determined and archived in the PDB. These targets also have extensive NMR data archived in the BioMagResDB (Doreleijers et al. 2003) (BMRB ids 5844 and 16385, respectively). For PfR193A, ^{15}N–^{1}H RDC data were also available, and were used in the rCS-Rosetta calculations as described elsewhere (Raman et al. 2010; Lange et al. 2012). These NMR restraint data were used to determine the structures of ZR18 and PfR193A using a standard NESG protocol, involving initial structure generation with CYANA followed by structure refinement with CNS in explicit water solvent (as described in detail at http://www.nmr2.buffalo.edu/nesg.wiki/). This standard NESG protocol is designated as the CNSw protocol. For this study, the ZR18 structure was downloaded from the PDB and re-refined using the CNSw protocol (since it had been deposited in the PDB before the standard CNSw protocol was adopted), while for PfR193A the CNSw-refined coordinates were those obtained from the PDB. The CNS restraints were then converted to rCS-Rosetta restraints using PDBStat, and rCS-Rosetta structures were generated using Rosetta ver. 3, as described in detail elsewhere (Mao, Tejero and Montelione, in preparation). The rCS-Rosetta calculations used restraint tolerance of 0.3 Å; i.e. loosening each restraint by 0.3 Å during the calculations to allow the structure to deviate slightly from experimental restraints in order to better satisfy the Rosetta energy function. rCS-Rosetta calculations required about 4,000 min to generate 10,000 decoys using 20 2.5 GHz processors, compared with CYANA structure generation followed by CNSw refinement, which required about 40 min using the same 20 2.5 GHz processors to generate

100 conformers with CYANA and to refine 20 conformers with CNSw.

The Restraint Summary and Restraint Violation Analysis provided by PDBStat, along with other knowledge-based structure quality scores provided by the PSVS (Bhattacharya et al. 2007) and RPF (Huang et al. 2005) programs, for the two structures, each solved with the two different protocols, are shown in Table 3. The resulting conformational ensembles are compared with each other, and with the corresponding X-ray crystal structures in Fig. 6, using the superimposition utilities of PDBStat. Average pairwise RMSDs within each ensemble, and between the NMR conformers and the corresponding X-ray crystal structure, are tabulated in Table 4, along with GDT-TS and GDT-HA backbone superimposition scores (Zemla 2003) assessing structural similarity of Cα atom positions. For both ZR18 and PfR193A, the rCS-Rosetta structures are more similar to the X-ray structure than the structures generated using the standard CYANA-CNSw protocol; for ZR18 the changes are relatively substantial, with movements of up to 2 Å, while for PfR193A the differences in backbone structures generated by the two methods are smaller. In both cases, the CYANA-CNSw and rCS-Rosetta NMR structures fit equally well to the NOESY peak list data (i.e. RPF and DP scores), as shown in Table 3. The rCS-Rosetta structures have significantly better knowledge-based structure quality scores than structures generated from the same restraint data using the standard NESG CYANA-CNSw protocol (Table 3). rCS-Rosetta structures, however, tend to have larger numbers of distance restraint and dihedral angle violations (Table 3), measured against the loosened restraints (i.e. the Rosetta restraints with 0.3 Å loosening of upper bounds). A single conformer in the ensemble of PfR193A structures has a significant dihedral angle violation of almost 60°. The CNSw-refined and rCS-Rosetta structures for ZR18 and rCS-Rosetta structure of PfR193A have been deposited in the PDB (2m6q and 2m8w for ZR18 and 2m8x for PfR193A, respectively).

A careful analysis of restraints in CNS and Rosetta formats, interconverted by PDBStat, and of the specific violations observed for the rCS-Rosetta structures, confirms that these restraint violations are characteristic of the rCS-Rosetta structures, rather than the result of errors in restraint conversion. For example, when the CNSw structures are assessed using the Rosetta restraints obtained following conversion from CNS format, as expected no residual restraint violations are observed (Supplementary Table S1). In addition NOE completeness calculations were done for both sets of restraints and the results are again the same. The modest number of small restraint violations in the rCS-Rosetta structures are indeed a feature of these structures.

