PROTEINS: Structure, Function, and Bioinformatics 66:778–795 (2007)

Evaluating Protein Structures Determined by Structural Genomics Consortia

Aneerban Bhattacharya,[1] Roberto Tejero,[1,2] and Gaetano T. Montelione[1]*
[1]*Center for Advanced Biotechnology and Medicine, Northeast Structural Genomics Consortium, Rutgers University and Robert Wood Johnson Medical School, Piscataway, New Jersey 08854*
[2]*Departmento de Quimica Fisica, Universidad de Valencia, Valencia, Spain*

ABSTRACT **Structural genomics projects are providing large quantities of new 3D structural data for proteins. To monitor the quality of these data, we have developed the protein structure validation software suite (PSVS), for assessment of protein structures generated by NMR or X-ray crystallographic methods. PSVS is broadly applicable for structure quality assessment in structural biology projects. The software integrates under a single interface analyses from several widely-used structure quality evaluation tools, including PROCHECK (Laskowski et al., J Appl Crystallog 1993;26:283–291), MolProbity (Lovell et al., Proteins 2003;50:437–450), Verify3D (Luthy et al., Nature 1992;356:83–85), ProsaII (Sippl, Proteins 1993;17: 355–362), the PDB validation software, and various structure-validation tools developed in our own laboratory. PSVS provides standard constraint analyses, statistics on goodness-of-fit between structures and experimental data, and knowledge-based structure quality scores in standardized format suitable for database integration. The analysis provides both global and site-specific measures of protein structure quality. Global quality measures are reported as Z scores, based on calibration with a set of high-resolution X-ray crystal structures. PSVS is particularly useful in assessing protein structures determined by NMR methods, but is also valuable for assessing X-ray crystal structures or homology models. Using these tools, we assessed protein structures generated by the Northeast Structural Genomics Consortium and other international structural genomics projects, over a 5-year period. Protein structures produced from structural genomics projects exhibit quality score distributions similar to those of structures produced in traditional structural biology projects during the same time period. However, while some NMR structures have structure quality scores similar to those seen in higher-resolution X-ray crystal structures, the majority of NMR structures have lower scores. Potential reasons for this "structure quality score gap" between NMR and X-ray crystal structures are discussed. Proteins 2007;66:778–795.**
© 2006 Wiley-Liss, Inc.

Key words: protein structure quality; ProsaII; PROCHECK; MolProbity; Verify3D; protein NMR structures; X-ray crystal structures

INTRODUCTION

Structural genomics aims to expand our understanding of protein structure and function by determining representative structures from protein sequence families for which no 3D structure is available. The goal of the United States Protein Structure Initiative (www.nigms.nih.gov/Initiatives/PSI) is to make the three-dimensional (3D) atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences. These efforts involve large-scale structure production, providing template 3D structures to be used as a basis for modeling members of structural domain families. Structural information about a protein can also sometimes help in elucidating its biochemical function by revealing distant homologous relationships that are not evident from sequence similarity information alone.[1]

In structural genomics efforts, and for structural biology in general, it is important to define broadly applicable and generally agreed upon measures of structure quality that can be compared for protein structures determined by different experimental and theoretical methods. Tools commonly used to evaluate different aspects of protein structure quality based on the knowledgebase of protein structural data include PROCHECK,[2,3] AQUA,[3] WHATIF/WHATCHECK,[4] MolProbity,[5] Verify3D,[6] ProsaII,[7] and the PDB validation software.[8] In addition, an assessment of how well the final

Abbreviations: 3D, three-dimensional; NESG, Northeast Structural Genomics Consortium; NIH, National Institutes of Health (USA); NMR, nuclear magnetic resonance; NOESY, nuclear overhauser effect spectroscopy; PDB, Protein Data Bank; PSI, protein structure initiative of the NIH; PSVS, protein structure validation software; rmsd, root mean square deviation; VdW, van der Waal interactions.

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.

Grant sponsor: NIH; Grant numbers: P50 GM62413, U54 GM074958; Grant sponsor: DGES (Spain); Grant number: PB98-1455.

*Correspondence to: Prof. Gaetano T. Montelione, CABM, Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854-5638.
E-mail: guy@cabm.rutgers.edu

Received 22 April 2006; Revision 5 July 2006; Accepted 6 July 2006

Published online 21 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21165

© 2006 WILEY-LISS, INC.

model (or ensemble of models) compares with the experimental data is an essential part of the structure validation process. In X-ray crystallography, structure quality can be globally assessed using the R-factor[9,10] and free R-factor[10,11] (a cross-validated R-factor), which measure the agreement between the structural model and experimental data. In NMR spectroscopy, a statistical summary of constraint violations, rmsd of atomic coordinates across the ensemble, and the numbers of constraints per residue are commonly used measures of structure quality. However, these measures of structural precision and accuracy have significant shortcomings (recently reviewed by Snyder et al.[12]). Recent efforts in NMR structure validation have included increased use of "NMR R factors"[13–15] to calculate the "goodness-of-fit" between the 3D protein NMR structures and experimental NOESY peak list or residual dipolar coupling data. However, even structures with reasonably good crystallographic R-factors or "NMR R-factors" exhibit local structural distortions that must be identified and assessed with respect to how well they are supported by the experimental data.

Structure quality assessment is especially critical for protein NMR structures. Some protein NMR structures show a range of local structural problems even though they exhibit low atomic rmsd across an ensemble, high numbers of constraints per residue, and "good values" of other commonly accepted quality factors.[12] It is quite easy to produce protein NMR structures with poor packing, and even incorrect folds, by misassignment of resonance frequencies and/or NOESY cross peaks. Such inaccuracies may not be detected by standard measures, such as residual constraint violations and atomic rmsd's. For example, Doreleijers et al.[16] have reported generally good correlations between different structural quality assessment measurements and experimental NMR data density. However, they also observed a significant number of outliers; for example in some cases, bond lengths, bond angles, and planarity of groups were found to deviate significantly from ideal values, while constraint violations and rmsd values appear quite good. Detection and analysis of unusual structural features is especially important in structure determinations using newly developed automated NMR data analysis methods.[17–21]

Homology models also exhibit many of these same types of inaccuracies, arising from misalignment of the target sequence to the homologous template sequence. The resulting structural inaccuracies can result in poor global or site-specific structural quality assessment scores and/or high conformational energies.[22,23] The process of X-ray crystal structure refinement, while primarily driven by the experimental data, can also be guided by analysis of local structural geometry and packing.[24]

Quality scores that are independent of experimental data, using only the final atomic coordinates as input, can be complementary to crystallographic and NMR R-factors for evaluating the quality of a structure. These knowledge-based methods[3–8] compare the value of particular structural parameters with those in amino acid, protein, or peptide structures determined with high resolution crystallographic data. Ideally, the parameters used for assessment should not have been constrained or optimized during refinement. These parameters can assess how 'normal' a structure is, compared with the standard values seen in accurate and precise structures. Useful quality assessment parameters include covalent geometry (bond length and bond angle), dihedral angle conformational distributions, nonbonded interactions, distributions of polar and nonpolar residues, hydrogen bonding patterns, and 3D packing profiles.

In this article, we present the Protein Structure Validation Software (PSVS) suite for consistent and rapid evaluation of the quality of protein structures, with a focus on NMR structures and homology models. PSVS provides a standardized set of quality scores and constraint analyses for each input structure. In addition to experimental constraints, this set encompasses a number of parameters evaluating different aspects of the structure, including fold and packing, local residue separation, deviations of bond lengths and bond angles, backbone and side-chain torsion angle stereochemistry, and steric overlaps between atoms. These data allow both global and site-specific structure quality assessments. A graphical user interface (GUI) runs the analysis and integrates information reported by several structure quality evaluation tools. Quality scores, calibrated with a set of high-quality X-ray crystal structures ($\leq$1.8 Å resolution), are summarized as Z scores for several structure validation analysis programs. The output consists of a standard set of tables and graphs and a concise validation report. The PSVS software is broadly applicable in structural biology projects. As a demonstration of the value of the PSVS server, we apply these tools to evaluate protein structures determined by different Structural Genomics Consortia, and compare the distributions of quality scores in these structures with X-ray crystal and solution NMR structures deposited in the PDB in recent years.

METHODS

Tools for Structure Quality Evaluation

PSVS incorporates published software developed by other research groups and in our own laboratory that have been integrated under a single graphical user interface. Table I summarizes the software tools supported by the current version of PSVS. PDBStat is a C++ program used to perform various statistical analyses given the Cartesian coordinates of a protein and a list of spatial constraints used to generate the structure (if available). The program is able to read and write coordinates and/or constraints in different standard formats (CONGEN,[29] XPLOR/CNS,[30,31] PDB,[8] and DYANA/CYANA[32]), handling the different hydrogen naming conventions, and can deal with proteins with multiple chains and/or models. PDBStat also produces a constraint satisfaction analysis for distance or dihedral con-

TABLE I. Tools Used by PSVS to Evaluate Different Aspects of Structure Quality

Tool(s)	Parameter(s) evaluated
PDBStat	Define ordered regions of the structure, and analyze numbers of conformationally-restricting constraints, violations of constraints, and rmsd of superimposed atomic coordinates
RPF[15]	Goodness-of-fit of protein NMR structure with NOESY peak list and resonance assignment data
DSSP[36]	Calculate secondary structure
PROCHECK G-factors[3]	Probability of dihedral angles of a residue type to be within a given range
MolProbity[5,26–28]	Calculate and visualize bad contacts, atomic overlaps, and Cβ position deviations
Verify3D[6]	Likelihood of the amino acid sequence to have the three-dimensional packing seen in the structure
ProsaII[7]	Pseudo energy of pair-wise interactions from the spatial separation of residues
PDB validation software[8]	Close contacts, deviations of bond lengths, and bond angles from ideality

straints, giving a summary with minimum, maximum, average, and root-mean-square violations, with violations classified in ranges. The program also provides a summary of experimental distance constraints, including the numbers of conformationally restricting distance constraints classified into intra and sequential (backbone/backbone, backbone/side-chain and side-chain/side-chain), long range, hydrogen bond, and disulfide bond constraints. PDBStat is also used to filter out conformationally nonrestricting intraresidue and sequential NOE constraints, if the constraint is too restricting, nonrestricting, or corresponds to a fixed distance, based on the ranges imposed by molecular geometry. In addition, it filters out duplicate constraints, and constraints for identical atom pairs with different bounds. PDBStat also generates contact maps based on coordinates or constraints, calculates atomic rmsd's for an ensemble of structures, and evaluates structural order parameters for backbone ϕ and φ dihedral angles[33] in order to assess how well local structure is defined across an ensemble of models. The program is also used to fit coordinates to a specified model, and translate and rotate coordinates to optimally superpose them for all or a selected set of atoms, over the average structure or an individual model. Some other functions of PDBStat include performing a simple close contact analysis, main chain and side chain (for Ile and Thr) chirality analysis, and an analysis of hydrogen bond satisfaction and classification (based on geometric parameters).

The PSVS report also includes RPF[15] "NMR R-factor" scores assessing the "goodness of fit" of the protein structure to NOESY peak list data. RPF calculates recall, precision and F-measures, statistical measures of the presence/absence of interproton distance relationships/NOESY peaks comparing the NMR-derived structure to NOESY peak list and resonance assignment data. Recall quantifies the fraction of short distance relationships expected from the NOESY data that are observed in the structure; the presence of NOESY peaks that cannot be explained by the 3D structure reduce the recall score. Precision quantifies the fraction of short distance relationships observed in the structure, which are consistent with the NOESY data; the absence of NOESY peaks expected for short interproton distances in the model structure reduce the Precision score. The F-measure combines the recall and precision scores to provide an assessment of overall fit between the experimental data and the model structure. The discriminating power (DP) score, a scaled F-measure, measures how well the available NOESY data distinguish the structure from a freely rotating chain model, and normalizes for the completeness of the NOESY data. A detailed description of the algorithms of RPF analysis are presented elsewhere.[15]

In addition to providing a GUI for PDBStat, PSVS integrates other structure validation programs under a single interface. PROCHECK_NMR[3] analyses the stereochemistry of the protein structure, and a number of other factors including distributions of backbone dihedral angles in the Ramachandran plot, bond lengths and bond angles, and locations of secondary structure elements. It also calculates a G-factor, based on the log-odds of observing specific backbone and side-chain dihedral angles for each residue type observed in the structure. The MolProbity[5,26–28,34] set of programs is used to regenerate the hydrogen atoms (using the *reduce* program), and calculate atomic clashes (using the *probe* program). It is also used to visualize any steric overlaps in the structure of the molecule (using the *prekin* program and *mage* modeler). Deviations in the positions of Cβ atoms, which are sensitive to incompatibilities between backbone and side-chain conformations,[5] are also calculated using the *prekin* program. Verify3D[6,35] compares the amino acid sequence with the 3D structure, and gives a score based on probability of finding a particular residue type in the environment observed in the structure. ProsaII[7,36] models a reduced-representation energy of pair wise interactions from the spatial separation of Cβ atoms of local residues. The validation software used by the Protein Data Bank is also used by PSVS to calculate close contacts, and to identify potential problems indicated by nonideal bond lengths, bond angles, and chirality.

GUI for Structure Quality Analysis

We developed a GUI, which runs on any web browser, to perform the structure quality analysis of a protein structure. The user provides a file containing atomic coordinates in the standard PDB format,[8] lists of con-

straints (for NMR structures), and some other information on the format of constraints and options for analysis of constraint violations. The user may also provide a file with secondary structure information, or choose to have this calculated using DSSP.[25] In addition, the user may optionally provide other annotating information about the protein, which will appear in the structure quality assessment reports.

The PSVS interface (version 1.2) runs the structure–quality analysis tools summarized in Table I, using a set of underlying Perl scripts. Supplementary Figure S1 shows a flow-chart describing this implementation. The analysis of a typical NMR structure and constraints, with 10 models of about 120 residues each, takes ~4 CPU minutes, while a typical X-ray structure of about 200 residues takes ~2 CPU minutes, on a Linux-based AMD Athlon 2.0 GHz CPU.

The output of PSVS includes of a Structure Quality Assessment Summary Table (see Fig. 1), together with a Concise Structure Quality Validation Report (see Fig. 2), providing a quick insight into key aspects of the structure quality assessment. PSVS provides plots for the regional variation of quality scores along the protein sequence, generated by the software listed in Table I. In addition, for multimodel NMR and homology-model structures, PSVS produces plots of dihedral angle order parameters,[29] and numbers of constraints per residue, along the protein sequence. The user can also view, through links provided by the Concise Structure Quality Validation Report, a large number of the detailed structure validation files produced by the individual validation programs that are integrated under PSVS.

Test Sets of X-Ray Crystal and NMR Structures

A sample of 557 X-ray crystal structures and 183 NMR structures that were deposited in the PDB by traditional structural biology groups between January 2000 and August 2005 was selected for assessing the utility of PSVS software tools for structure validation. The X-ray crystal structures are 50–500 residues long, have ≤50% sequence identity between each other, and are monomers or dimers. A restriction on sequence identity was imposed to avoid any bias toward the selection of structures belonging to particular families, which might contribute to a larger proportion of high-resolution structures in the PDB. Since the analysis is being performed primarily for NMR structures and homology models, which are usually of shorter length, a restriction on sequence length was imposed to avoid any bias from scores from longer structures in the training set. The X-ray crystal structures were classified into resolution-based groups, with the following definitions: (i) High resolution: resolution ≤1.80 Å, R factor ≤0.25, R-free ≤0.28 (252 structures); (ii) Medium resolution: resolution >1.80 Å and ≤2.50 Å, R factor ≤0.28, R-free ≤0.32 (244 structures); (iii) Low resolution: resolution >2.50 Å and ≤3.50 Å (61 structures). The NMR structures are 40–370 residues long, and are monomers. Protein structures included in each of these four sets (3 X-ray, 1 NMR) are listed in Supplementary Table S1. Protein structures deposited in PDB by structural genomics consortia were obtained from the TargetDB database,[37] and include 404 NMR and 985 X-ray crystal structures deposited in the PDB by structural genomics consortia between January 1, 2000 and August 31, 2005 (the end date of the first stage of the U.S. Protein Structure Initiative project). Protein structures included in each of these two sets (1 X-ray, 1 NMR) are listed in Supplementary Table S2.

PSVS Software Suite Availablity

The web-based PSVS server can be accessed at http://www-nmr.cabm.rutgers.edu/bioinformatics/index.html. Use of the server requires licenses for using associated software. The Perl scripts underlying PSVS as well as the PDBStat and RPF[15] programs used by PSVS are also available upon request from the authors.

RESULTS

Tools of PSVS

PSVS incorporates published software developed in our own laboratory and by other research groups, integrated under a single graphical user interface. These tools are summarized in Table I. Descriptions of the information generated in these analyses are presented in the Methods section.

Comparing Quality Scores to Scores From 'Good Quality' X-Ray Crystal Structures

A set of 252 X-ray crystal structures with resolution ≤1.80 Å, R-factor ≤0.25, and R-free ≤0.28 was taken from the PDB[8] (the high resolution set, as described in the Methods section) and used as a set of 'good' structures. The Kolmogorov-Smirnov[38] test was used to assess the normality of distributions of scores in this test set. Using a confidence level of $\alpha = 0.05$ (that is, probability of Type I error is 5%), the distribution of scores from each structure validation tool was found to be significantly close to normal. The mean and standard deviation from the distribution of scores for each tool[3,5–7,26] was then used to calculate the Z-score (independently, for each tool) for each structure being analyzed,

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

where for a particular tool, Z is the Z-score, X is the observed score for a given protein, μ is the mean score for that tool for the set of 'good' structures, and σ is the standard deviation for the scores of the 'good' structures for the same assessment tool.

Scores in Local Regions of the Structure

Global quality scores have limited utility in identifying the causes of strain and 'unusual' properties seen in the

Summary of conformationally-restricting constraints and structure quality factors

Summary of conformationally-restricting experimental constraints [a]

NOE-based distance constraints:

Total	1018
intra-residue [i = j]	147
sequential [| i - j | = 1]	311
medium range [1 < | i - j | < 5]	126
long range [| i - j | ≥ 5]	434
NOE constraints per restrained residue [b]	15.2

Hydrogen bond constraints:

Total	50
long range [| i - j | ≥ 5]	42

Dihedral-angle constraints:	151
Total number of restricting constraints [b]	1219
Total number of restricting constraints per restrained residue [b]	18.2
Restricting long-range constraints per restrained residue [b]	7.1
Total structures computed	56
Number of structures used	10

Residual constraint violations [a,c]

Distance violations / structure

0.1 - 0.2 Å	4.8
0.2 - 0.5 Å	3.9
> 0.5 Å	0.4
RMS of distance violation / constraint	0.02 Å
Maximum distance violation [d]	0.83 Å

Dihedral angle violations / structure

1 - 10 °	2.1
> 10 °	0
RMS of dihedral angle violation / constraint	0.20 °
Maximum dihedral angle violation [d]	2.30 °

RPF scores

Recall	Precision	F-measure	DP-score
0.978	0.916	0.946	0.794

RMSD Values

	all residues	ordered residues [e]
All backbone atoms	1.8 Å	0.4 Å
All heavy atoms	2.5 Å	1.0 Å

Structure Quality Factors - overall statistics

	Mean score (all models)	Standard Deviation	Z-score [f]
Procheck G-factor [e] (phi / psi only)	-0.66	N/A	-2.28
Procheck G-factor [e] (all dihedral angles)	-0.48	N/A	-2.84
Verify3D	0.37	0.0259	-1.44
ProsaII (-ve)	0.42	0.0726	-0.95
MolProbity clashscore	27.64	4.1668	-3.22

Ramachandran Plot Summary from Procheck [e]

Most favoured regions	90.9%
Additionally allowed regions	9.1%
Generously allowed regions	0.0%
Disallowed regions	0.0%