Discussion

The PDBStat program is a central component of the NESG NMR structure production pipeline, and of the PSVS (Bhattacharya et al. 2007) structure quality assessment server, and has been used as part of the structure determination and structure validation process on over 450 protein NMR structures. It is a user-friendly software package that integrates many of the computational tools needed to generate and assess protein structures from NMR restraint lists. The PDBStat software is easy to install on laptops or computers in small NMR lab groups, and has minimal requirements in terms of disk space and CPU speed. The software provides a uniform restraint converter that allows the same restraint data to be used with several different structure generation programs, including CNS/XPLOR, various versions of DYANA and CYANA, and Rosetta. Some of these features are also provided by the CING (Doreleijers et al. 2012a, b) and CCPN (Vranken et al. 2005) software packages. In our experience, however, the restraint conversions and restraint violation analysis tools provided by PDBStat are much more extensive and easier to use.

A special feature of PDBStat is the accurate conversion of restraint lists prepared from one program (e.g. CYANA) into restraint lists that can be used to run another structure generation program (e.g. CNS or rCS-Rosetta). This versatility underlies an evolving approach in which once a protein NMR structure is determined using one software package, it can be validated by rapid redetermination with other software packages. This approach can also be used to validate restraint lists generated by different automatic NOESY analysis programs. PDBStat's universal restraint and coordinate conversion utilities thus provide the basis for the use of many NMR data analysis and software generation programs in parallel, and standardized assessment of restraint violations generated by the different structure generation programs.

Of special interest is the conversion of CYANA or CNS restraint lists into input for rCS-Rosetta. The resulting rCS-Rosetta structures are observed to have better knowledge-based structure quality scores, better agreement with the corresponding crystal structure, and about equally good global scores in matching to the NOESY peaks lists (i.e. RPF scores), as structures generated from the same restraint data using our standard protocol of CYANA structure generation followed by CNSw refinement. However, these rCS-Rosetta structures have a larger number of small restraint violations. Careful analysis of these restraint violations demonstrates that they are not the result of inaccurate restraint conversion by PDBStat; indeed when the rCS-Rosetta restraints are used to assess the CNSw-refined structures there are essentially no serious restraint violations and the NOE completeness calculations are the same, as expected (Supplementary Table S1). Hence, the

Table 3 NMR structure statistics for the CNSw and rCS-Rosetta structures of PfR193A and ZR18

	PfR193A (CNSw)	PfR193A (rCS-Rosetta)	ZR18 (CNSw)	ZR18 (rCS-Rosetta)
Completeness of resonance assignments[a]				
Backbone (%)	93.46	93.46	93.99	93.99
Side chain (%)	90.34	90.34	78.02	78.02
Aromatic (%)	100	100	100	100
Stereospecific methyl (%)	88.46	88.46	100	100
Conformationally-restricting restraints[b]				
Distance restraints				
Total	2,719	2,719	1,137	1,137
Intra-residue ($i = j$)	523	523	168	168
Sequential ($\|i–j\| = 1$)	686	686	337	337
Medium range ($1<\|i–j\| < 5$)	271	271	217	217
Long range ($\|i–j\| \geq 5$)	1,239	1,239	415	415
Dihedral angle restraints	165	165	179	179
Hydrogen bond restraints	0	0	54	54
No. of restraints per residue	26.7	26.7	15.7	15.7
No. of long range restraints per residue	11.5	11.5	5.1	5.1
Residual restraint violations[b]				
Average no. of distance viol per structure				
0.1–0.2 Å	2.60	16.05	9.50	8.95
0.2–0.5 Å	0.05	10.95	5.10	5.35
>0.5 Å	0	2.25	0.50	3.20
Largest violation (Å)	0.22	1.72	0.95	1.30
Average no. of dihed angle viol per structure				
1–10°	19.90	2.55	9.7	1.35
>10°	0	1.55	0.1	0.15
Largest violation (°)	9.6	59.7	11.0	22.2
Model quality[b]				
RMSD backbone atoms (Å)[c]	0.4	0.4	0.7	0.6
RMSD heavy atoms (Å)[c]	0.6	0.6	1.0	0.8
RMSD bond lengths (Å)	0.008	0.010	0.019	0.010
RMSD bond angles (°)	0.6	0.5	1.3	0.4
MolProbity Ramachandran statistics[b,c]				
Most favored regions (%)	96.7	98.4	88.9	97.1
Allowed regions (%)	3.2	1.3	9.8	2.9
Disallowed regions (%)	0.1	0.3	1.3	0.0
Global quality scores (Raw/Z-score)[b]				
Verify3D	0.37/−1.44	0.41/−0.80	0.33/−2.09	0.43/−0.48
ProsaII	0.35/−1.24	0.35/−1.24	0.43/−0.91	0.67/0.08
Procheck (phi–psi)	−0.42/−1.34	−0.30/−0.87	−0.70/−2.44	−0.20/−0.47
Procheck (all)[c]	−0.27/−1.60	−0.05/0.30	−0.43/−2.54	0.12/0.71
MolProbity clash score	15.93/−1.21	10.36/−0.25	12.95/−0.70	6.27/0.45
RPF scores[d]				
Recall/Precision	0.967/0.955	0.967/0.958	0.927/0.779	0.931/0.772
F-measure/DP-score	0.961/0.874	0.963/0.881	0.847/0.747	0.844/0.747
Model contents				
Ordered residue range[c]	436–541	436–520, 523–541	3–5, 14–18, 28–32, 38–50, 53–68, 70–82	2–11, 14–22, 27–83
BMRB accession number	16,385	16,385	5,844	5,844
PDB id	2kl6	2m8x	2m6q	2m8w