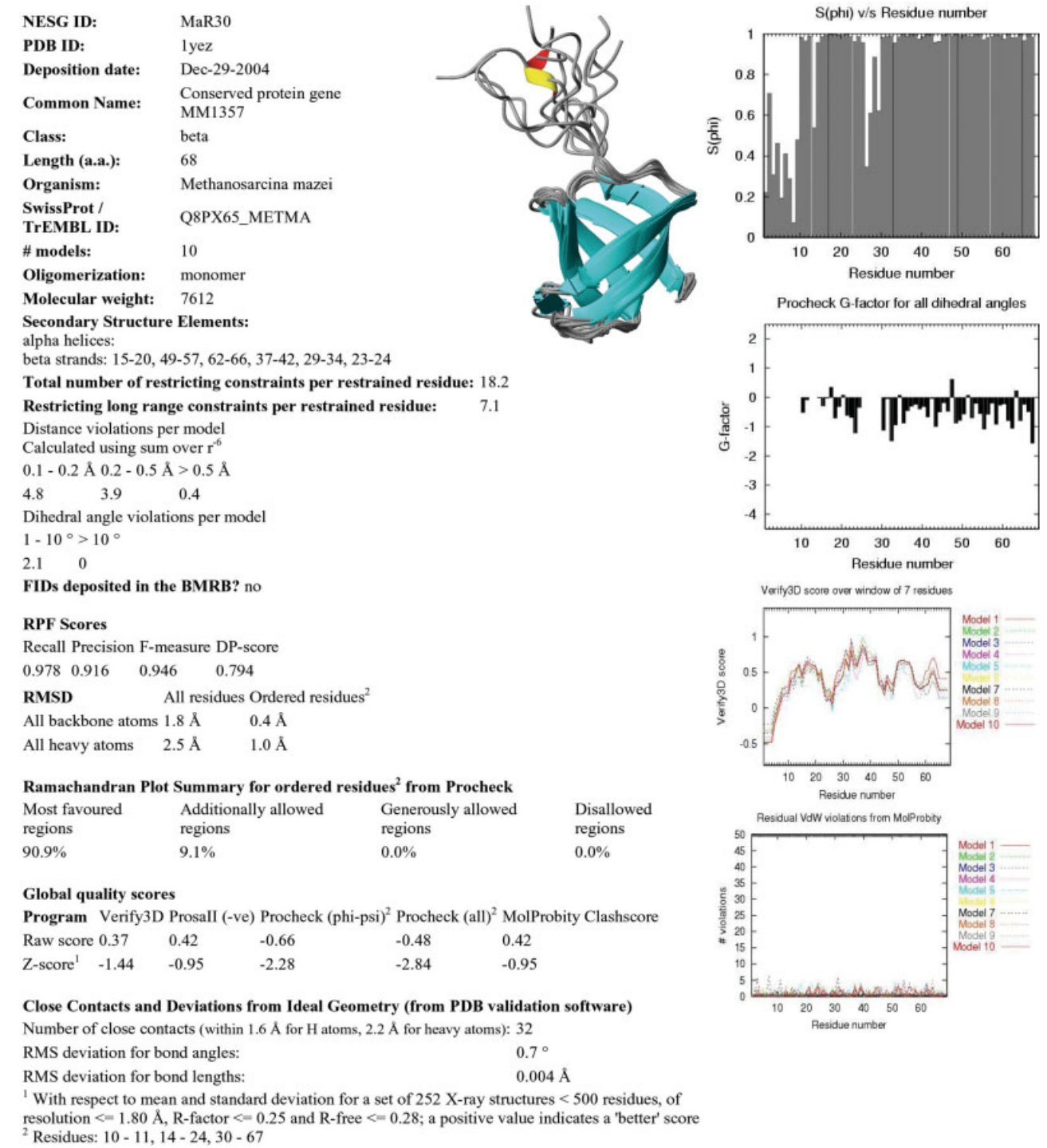
[a] Analysed for residues 1 to 68
[b] There are 67 residues with conformationally restricting constraints
[c] Calculated for all constraints for the given residues, using sum over r^{-6}
[d] Largest constraint violation among all the reported structures
[e] Residues with sum of phi and psi order parameters > 1.8

Ordered residue ranges: 10 - 11, 14 - 24, 30 - 67

[f] With respect to mean and standard deviation for for a set of 252 X-ray structures < 500 residues, of resolution <= 1.80 Å, R-factor <= 0.25 and R-free <= 0.28; a positive value indicates a 'better' score

Generated using PSVS 1.1

Fig. 1. PSVS structure quality assessment summary table.



NESG ID: MaR30
PDB ID: 1yez
Deposition date: Dec-29-2004
Common Name: Conserved protein gene MM1357
Class: beta
Length (a.a.): 68
Organism: Methanosarcina mazei
SwissProt / TrEMBL ID: Q8PX65_METMA
 # models: 10
Oligomerization: monomer
Molecular weight: 7612
Secondary Structure Elements:
alpha helices:
beta strands: 15-20, 49-57, 62-66, 37-42, 29-34, 23-24
Total number of restricting constraints per restrained residue: 18.2
Restricting long range constraints per restrained residue: 7.1
Distance violations per model
Calculated using sum over r^{-6}

0.1 - 0.2 Å	0.2 - 0.5 Å	> 0.5 Å
4.8	3.9	0.4

Dihedral angle violations per model

1 - 10 °	> 10 °
2.1	0

FIDs deposited in the BMRB? no

RPF Scores

Recall	Precision	F-measure	DP-score
0.978	0.916	0.946	0.794

RMSD	All residues	Ordered residues[2]
All backbone atoms	1.8 Å	0.4 Å
All heavy atoms	2.5 Å	1.0 Å

Ramachandran Plot Summary for ordered residues[2] from Procheck

Most favoured regions	Additionally allowed regions	Generously allowed regions	Disallowed regions
90.9%	9.1%	0.0%	0.0%

Global quality scores

Program	Verify3D	ProsaII (-ve)	Procheck (phi-psi)[2]	Procheck (all)[2]	MolProbity Clashscore
Raw score	0.37	0.42	-0.66	-0.48	0.42
Z-score[1]	-1.44	-0.95	-2.28	-2.84	-0.95

Close Contacts and Deviations from Ideal Geometry (from PDB validation software)
Number of close contacts (within 1.6 Å for H atoms, 2.2 Å for heavy atoms): 32
RMS deviation for bond angles: 0.7 °
RMS deviation for bond lengths: 0.004 Å

[1] With respect to mean and standard deviation for a set of 252 X-ray structures < 500 residues, of resolution <= 1.80 Å, R-factor <= 0.25 and R-free <= 0.28; a positive value indicates a 'better' score
[2] Residues: 10 - 11, 14 - 24, 30 - 67

Fig. 2. Abbreviated PSVS concise structure quality validation report.

structure. A regional score, either per-residue or based on a multiple-residue window, can identify the regions of a structure that contribute to a poor score. Care must be taken in interpreting such local structure quality scores, as packing and dihedral angle distributions are expected to be poorer for residues that are disordered or otherwise not well-defined in the structural model(s).

Representative Structure Quality Assessment Reports

Given atomic coordinates, together with constraint lists and NOESY peaks list (where available), PSVS generates two standardized structure quality reports, including a concise structural quality validation report and a structure quality assessment table. Examples of

these two reports are described in the following two sections.

Solution NMR structure of NESG target MaR30—a bacterial TRAM domain (1yez)

The *Methanosarcina mazei* protein (SwissProt ID: Q8PX65_METMA; PDB ID: 1yez; NESG ID MaR30) is a small 68-residue β-barrel protein targeted by the Northeast Structural Genomics (NESG) Consortium. The PSVS Structure Quality Assessment Summary Table for MaR30/1yez, suitable for direct use in publications, is shown in Figure 1, and a Concise Structural Quality Validation Report, designed for web site diplay, is shown in Figure 2. A more extensive and detailed report for the structure generated by PSVS is available on line at http://spine.nesg.org:7000/gallery/reports/MaR30/reports/fsvr/MaR30.html.

Using standardized methods (described in the Methods section), PSVS reports that the MaR30 structure was determined using 18.2 conformationally-restricting constraints per restrained residue, with 7.1 restricting long-range constraints per restrained residue. The resulting NMR structure has atomic rmsd of 0.4 Å for ordered backbone atoms (where ordered residues are defined by the dihedral angle order parameter,[33] as described in the Methods section). MaR30 has RPF[15] scores, which compares the 3D structure with the NOESY peak list, of 0.978 for Recall, 0.916 for Precision, F-measure of 0.946, and DP score of 0.794, values indicative of a good-quality NMR structure. MaR30 has Z-scores of −1.44 for Verify3D, −0.95 for ProsaII, −2.28 for PROCHECK G-factor (ϕ–ψ), −2.84 for PROCHECK G-factor (all dihedrals), and −0.95 for MolProbity clashscore. The PDB validation tool reports an rmsd from ideal values of 0.7° for bond angles and 0.004 Å for bond lengths. Overall, these statistics indicate that the structure is very good quality compared with typical NMR structures that we have analyzed. Except in polypeptide segments of residues 1–14 and 25–29, which are poorly defined with few experimental constraints, the structure shows good global and local structure quality scores (see Fig. 2). The loops of residues 44–48 and 57–59, between β strands, have poorer Verify3D and ProsaII scores, although the dihedral angle distributions are satisfactory in these regions, and steric clashes, identified by MolProbity, are not significant.

X-ray crystal structure of NESG target HR2118—human 60S ribosome biogenesis protein NIP7 (1t5y)

The 2.5 Å crystal structure of the 165-residue human 60S ribosome subunit biogenesis protein NIP7 (SwissProt ID: Q9Y221; PDB ID: 1t5y; NESG ID: HR2118) is an α + β fold protein. A Concise Structure Quality Validation Report for HR2118/1t5y is included as Supplementary Figure S2. The structure was refined to R-factor of 0.213 and R-free of 0.294. Z-scores for this structure are −0.16 for Verify3D, 0.12 for ProsaII, −1.22 for PROCHECK (ϕ–ψ only), −1.66 for PROCHECK (all dihedrals), and −2.40 for MolProbity clashscore. The PDB validation software reports an rmsd from ideal geometry of 1.2° for bond angles and 0.008 Å for bond lengths. These global scores are typical of a medium-resolution X-ray crystal structure. A more detailed look at the regional variation of scores identifies a region with strain around residues 54–60 (loop before α helix), which is identified by all the quality evaluation tools. In addition, residues 22–24 (loop between α helices) and residues 42–44 (at the start of a β strand) have poor PROCHECK scores, while residues 106–110 (end of an α helix) have poor Verify3D and ProsaII packing scores. Such structure quality assessment data can potentially be used to further refine the 3D structure by iterative model correction and refinement of diffraction phases.[24]

Database of PSVS Reports for the NESG Consortium

Concise Structure Quality Validation Reports for all protein structures determined by the NESG Consortium are accessible through a link in the NESG Structure Gallery at http://www.nesg.org. These reports are generated for all structures deposited in the PDB by the NESG. They provide an archive documenting a standard structure quality analysis for all NESG protein structures. The NESG Structure Gallery also allows one to access extensive and detailed validation reports generated by PROCHECK, MolProbity, ProsaII, Verify3D, and other programs for each NESG protein structure.

Distributions of Crystallographic R Factors and PSVS Z-Scores for Proteins Determined by Structural Genomics Projects

Distributions of X-ray crystal structure R factors and PSVS structure quality Z-scores determined using PROCHECK G-factor (ϕ–ψ backbone only), PROCHECK G-factor (all dihedral angles), MolProbity clashscore, Verify3D, and ProsaII are compared in Figures 3–5 for protein structures deposited in the PDB over the time period January 1, 2000 through August 31, 2005. Included in this analysis were (i) protein structures determined by the NESG consortium in phase 1 of the NIH protein structure initiative (PSI-1) project (115 X-ray and 85 NMR structures) (ii) structures determined by all other international structural genomics (SG) consortia (870 X-ray and 319 NMR structures) during this time period, and (iii) a random sample of protein structures determined by traditional structural biology groups (non-SG) and deposited into the PDB during the same time period (557 X-ray and 183 NMR structures). Structures were classified into eight groups, defined in Table II, based on crystallographic resolution and method of determination.

R-factors computed for X-ray crystal structures deposited in the PDB by structural genomics consortia have distributions similar to those of high- and medium-resolution X-ray crystal structures generated in traditional structural biology efforts (see Fig. 3). In particular, X-ray crystal structures produced by the NESG consortium

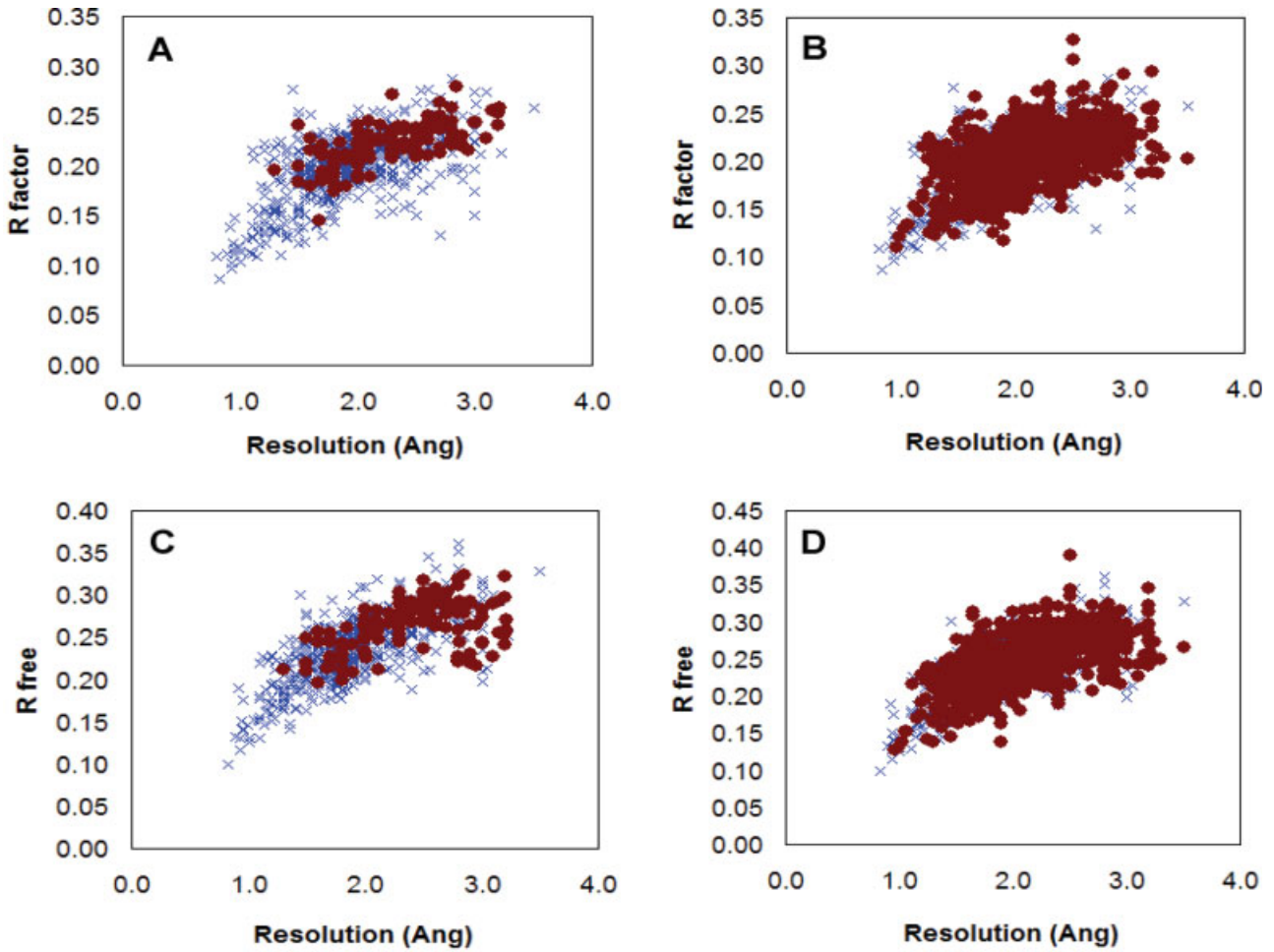


Fig. 3. X-ray crystal structure resolution vs. R and R_{free}. Crystallographic R factor (**A**) and R_{free} (**C**) vs. data resolution for a representative set of 557 non-SG X-ray crystal structures (blue crosses) compared with data for 115 X-ray crystal structures determined by the NESG consortium in PSI-1 (brown circles). Crystallographic R factor (**B**) and R_{free} (**D**) vs. data resolution for the same 557 non-SG X-ray crystal structures (blue crosses) compared with data for 985 structures from structural genomics consortia deposited through August 31, 2005 (brown circles).

[Figs. 3(A,C)] have R-factor vs. resolution and R-free vs. resolution plots similar to those for structures produced in traditional structural biology projects. NESG crystal structures have been mostly solved with data at limiting resolutions of 1.5–3.5 Å, while other structural genomics and traditional structural biology groups have a higher percentage of X-ray crystal structures solved with diffraction data at <1.5 Å.

Distributions of PSVS structure quality scores for these same protein X-ray crystal structures are compared in Figure 4, and mean values and standard deviations for these distributions are tabulated in Table II. The SG and non-SG X-ray crystal structures were classified as (i) high-resolution (< 1.8 Å), (ii) medium-resolution (1.8–2.5 Å), and (iii) low-resolution (2.5–3.5 Å). Overall, within each of these three resolution classes, the distributions of the PSVS structure quality Z-scores are similar for NESG, SG, and non-SG X-ray crystal structures.

The analyses presented in Figure 4 and Table II also document the sensitivities of the various structure quality assessment scores in distinguishing between low-, medium-, and high-resolution X-ray crystal structures. PROCHECK G-factors [Figs. 4(A,B)] and MolProbity clashscores [Fig. 4(C)] can distinguish high-, medium-, and low-resolution crystal structures, indicating that these are indeed sensitive measures of X-ray crystal structure accuracy. PROCHECK G-factors for all dihedral angles [Fig. 4(B)] provide a somewhat stronger correlation with crystal structure resolution than those for backbone dihedrals alone [Fig. 4(A)]. On the other hand, the distributions of ProsaII scores are similar across all categories of X-ray crystal structures, demonstrating that this metric has limited value in distinguishing between different classes of proteins that have correct folds [Fig. 4(D)]. ProsaII and Verify3D scores are, however, able to distinguish proteins modeled with correct folds from those intentionally modeled with incorrect folds (A. Bhattacharya, Z. Wunderlich, R. Tejero, and G. T. Montelione, manuscript in preparation).

Figure 4 also documents the distributions of PSVS Z-scores for NMR structures produced by SG and tradi-

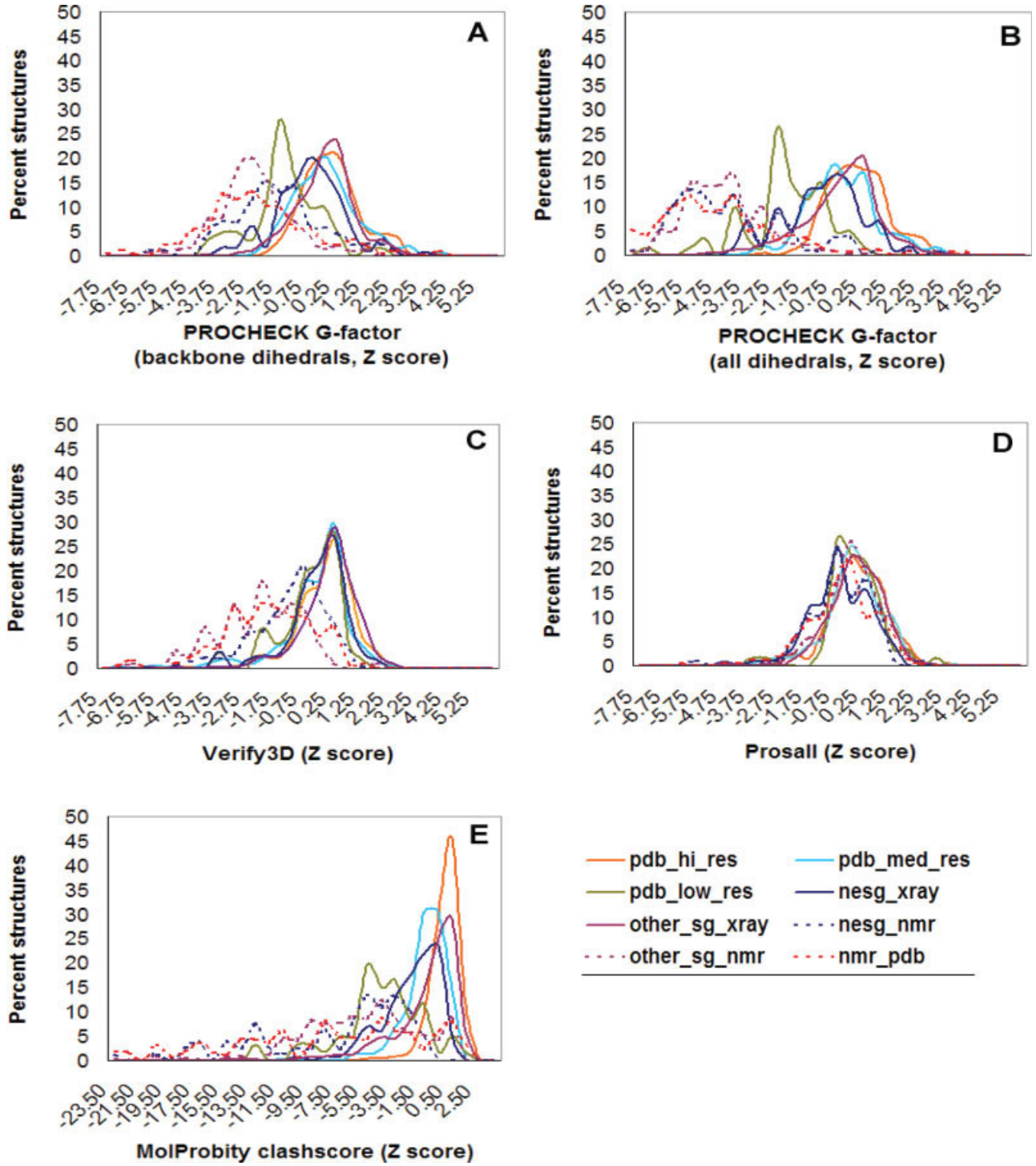


Fig. 4. Distributions of PSVS Z scores for different structure validation tools. Shown are the distribution of scores for each of the eight classes of structures described in Table II for (**A**) PROCHECK G-factor (backbone φ–ψ), (**B**) PROCHECK G-factor (all dihedrals), (**C**) Verify3D, (**D**) Prosall, (**E**) MolProbity clashscore. For all plots, a few structures whose scores are more than 6 standard deviations away from the mean score of high-resolution X-ray structures (18 s.d. for MolProbity clashscore) are excluded to provide clarity.