Structural statistics computed for the ensemble of deposited structures

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

[b] Calculated using PSVS 1.5 (Bhattacharya et al. 2007). Average distance violations were calculated using the sum over r^{-6}

[c] For ordered residues with [S(phi) + S(psi) ≥ 1.8]

[d] RPF scores (Huang et al. 2005b, 2012) reflecting the goodness-of-fit of the final ensemble of structures (including disordered residues) to the NOESY data and resonance assignments

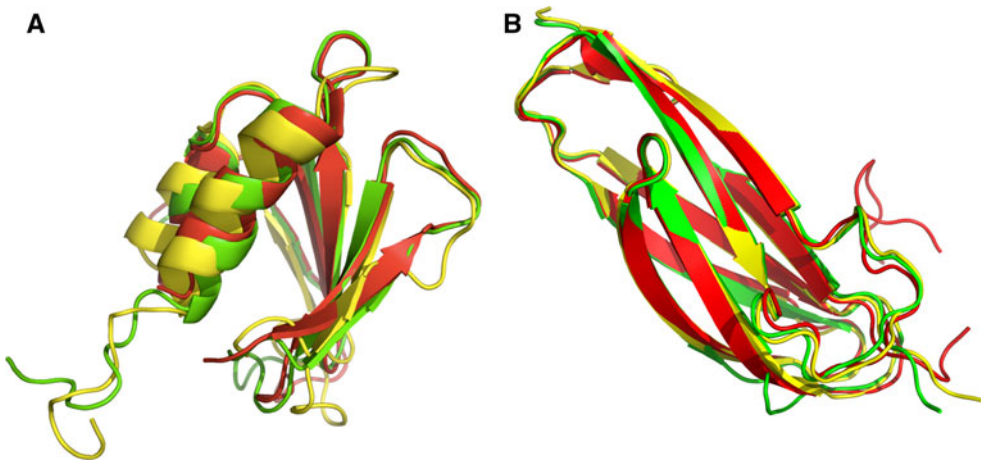

Fig. 6 Comparison of small protein structures generated with either a standard CYANA-CNSw protocol or with restrained CS-Rosetta (rCS-Rosetta). Results are shown for NESG target proteins ZR18 (**a**) and PfR193A (**b**). For each protein target, backbone structures are shown for the X-ray crystal structure (*red*), a standard CYANA-CNSw refined structure (*yellow*), and a structure generated using rCS-Rosetta following restraint conversion using PDBStat (*green*). As quantified in Tables 3 and 4, for these small proteins the rCS-Rosetta structures have better knowledge-based structure quality scores, equally good agreement with the NOESY peak list data, and are slightly more similar to the corresponding X-ray crystal structures