tional structural biology projects. The higher sensitivity of PROCHECK G-factors (all dihedrals) and MolProbity clashscore in distinguishing among structures is particularly evident in these data for NMR structures. While some NMR structures have structure quality scores similar to low- and medium-resolution crystal structures, most exhibit PROCHECK, MolProbity, and (to a lesser degree) Verify3D structure quality scores that are lower

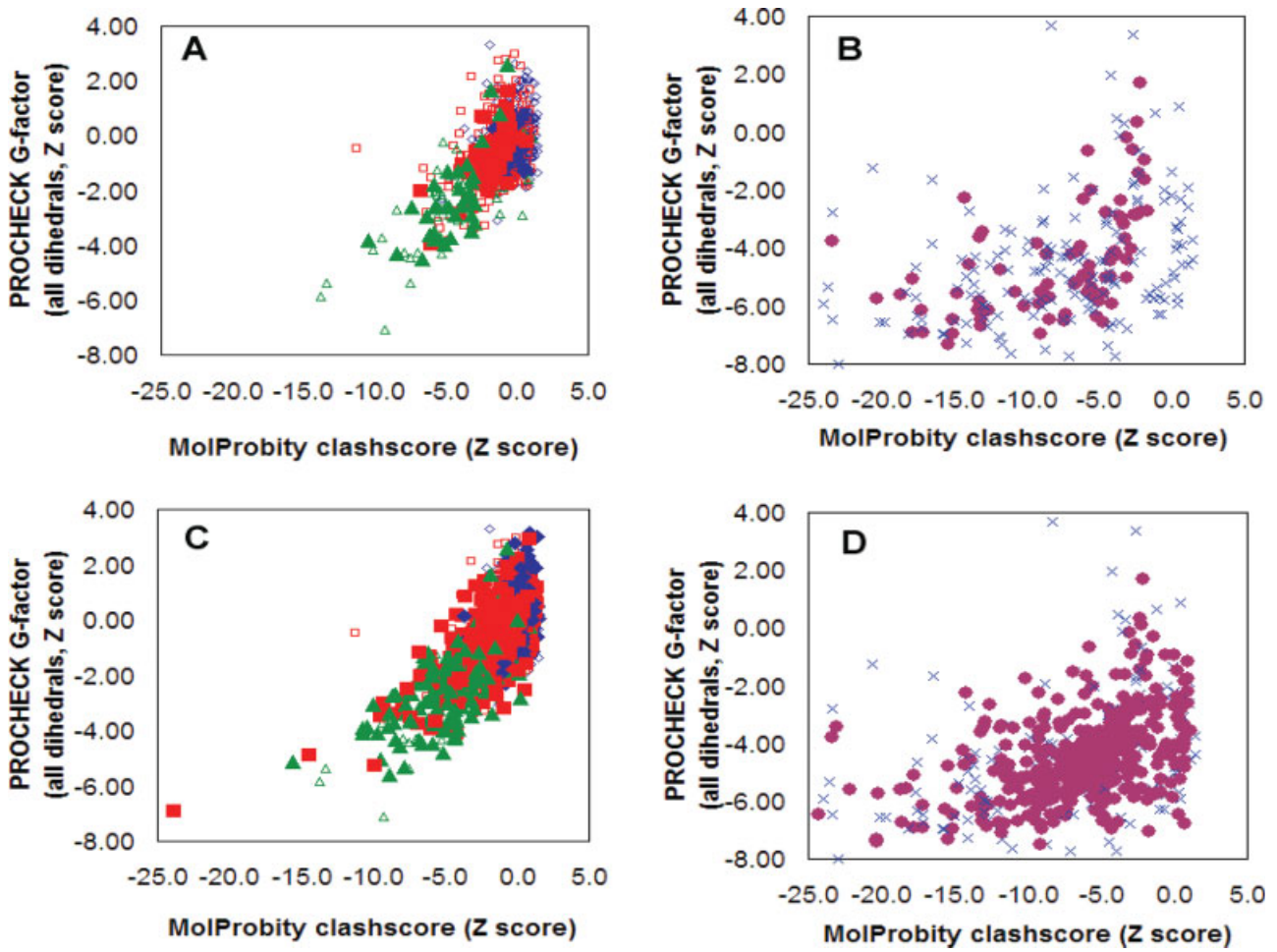


Fig. 5. Bivariant plots of PROCHECK G-factor (all dihedrals) and MolProbity clashscore for (**A**) 115 X-ray crystal structures determined by NESG and (**C**) 870 X-ray crystal structures determined by other international structural genomics consortium, each compared with a representative set of 557 non-SG X-ray structures. Structures were classified into high-resolution (blue diamonds), medium-resolution (red squares), and low-resolution (green triangles) classes, as defined in footnote *a* of Table II. The filled shapes denote SG structures, and open shapes denote the non-SG crystal structures. Also shown are bivariant plots for (**B**) 85 NMR structures determined by NESG in PSI-1or (**D**) 319 NMR structures determined by other international structural genomics consortia, compared with a representative set of 183 NMR structures determined by non-SG groups in the same time period. The filled brown circles represent the SG structures, and the blue crosses the non-SG structures.

than those of X-ray crystal structures. PROCHECK (all dihedral angles) Z scores for NMR structures, determined only for ordered residues as described in the Methods section, exhibit a wider distribution of scores than PROCHECK scores based on backbone dihedrals alone [*cf.* Figs. 4(A,B)]. This observation is consistent with the conclusion reached with X-ray crystal structures that PROCHECK G-factors for all dihedral angles have a somewhat stronger correlation with crystallographic resolution than those for backbone dihedral angles alone. MolProbity clashscores for NMR structures exhibit an even wider distribution than the PROCHECK G-factor scores [Fig. 4(E)]. ProsaII scores, however, do not distinguish between these NMR and X-ray crystal structures [Fig. 4(D)]. Overall, these results for both X-ray and NMR structures demonstrate that PROCHECK G-factor (for all dihedral angles) and MolProbity clashscores are the most sensitive of the knowledge-based Z-scores computed by PSVS for distinguishing among the NMR structures.

In the three most-sensitive measures [PROCHECK G-factor (backbone), PROCHECK G-factor (all dihedrals), and MolProbity clashscore], NMR structures generally exhibit scores comparable to those of low-resolution (2.5–3.5 Å) X-ray structures (and in some cases, the scores for NMR structures are much lower than these). Some NMR structures have scores similar to those obtained for 1.8–2.5 Å X-ray structures. These data document a significant gap in quality scores of NMR and X-ray crystal structures that have been deposited in the PDB over the last several years. In contrast, some of these NMR structures are for proteins which have never been successfully crystallized, and thus represent the only "high resolution" structural data obtainable. On the

TABLE II. Comparison of Mean PSVS Z Scores for Different Groups of Structures

Group[a]	PROCHECK G-factor (φ–ψ)	PROCHECK G-factor (all dihedrals)	Verify3D	ProsaII	MolProbity clashscore
pdb_hi_res	0.00 (1.00)[b]	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)
pdb_med_res	−0.18 (1.14)	−0.53 (1.19)	−0.43 (1.28)	−0.17 (1.03)	−1.39 (1.54)
pdb_low_res	−1.57 (1.21)	−2.46 (1.42)	−0.59 (0.99)	−0.14 (0.99)	−4.51 (3.10)
nesg_xray	−0.73 (1.20)	−1.21 (1.38)	−0.38 (1.06)	−0.70 (1.05)	−2.45 (2.07)
other_sg_xray	−0.22 (1.03)	−0.58 (1.34)	0.09 (0.99)	−0.09 (0.93)	−1.28 (2.44)
nesg_nmr	−1.96 (1.57)	−4.64 (2.04)	−2.36 (1.48)	−0.68 (1.12)	−8.94 (7.52)
other_sg_nmr	−2.64 (1.31)	−4.71 (1.51)	−1.31 (1.24)	−0.17 (0.92)	−6.59 (6.02)
pdb_nmr	−2.55 (1.97)	−5.08 (2.50)	−2.15 (1.74)	−0.62 (1.26)	−10.74 (10.36)

[a]Protein structures deposited in the Protein Data Bank between Jan 1, 2000 and Aug. 31, 2005, divided into the following groups (additional selection details are provided in the Methods section):
pdb_hi_res: 252 non-SG X-ray crystal structures refined to resolution ≤ 1.80 Å, R-factor ≤ 0.25, R-free ≤ 0.28.
pdb_med_res: 244 non-SG X-ray crystal structures refined to 1.80 Å < resolution ≤ 2.50 Å, R-factor ≤ 0.28, R-free ≤ 0.32.
pdb_low_res: 61 non-SG X-ray crystal structures refined to 2.50 Å < resolution ≤ 3.50 Å.
other_sg_xray: 870 X-ray crystal structures determined by structural genomics consortia other than NESG.
nesg_xray: 115 X-ray crystal structures determined by NESG.
other_sg_nmr: 319 NMR structures determined by structural genomics consortia other than NESG.
nesg_nmr: 85 NMR structures determined by NESG.
pdb_nmr: 1 83 NMR structures determined by non-SG groups.
[b]Mean value (standard deviation).

basis of the data summarized in Figure 4 and Table II, we conclude that, on average, the NMR structures generated by structural genomics consortia, including the NESG, have structure quality scores comparable to NMR structures deposited into the PDB by traditional structural biology groups in 2000–2005.

Bivariant Plots of PROCHECK G (all dihedral) vs. MolProbity Clashscore

Considering the observation that the MolProbity clashscore and PROCHECK G-factor (all dihedrals) are the most sensitive of our Z-scores in distinguishing between X-ray crystal structures refined to different resolutions, we next examined the correlations between MolProbity clashscore, PROCHECK G-factor (all dihedrals), and X-ray crystal structure resolution [Figs. 5(A,C)]. In these bivariant plots, the high- (blue) and medium- (red) resolution X-ray structures cluster together, with Z scores ranging from −5 to +3. The highest-resolution X-ray structures (blue) tend to cluster in the upper right corner of the plot, with highest Z values. Lower-resolution X-ray structures by contrast are scattered in regions of poorer values (Z < −2). This analysis demonstrates the value of these bivariant plots for distinguishing low-resolution from medium-/high-resolution X-ray crystal structures.

NMR structures were also assessed using these bivariant PROCHECK G (all dihedral) vs. MolProbity clashscore plots [Figs. 5(B,D)]. Shown in Figure 5(B) are the distributions of non-SG NMR structures together with the complete set of NMR structures determined by the NESG consortium in PSI-1; in Figure 5(D) are distributions for all SG NMR structures deposited in the same time period, together with the same set of non-SG structures. These distributions of quality scores are similar for NESG, SG, and non-SG NMR structures; 15–25% of NMR structures fall in the region of these bivariant plots (Z > −5) typical of medium- and high-resolution crystal structures, while the rest have scores similar (or much poorer than) lower resolution X-ray crystal structures. Potential reasons for these significant differences in PSVS Z-scores between X-ray crystal and NMR structures are addressed in the Discussion section.

Site-Specific Structure Quality Assessment

In addition to providing several global quality scores, the PSVS report provides extensive site-specific quality assessment information (some shown in Fig. 2). These include assessments of atomic clashes, detailed summaries of NMR constraint violations, and the RPF precision score,[15] which identifies short interproton distances in the NMR structure, which are not supported by corresponding data in the NOESY peak lists. Such site-specific quality scores are particularly valuable not only for structure quality assessment[39] but also as a guide for structure refinement.

An illustrative example of the value of the PSVS analysis, and the use of site-specific quality assessment data, involves the NMR structure refinement of the human dynein light chain 2A (hDLC-2A), one of the targets for the NESG consortium (NESG target ID HR2106). hDLC-2A was initially solved as a monomeric structure, and deposited in the PDB (PDB ID, 1tgq) prior to the introduction of PSVS analysis in the NESG structure production pipeline. However, subsequent to depositing the 1tgq structure into the PDB, we learned that a second group had solved the 3D structure of the same hDLC-2A protein[40] (PDB ID: 1z09), which is in fact a homodimer. Inspection of our gel filtration and light scattering data for hDLC-2A revealed that it is indeed largely dimeric in solution under the conditions of the NMR analysis, and indicated that the monomeric structure described by

PDB id 1tgq is incorrect. Accordingly, the structure of hDCL-2A was reanalyzed using X-filtered NOESY data (G. Liu, R. Xiao, G. T. Montelione, and T. Szyperski, in preparation), which distinguish intrachain from interchain NOEs, and this revised dimeric structure was deposited into the PDB (PDB ID 2b95) as a replacement for 1tfq. This same 1tfq monomer structure was subsequently used as an example of the challenges encountered by NMR studies in reliably distinguishing symmetric dimer from monomer structures using standard NMR methods.[39]

Although not yet in use by the NESG at the time 1tgq was deposited into the PDB, the PSVS analysis did flag problems with this monomeric structure of hDCL-2A. PSVS global Z-scores for PROCHECK G-factor (all dihedrals) and MolProbity clashscore are −6.7 and −13.1, respectively, for the incorrect 1tgf monomer structure, and −4.5 and −3.5, respectively, for the correct (2b95) dimer structure. Similarly, the global RPF fold discrimination score (DP score) is 0.69 for the monomer (a borderline DP quality score[15]), and increases to 0.75 for the dimer structure using the same NOESY peak list data. However, such global measures have limited sensitivity since the overall fold of hDCL-2A is similar in both the incorrect monomer and the dimer structures. As is shown in Figure 6, site-specific structure quality assessment reveals even more significant differences between the incorrect monomer and correct dimer structures. In particular, helix 2 (residues 45–71), which has antiparallel interactions across the dimer interface, is significantly distorted in the incorrect monomer structure of hDLC-2A, resulting in extensive local clashes [indicated in Fig. 6(A)] and poor local RPF precision scores [i.e. short distances in the structure not consistent with the NOESY peak list data, Fig. 6(C)]. This helix is not distorted in the dimer structure, as the NOEs, which caused the distortion, were determined to arise from interchain interactions. The corresponding site-specific analyses of the dimer structure, shown on the coordinates of one monomer from the dimer [Figs. 6(B,D)], shows many fewer clash and RPF violations. Thus, the incorrect monomer structure of hDCA-2A is clearly flagged by PSVS analysis; while the dimer structure still exhibits some minor clashes and precision violations, its site-specific quality scores are greatly improved. Although it is possible to use energy refinement with molecular dynamics to improve the MAGE clash scores, Precision and Recall violations[15] (assessing how well interproton distances are explained by NOESY peak list data) remain even after extensive molecular dynamics "refinement", and are sensitive in distinguishing the incorrect monomer from correct dimer structures. Evidently, had PSVS been in use by the NESG consortium at the time of this PDB deposition, it could have prevented deposition of the incorrect monomer structure.

DISCUSSION

Standardized Structure Quality Assessment

A primary requirement for a structure validation method is that it provides an objective evaluation of structure quality using criteria that are largely independent of the methods used to generate the structure. Validation must also include assessments of how well the structure satisfies the experimental data from which it is derived, such as crystallographic R-factors, NMR RPF scores,[15] and constraint violation analyses. Validation may also involve comparing structural parameters (e.g., bond length and bond angle deviations, hydrogen bond geometry, and core packing) with respect to the corresponding values observed in high-resolution and accurate protein structures. The parameters assessed should span multiple measures of experimental consistency, as well as a wide range of knowledge-based structural assessment criteria, since inaccurate structures may score well using certain measures of structural reliability, but poorly in other measures.

The PSVS software suite and server presented in this article provides a single-point interface to assess structure quality for protein structures. It performs this analysis with a standardized set of tools, covering a wide range of different characteristics of a protein structure that can be used to evaluate its quality. These analyses provide benchmarks applicable for assessing structures determined by X-ray crystallography and NMR spectroscopy. The PSVS software runs fast, and is easy to use. It is accessible either as stand-alone software, or as an internet server. Using simple input, including the PDB coordinates and (for NMR structures) the NMR constraint files and NOESY peak lists, the interface generates a standardized set of reports. Where appropriate (e.g., for NMR structures), residues that are not well-defined across the ensemble of structures reported as the "PDB structure" are excluded from PROCHECK and other analyses. For some key structure quality assessment tools, PSVS reports Z-scores normalized with scores obtained for a set of high-resolution X-ray crystal structures; these normalized Z scores facilitate comparison of these different methods of structure quality assessment. Extensive structural statistics specific to NMR structures (e.g. constraint violations, restraining constraints per residue, superimposed rmsd values, etc.) are determined and reported in a consistent and standardized fashion. PSVS groups and organizes all of these data and reports in one location, making them easy to survey and comprehend. Key structure quality parameters are also output in eXtensible Markup Language (XML) format, suitable for export into database programs. In addition to global quality scores for the entire protein, scores along a residual window, or on a per-residue basis, show regions of possible strain in the structure (e.g., localized poor fit to the experimental NOESY data) that contribute to its poorer quality assessment scores. The wide range of information presented in these reports provides a comprehensive view of different aspects of quality for the structure under study.

Using Information From PSVS

Scores from the different structure quality evaluation tools, along with the detailed reports that are also run

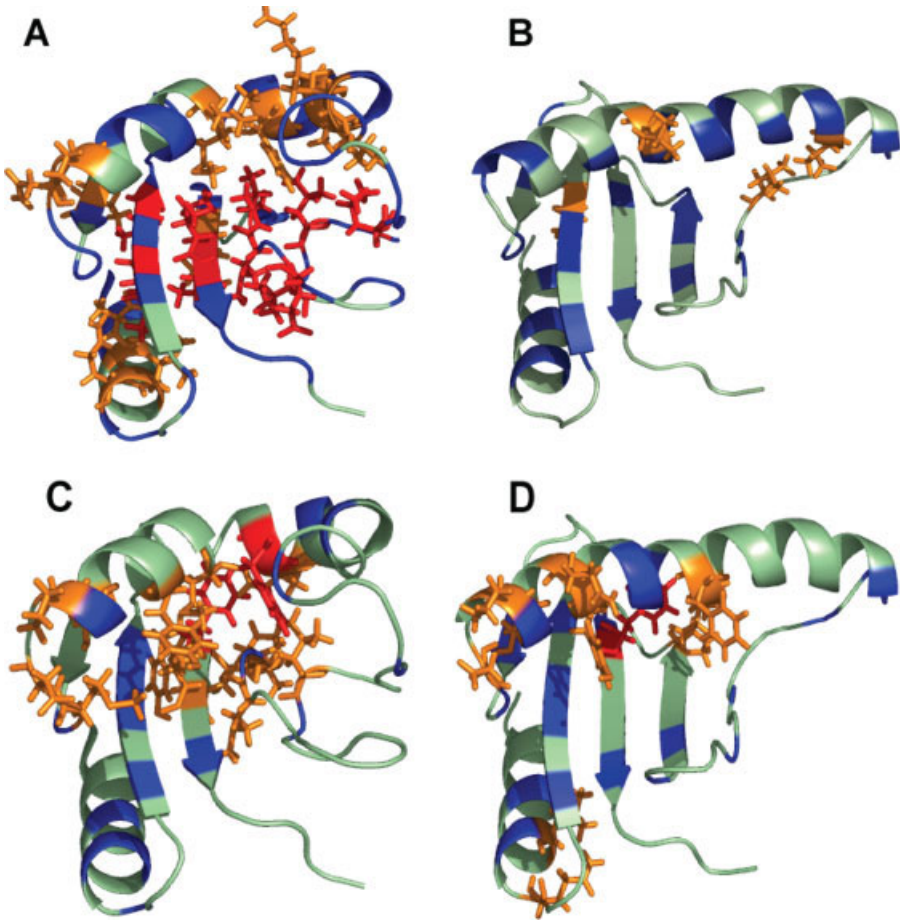


Fig. 6. Site-specific structure quality assessments for hDLC-2A. Numbers of severe VdW clashes (> 0.4 Å) per residue for (**A**) incorrect monomer structure (PDB ID 1tgq) and (**B**) correct dimer structure (PDB ID 2b95). Residues are colored red (8 or more severe VdW clashes/residue), orange (5–7 VdW clashes/residue), blue (1–4 clashes/residue) or green (no clashes). Side chains are also shown for those residues (red and orange) with largest numbers of VdW clashes. Numbers of interproton precision violations (short distances with no supporting data in the NOESY peak list[15]) for (**C**) incorrect monomer and (**D**) correct dimer structures are shown. Residues are color coded to indicate extensive (red), moderate (orange), minimal (blue) numbers of RPF precision violations, or essentially no precision violations (green); side chains are shown for residues with moderate or extensive numbers of precision violations.