Table 4 Backbone (N, Cα, C′) RMSD and GDT scores comparing structures of NESG target proteins ZR18 and PfR193A generated with the standard NESG CYANA-CNSw protocol, with the same structures refined with rCS-Rosetta following restraint conversion using PDBStat. These ensembles of NMR conformers are compared with the corresponding X-ray crystal structure coordinates

	Number of residues[a]	rCS-Rosetta versus CNSw RMSD/GDT-TS/GDT-HA	rCS-Rosetta versus X-ray RMSD/GDT-TS/GDT-HA	CNSw versus X-ray[b] RMSD/GDT-TS/GDT-HA
ZR18	91	1.18 Å/0.86/0.68	0.98 Å/0.94/0.79	1.40 Å/0.82/0.62
PfR193A	114	0.42 Å/0.97/0.86	0.57 Å/0.99/0.93	0.64 Å/0.99/0.89

[a] Well-defined residues were determined by PDBStat using the FindCore (Snyder and Montelione 2005) method

[b] For target ZR18, the PDB id of the reference X-ray crystal structure is 2ffm. For target PfR193A, the PDB id of the reference X-ray crystal structure is 3idu

modest restraint violations observed in the rCS-Rosetta structures reflect the inconsistency between the NMR restraints and the conformations preferred in the Rosetta force field, which are generally closer to the crystal structure. The violations of Rosetta restraints were measured against the loosened restraints (i.e. the Rosetta restraints with 0.3 Å loosening of upper bounds).

Similar observations have been observed in unrestrained refinement using Rosetta (Ramelot et al. 2009), where it was first suggested that such analyses can be used to correct misinterpretation and miscalibration of restraints derived from the NOESY peak list data due to misassignment of NOESY cross peaks, the effects of conformational averaging, and/or attenuation of cross peak intensities due to exchange broadening. Indeed, in a systematic study of some 40 pairs of structures determined by both NMR and X-ray crystallography (Mao, Tejero, and Montelione, in preparation), we consistently observe that as the accuracy of the NMR structure relative to the crystal structure improves, a small number of NMR-derived restraints become violated. It is not clear if these represent inaccuracies in the restraints or the effects of dynamic averaging in solution. However, these observations demonstrate the tremendous power of the PDBStat universal restraint converter in allowing a simple conversion between restraint formats, allowing users to rapidly and easily exploit the unique strengths of different structure generation packages using the same experimental data. In this way, users can consider to determine NMR structures in parallel with several different structure generation methods, and to use consensus methods to improve the accuracy of the NOESY data interpretation and the precision and accuracy of the resulting NMR structure models.

Acknowledgments We thank all the members of the NMR groups of the Northeast Structural Genomics Consortium who contributed constructive criticisms and test data sets used in the development of PDBStat. Special thanks to C. Arrowsmith, J. Cort, A. Eletsky, L. Fella, Y. J. Huang, A. Lemak, M. Kennedy, G. Liu, J. Prestegard, T. Ramelot, A. Rosato, G.V.T. Swapna, T. Szyperski, Y. Tang, and B. Wu for useful discussions. This work was supported by a Grant from the Protein Structure Initiative of the National Institutes of Health (U54-GM094597). RT also acknowledges suppport from CONSOLIDER INGENIO CSD2010-00065 and Generalitat Valenciana

PROMETEO 2011/008. DS also acknowledges support from the Research Corporation for Science Advancement, College Cottrell Grant, Award #19803.

References

Bahrami A, Assadi AH, Markley JL, Eghbalnia HR (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. PLoS Comput Biol 5:e1000307

Baran MC, Huang YJ, Moseley HN, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104:3541–3556

Baran MC, Moseley HN, Aramini JM, Bayro MJ, Monleon D, Locke JY, Montelione GT (2006) SPINS: a laboratory information management system for organizing and archiving intermediate and final results from NMR protein structure determinations. Proteins 62:843–851

Bassolino-Klimas D, Tejero R, Krystek SR, Metzler WJ, Montelione GT, Bruccoleri RE (1996) Simulated annealing with restrained molecular dynamics using a flexible restraint potential: theory and evaluation with simulated NMR constraints. Protein Sci 5:593–603

Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. Proteins 66:778–795

Bhattacharya A, Wunderlich Z, Monleon D, Tejero R, Montelione GT (2008) Assessing model accuracy using the homology modeling automatically (HOMA) software. Proteins 70:105–118

Braun W, Go N (1985) Calculation of protein conformations by proton-proton distance constraints: a new efficient algorithm. J Mol Biol 186:611–626

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

Cornilescu G, Marquardt JL, Ottiger M, 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

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

Doreleijers JF, Raves ML, Rullmann T, Kaptein R (1999) Completeness of NOEs in protein structure: a statistical analysis of NMR data. J Biomol NMR 14:123–132

Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Markley JL, Ulrich EL (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. J Biomol NMR 26:139–146

Doreleijers JF, Sousa da Silva AW, Krieger E, Nabuurs SB, Spronk CA, Stevens TJ, Vranken WF, Vriend G, Vuister GW (2012a) CING: an integrated residue-based structure validation program suite. J Biomol NMR 54:267–283

Doreleijers JF, Vranken WF, Schulte C, Markley JL, Ulrich EL, Vriend G, Vuister GW (2012b) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. Nucleic Acids Res 40:D519–D524

Güntert P, Braun W, Wüthrich K (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. J Mol Biol 217:517–530

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

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57

Havel TF, Wüthrich K (1985) An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. J Mol Biol 182:281–294

Hendrickx PM, Gutmanas A, Kleywegt GJ (2013) Vivaldi: visualizaton and validation of biomacromolecular NMR structures from the PDB. Proteins 81:583–591

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

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

Huang YJ, Powers R, Montelione GT (2005b) 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 YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins. 62:587–603

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

Hyberts SG, Goldberg MS, Havel TF, 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

Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallogr A 32:922–923

Kabsch W (1978) A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr A 34:827–828

Kirchner DK, Güntert P (2011) Objective identification of residue ranges for the superposition of protein structures. BMC Bioinformatics 12:170

Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, 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

Liu G, Shen Y, Atreya HS, Parish D, Shao Y, Sukumaran DK, Xiao R, Yee A, Lemak A, Bhattacharya A, Acton TA, Arrowsmith CH, Montelione GT, Szyperski T (2005) NMR data collection and analysis protocol for high-throughput protein structure determination. Proc Natl Acad Sci USA 102:10487–10492

Losonczi JA, Andrec M, Fischer MW, Prestegard JH (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson 138:334–342

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

Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes B, Wright P, Wüthrich K (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. Pure Appl Chem 70:117–142

Moseley HN, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 9:635–642

Moseley HN, Monleon D, Montelione GT (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Methods Enzymol 339:91–108

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

Nabuurs SB, Spronk CA, Vuister GW, Vriend G (2006) Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. PLoS Comput Biol 2:e9

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 OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) NMR structure determination for larger proteins using backbone-only data. Science 327:1014–1018

Ramelot TA, Raman S, Kuzin AP, Xiao R, Ma LC, Acton TB, Hunt JF, Montelione GT, Baker D, Kennedy MA (2009) Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. Proteins 75:147–167

Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Gutmanas A, Güntert P, He Y, Herrmann T, Huang YJ, Jaravine V, Jonker HR, Kennedy MA, Lange OF, Liu G, Malliavin TE, Mani R, Mao B, Montelione GT, Nilges M, Rossi P, van der Schot G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, de Vries S, Vuister GW, Wu B, Yang Y, Bonvin AM (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20:227–236

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73

Snyder DA, Montelione GT (2005) Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. Proteins 59:673–686

Struyf A, Hubert M, Rousseeuw P (1996) Clustering in an object-oriented environment. J Stat Softw 1:1–30

Tejero R, Bassolino-Klimas D, Bruccoleri RE, Montelione GT (1996) Simulated annealing with restrained molecular dynamics using CONGEN: energy refinement of the NMR solution structures of epidermal and type-α transforming growth factors. Protein Sci 5:578–592

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

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59:687–696

Williams T, Kelley C (2011) Gnuplot 4.5: an interactive plotting program. http://gnuplot.info

Williamson MP, Havel TF, Wüthrich K (1985) Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. J Mol Biol 182:295–315

Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 31:3370–3374

Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269:592–610

Zweckstetter M, Bax A (2000) Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. J Am Chem Soc 122: 3791–3792