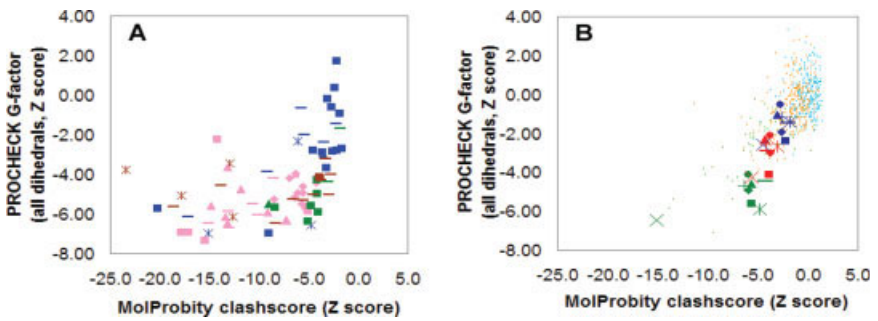


Fig. 7. Bivariant plots of PROCHECK G-factor (all dihedrals) and MolProbity clashscores for different NMR structure determination protocols. (**A**) Comparison of different protocols for NOE assignment and structure refinement. The shapes denote structures determined using different methods for NOESY cross peak assignment, including AutoStructure[20] (square), Candid[19] (triangle), Aria[17] (cross), consensus of AutoStructure and Candid[52] (diamond), and manual (dash) NOESY assignment protocols. The colors denote molecular mechanics programs used for structure refinement, including DYANA/CYANA[32] (pink), XPLOR[31] (brown), NIH-XPLOR[53] (green), and CNS[30] (blue). (**B**) Monitoring the refinement of 8 protein NMR structures, adopted from Liu et al.[52] The shapes represent different NESG target proteins. The colors refer to the different programs used in the final refinement, including CYANA 2.0 (green), CNS with energy minimization in vacuum (red), CNS with explicit solvent (blue). The small dots are the same non-SG protein crystal structure data shown in Figure 4, including the *pdb_hi_res* (lt. blue), *pdb_med_res* (gold), and *pdb_low_res* (green) data sets described in Table II.

and organized by PSVS, can be used to assess how much specific structural parameters deviate from values expected for accurate structures. In addition to evaluating the quality of the final structural model, PSVS is tremendously useful for assessing structure quality during the process of protein structure refinement. For example, Verify3D, ProsaII, and RPF fold discrimination power (DP) scores, which assess the quality of the overall protein fold, are particularly useful in detecting a wrong fold generated at the initial stages of NMR structure determination. It is also possible to use PSVS analysis to select among alternate models generated during a NMR or X-ray structure refinement process, and to thus affect the trajectory of the structure refinement. Parameters assessing local structural features, such as stereochemistry, close contacts, or RPF precision scores, may not be as useful in the early stages of refinement, but are essential in evaluating models in the final stages of refinement. Site-specific MolProbity clash and RPF precision scores (see Fig. 6) can also be used to refine NOESY peak lists, and the corresponding assignments, allowing more accurate analysis of the protein structure.

Value of Multiple Structure Quality Assessment Measures

Overall, protein structure quality assessment is best made using multiple assessment metrics, including both knowledge-based methods and measures of the goodness-of-fit between the 3D protein structure and experimental data. The different tools used by PSVS to evaluate structure quality also exhibit different sensitivities for quality assessment when applied to structures with different degrees of precision and accuracy. For example, ProsaII, while useful for identifying grossly misfolded proteins,[7,36] cannot be used to differentiate between X-ray structures with the correct global fold but different levels of crystallographic resolution. In contrast, even highly refined NMR structures may have inaccurate side-chain conformations and exhibit low MolProbity clashscores and PROCHECK G-factor. The MolProbity clashscores and PROCHECK G-factors (for all dihedrals) roughly correlate with crystallographic resolution, suggesting they can distinguish highly accurate from less accurate X-ray or NMR structures. However, a completely incorrect representation of the protein structure as an energy-minimized extended chain, with poor ProsaII and Verify3D scores, would also exhibit good MolProbity clash and dihedral angle distribution scores. Good quality structures exhibit good structure quality scores across all of the assessment tools used by PSVS. Moreover, as demonstrated by the data in Figure 6, local assessments of structural irregularities are particularly valuable in identifying subtle structural problems, which may not be adequately reflected in global structure quality scores.

Good stereochemical and/or packing scores alone do not necessarily demonstrate that a protein structure is accurate; in addition the structure must be consistent with the experimental data. While crystallographic R-factors (particularly R-free) provide broadly accepted criteria for concordance with the diffraction data, the corresponding assessment measure(s) for NMR structures are not yet universally agreed upon. 'Traditional' NMR quality scores reported by PSVS, including summaries of distance constraint violations, constraints-per-residue, and convergence across the conformational ensemble (rmsd), are important measures of structure quality. However, such assessments provide necessary, but not sufficient, criteria for good quality structures.[12] In particular, assessments of constraint violations do not provide information on the accuracy of the structure, or how well it fits to all of the experimental data; e.g., some NOESY data may not be interpreted as specific distance constraints. Inaccurate structures may also exhibit good convergence (low rmsd) with a network of incorrectly interpreted, but largely satisfied, constraints. These issues are addressed to some degree by proposed "NMR R-factors"[13–15] comparing observed and back-calculated NOESY or RDC data, such as the RPF[15] analysis provided by PSVS.

Limitations of PSVS Analyses

PSVS runs multiple software tools for structure quality assessment, each of which has its own limitations in assessing structural accuracy. Moreover, PSVS uses a limited number of the available structure validation tools. While the set described here includes most of the commonly considered structure-quality-assessment parameters, there is a larger set of tools, including some dealing with hydrogen-bonding geometry[4,41] and comparisons with residual dipolar coupling NMR data,[14] that are not yet included in the PSVS automated analysis.

The knowledge-based potentials used by many of the PSVS structure evaluation tools also suffer from limitations arising from the particular data sets used in their training. Most of these scoring functions were derived from data on well-packed, soluble, globular X-ray crystal structures. New structural features of an accurate experimental structure will be flagged by the PSVS tools, particularly the site-specific analyses, as outliers. Such features require expert assessment of the reliability of the structure considering the available sample-specific experimental data. In addition, the Z scores of PSVS were calibrated on a set of relatively small (<500-residue) high-resolution crystal structures. Accordingly, these Z-scores may not be applicable for analysis of non-globular (e.g. coiled-coil) or membrane proteins.

Site-specific analyses provided by PSVS give tentative identifications of conformationally strained regions, which often result from errors in the structure. However, there are a number of biologically critical structural features that are indeed strained in native protein structures. Protein dynamics may also result in ensemble-averaged effects, which are interpreted as conformationally strained local protein structures.[12,15] For exam-

ple, RPF precision scores, identifying interproton distances that are close in the 3D structure but do not have corresponding data in the NOESY peak lists, may also reflect resonance broadening because of conformational exchange, rather than structural inaccuracy.[15] For these reasons, conformationally dynamic regions of the protein structure, which may be involved in ligand binding or transient oligomer interactions, may be inaccurately reported as conformationally strained. As such, information output by PSVS from its underlying assessment tools must be interpreted in the context of specific properties of the protein under study.

Finally, while PSVS provides Z scores for each structure quality assessment tool used, there is no clear threshold score to define 'good' and 'bad' structures. While further research is required to establish such thresholds, beginning in 2006 the NESG consortium has required that all NMR and X-ray crystal structures have Z scores > -5 for the four structure quality assessment tools used in this work, and (for NMR structures) RPF DP scores > 0.70.[15] Such thresholds should avoid the regrettable case of hDLC-2A, in which the incorrect monomer structure was initially deposited in the PDB.[39]

Comparison With Other Structure Quality Validation Interfaces

PSVS is not unique in so far as other software and servers also provide various levels of structure quality assessment. For example, the WHATCHECK module of the WHATIF[4] program provides analysis on a variety of structure quality parameters. These include analyses of bond length and bond angles, buried hydrogen bonds, atomic overlap, packing quality (using a threading potential), backbone dihedral angles, rotamer states, and a residue environment profile. In addition, it checks for atomic nomenclature, and performs other analyses more relevant to X-ray crystal structures. For some of the above parameters, WHATCHECK also reports Z-scores calculated from its database of high and medium resolution crystal structures.

The Biotech Validation Suite (http://biotech.ebi.ac.uk:8400) is a comprehensive interface that performs structure quality validation using a series of tools, including PROCHECK[2,3] to perform geometric analyses, including bond length and bond angles, Ramachandran plots, distorted geometry, and distances from planarity; WHATIF[4] to analyze bond length and bond angle, nomenclature, peptide flips, directional atomic contacts, torsion angles, and some other parameters; and PROVE[42] to analyze atomic volume deviations. In addition, the interface runs AQUA[3] to analyze completeness of NOE constraints, violations of constraints, and redundancy of intraresidual NOE constraints for NMR structures. It outputs the plots and tables generated by these tools.

VADAR is a web interface (http://redpoll.pharmacy.ualberta.edu/vadar) that calculates a large number of key structural quality parameters, compiling results from a number of structure quality evaluation programs. The programs incorporated into VADAR include those evaluating excluded volume,[43] accessible surface area,[44] stereochemical quality analysis,[45] secondary structure,[25,46,47] hydrogen bond geometry and energetics,[25,48] solvation free energy[49,50] as well as fold quality.[6] The program produces a set of graphs and tables with indices showing the quality of the protein structure.

The PDB[8] runs its own validation software as part of the ADIT tool used for structure deposition. The parameters evaluated include bond lengths and bond angles, torsion angles, chirality, atomic nomenclatures, missing atoms, and missing residues. These PDB assessment tools are also run by PSVS. PDB also provides PROCHECK analyses for all submitted protein structures. The BioMagResDataBase[51] runs AQUA, described above for the Biotech Validation Suite.

There are also a number of web servers that run structure quality validation analyses using specific tools. Verify3D[6] is accessible at http://www.doe-mbi.ucla.edu/Services/Verify_3D/, and generates a table and plot showing its output. The MolProbity[5] server builds H atoms, performs all-atom contact analysis and calculates Cβ deviations. It can be accessed at http://152.16.14.32. Both of these tools are also run by PSVS.

The PSVS server, presented here, attempts to perform key structure quality assessment analyses, with an emphasis toward those that are most useful in assessing structures generate by NMR or homology modeling methods. PSVS output includes data from the each of the underlying tools used, as well as summary tables and plots suitable for inclusion in a publication. The Concise Structure Quality Validation Report can be combined with text to directly produce an initial draft of a short publication. It will be expanded in the future to include other structure quality assessment reports.

Structure Quality Gap Between Crystal and NMR Structures

Many of the NMR structures reported between 2000 and 2005 by both SG and non-SG groups have PSVS Z scores similar to those observed for X-ray crystal structures refined to 2.5–3.5 Å resolution. However, in many other cases, the scores for NMR structures are poorer than even those observed for low-resolution X-ray structures. Only a few NMR structures have scores comparable to those observed for medium- and high-resolution (<2.5 Å) X-ray structures. We have made some initial efforts to understand the reasons for the improved stereochemical assessments of some NMR structures, and the apparent gap in structure quality between NMR and X-ray crystal structures.

Although in fact there is no *a priori* reason to believe that these quality scores should differ in sensitivity to accuracy when analyzing proteins in crystalline and solution environments, we cannot completely exclude that at least part of the reason for the large differences in Z score distributions between most NMR structures and high-resolution X-ray crystal structures may be *bonafide*

differences in conformational preferences in these environments. PSVS Z scores may also be compromised in regions of the NMR structure where the data is derived from a dynamic ensemble of rapidly interconverting structures; such dynamics may be more extensive for proteins in solution compared with those in crystalline environments. Accordingly, the lower scores observed for NMR structures simply reflect the dynamic nature of protein structures in solution.

Another potential cause for lower scores is that the NMR structures have inaccuracies resulting from incorrect interpretation of distance constraints. PSVS Z scores are particularly sensitive to structural distortions, which can result from incorrect interpretation of NMR data as distance constraints. In Figure 7(A), we compare different protocols for assigning NOESY peaks using the PROCHECK G-factor (all dihedrals) – MolProbity bivariant plot. In these data, there is no significant difference in global quality scores for structures determined with manually and automatically analyzed NOESY data, nor any significant difference when automated NOESY assignments were determined using the programs AutoStructure,[20] Candid,[19] or ARIA.[17,31] While detailed features of NMR structures do sometimes vary depending on the method of automated NOESY data analysis (unpublished results), these effects are smaller than other factors that determine the wide range of structure quality scores obtained for NMR structures.

Next, we examined the impact of structure generation protocols on structural quality assessment scores. As can also be seen in Figure 7(A), structures refined using different structure generation protocols and potential energy functions do indeed fall into clusters, with some overlap. The best scores are observed for NMR structures refined using the CNS force field.[30] Most of these highest-scoring NMR structures were refined with explicit solvent, using the protocol of Nederveen et al.[54] While this data set is too small to make statistically reliable conclusions, and no effort was made to further distinguish between the exact protocols that were used within the structure refinement programs, we observed that structures refined using CNS[30] have structure quality scores (particularly steric clashes and distributions of all dihedral angles) closer to those observed in high- and medium-resolution X-ray structures. Some structures refined using XPLOR (both the original[31] and the NIH version[53]) also have high scores for these two parameters.

Many of the NMR structures with poorer scores were refined using the protein structure analysis software DYANA/CYANA.[19,32] DYANA/CYANA uses a minimal force field, with bond length and bond angle parameters for amino acid residues that differ from those used in programs such as XPLOR, CNS, or PROCHECK. Over the last several years, there have been improvements in potential functions used by different versions of DYANA and CYANA. The PDB files available for the analyses presented in Figure 7(A) do not all provide DYANA/CYANA version numbers, and minimization protocols are not generally documented in the PDB header files. For these reasons, data for structures refined with various DYANA and CYANA versions are aggregated together in the plots of Figure 7(A).

Obviously, energy minimization can improve the structure quality scores based on atomic packing and dihedral angle distributions. However, if incorrect spatial constraints are incorporated, the best achievable scores (or lowest achievable energies) will be limited by the inaccurate constraints.[23] For this reason, it is useful to use the PSVS scores to track the refinement of a structure. For example, as part of the technology development efforts of the NESG Consortium, Liu et al.[52] have generated 8 protein NMR structures using automated NOESY analysis and structure generation with CYANA ver 2.0. These structures exhibited medium to low quality PROCHECK G-factor and MolProbity clashscore Z-scores [Fig 7(B)]. These eight structures were then further refined using constrained energy minimization with CNS, without or with explicit solvent. Figure 7(B) shows how this refinement process was tracked using PROCHECK G-factor—MolProbity clashscore bivariant plots. Simple energy minimization and molecular dynamics in the CNS refinement in vacuum protocol improves the values of these scores. Indeed, using an explicit solvent term in the CNS refinement protocol yields scores approaching those obtained for 1.8–2.5 Å X-ray crystal structures. More generally, site-specific structure–quality assessment scores [e.g. MolProbity clashscores and RPF precision scores, as shown in Fig. 6] can be used at this point to refine and correct the interpretation of NOESY data in terms of spatial constraints. In this way, it is possible to use PSVS, energy refinement, and reinterpretation of NOESY data to iteratively refine the experimental constraint lists and 3D protein structure.

CONCLUSIONS

The PSVS suite provides a standardized evaluation of protein structure quality, and can be used to compare the quality of X-ray and NMR structures. Z scores for each quality parameter allow comparisons of these several aspects of structure quality with the corresponding values in a set of high-resolution crystal structures. They provide both global and site specific measures of structural regularity relative to the knowledge-base of protein structures. The PSVS server can be accessed via an easy to use web-based graphical user interface.

Structural genomics projects make extensive use of newly developed automated methods of X-ray and NMR structure analysis to speed-up and standardize structure determination. Within the context of such structural genomics efforts, rapid and easy-to-use servers like PSVS are critical to ensure a consistent quality assessment, and to help identify structures, particularly those generated with NMR methods, that may require further structure refinement. PSVS analysis also provides useful information for improving electron density interpreta-

tions. The software suite is used to evaluate all protein structures determined by the NESG Consortium. These reports are posted in the publicly-accessible NESG Structure Gallery. Significantly, one key conclusion from this work is that, on average, the quality of small (< 500-residue) X-ray crystal and NMR structures generated by structural genomics consortia between 2000 and 2005 are similar to those determined with these methods by traditional structural genomics projects during this same period. However, using the set of structure–quality assessment parameters evaluated here, there is an apparent quality gap between most NMR structures and X-ray crystal structures determined at high- and medium-resolution. This quality gap can be reduced using conformational energy minimization protocols together with site-specific structure quality assessments to refine the NMR constraint lists. PSVS thus provides important tools for improving and ensuring the quality of protein X-ray crystal and NMR structures generated in both structural genomics and traditional structural biology projects.

ACKNOWLEDGMENTS

We thank J. Aramini, J. Cort, J. Locke, D. Monleon, D. Snyder, T. Szyperski, T. Ramelot, Y. J. Huang, and D. Hang for helpful discussions. We also thank J. Richardson and D. Richardson for their expert advice and suggestions.

REFERENCES

1. Montelione GT, Anderson S. Structural genomics: keystone for a human proteome project. Nat Struct Biol 1999;6:11–12.
2. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
3. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996;8:477–486.
4. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8:29,52–56.
5. Lovell SC, Davis IW, Arendall WB, III, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: ϕ, φ and Cβ deviation. Proteins 2003; 50:437–450.
6. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–85.
7. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
9. Stout GH, Jensen LH. X-ray structure determination: a practical guide. New York: Wiley; 1989. pp 343–378.
10. Brünger A. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 1992;355:472–475.
11. Brünger A. Free R value: cross-validation in crystallography. In: Carter CW, Sweet RM, editors. Macromolecular crystallography. San Diego, CA: Academic Press; 1997.
12. Snyder DA, Bhattacharya A, Huang YJ, Montelione GT. Assessing precision and accuracy of protein structures derived from NMR data. Proteins 2005;59:655–661.
13. Gronwald W, Kirchhofer R, Gorler A, Kremer W, Ganslmeier B, Neidig KP, Kalbitzer HR. RFAC, a program for automated NMR R-factor estimation. J Biomol NMR 2000;17:137–151.
14. Clore GM, Garrett DS. R factor, free R, and complete cross validation for dipolar coupling refinement of NMR structures. J Am Chem Soc 1999;121:9008–9012.
15. Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 2005;127:1665–1674.
16. Doreleijers JF, Rullmann JA, Kaptein R. Quality assessment of NMR structures: a statistical survey. J Mol Biol 1998;281:149–164.
17. Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H. Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. J Mol Biol 1997;269:408–422.
18. Nilges M. Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. J Mol Biol 1995;245: 645–660.
19. Herrmann T, Guntert P, Wuthrich K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 2002;319:209–227.
20. Huang YJ, Tejero R, Powers R, Montelione GT. A topology-constrained distance network algorithm for protein structure determination from NMR data. Proteins: 2006;62:587–603.
21. Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM. Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear Overhauser effect spectra and chemical shift assignments. J Am Chem Soc 2004;126:6258–6273.
22. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.
23. Sahasrabudhe PV, Tejero R, Kitao S, Furuichi Y, Montelione GT. Homology modeling of an RNP domain from a human RNA-binding protein: homology-constrained energy optimization provides a criterion for distinguishing potential sequence alignments. Proteins 1998;33:558–566.
24. Arendall WB, III, Tempel W, Richardson JS, Zhou W, Wang S, Davis IW, Liu ZJ, Rose JP, Carson WM, Luo M, Richardson DC, Wang BC. A test of enhancing model accuracy in high-throughput crystallography. J Struct Funct Genomics 2005;6: 1–11.
25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
26. Word JM, Bateman RC, Jr, Presley BK, Lovell SC, Richardson DC. Exploring steric constraints on protein mutations using MAGE/PROBE. Protein Sci 2000;9:2251–2259.
27. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 1999;285:1711–1733.
28. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.
29. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.
30. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54: 905–921.
31. Brünger AT. X-PLOR, Version 3.1: a system for X-ray crystallography and NMR. New Haven: Yale University Press; 1992. xvii, 382 p.
32. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 1997;273:283–298.
33. Hyberts SG, Goldberg MS, Havel TF, Wagner G. The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. Protein Sci 1992;1:736–751.

34. Richardson DC, Richardson JS. The kinemage: a tool for scientific communication. Protein Sci 1992;1:3–9.
35. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
36. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990; 213:859–883.
37. Chen L, Oughtred R, Berman HM, Westbrook J. TargetDB: a target registration database for structural genomics projects. Bioinformatics 2004;20:2860–2862.
38. Sokal RR, Rohlf FJ. Biometry: the principles and practice of statistics in biological research. New York: W.H. Freeman; 1995. pp 429–434, 708–715.
39. Nabuurs SB, Spronk CA, Vuister GW, Vriend G. Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. PLoS Comput Biol 2006;2:e9.
40. Ilangovan U, Ding W, Zhong Y, Wilson CL, Groppe JC, Trbovich JT, Zuniga J, Demeler B, Tang Q, Gao G, Mulder KM, Hinck AP. Structure and dynamics of the homodimeric dynein light chain km23. J Mol Biol 2005;352:338–354.
41. Grishaev A, Bax A. An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. J Am Chem Soc 2004;126:7281–7292.
42. Pontius J, Richelle J, Wodak SJ. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 1996;264:121–136.
43. Richards F. Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng 1977;6:151–176.
44. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
45. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. Proteins 1992;12:345–364.
46. Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 1988;3:71–84.
47. Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. J Mol Biol 1977;114:181–239.
48. Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 1984;44:97–179.
49. Chiche L, Gregoret LM, Cohen FE, Kollman PA. Protein model structure evaluation using the solvation free energy of folding. Proc Natl Acad Sci USA 1990;87:3240–3243.
50. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
51. Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Markley JL, Ulrich EL. 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 2003;26:139–146.
52. 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. NMR data collection and analysis protocol for high-throughput protein structure determination. Proc Natl Acad Sci USA 2005;102:10487–10492.
53. Schwieters CD, Kuszewski JJ, Tjandra N, Marius Clore G. The Xplor-NIH NMR molecular structure determination package. J Magn Reson 2003;160:65–73.
54. Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CA, Nabuurs SB, Guntert P, Livny M, Markley JL, Nilges M, Ulrich EL, Kaptein R, Bonvin AM. RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins 2005;59:662–672.