REVIEW

A community resource of experimental data for NMR / X-ray crystal structure pairs

John K. Everett,[1] Roberto Tejero,[2] Sarath B. K. Murthy,[1] Thomas B. Acton,[1] James M. Aramini,[1] Michael C. Baran,[1] Jordi Benach,[3] John R. Cort,[4] Alexander Eletsky,[5] Farhad Forouhar,[3] Rongjin Guan,[1] Alexandre P. Kuzin,[3] Hsiau-Wei Lee,[6] Gaohua Liu,[1] Rajeswari Mani,[1] Binchen Mao,[1] Jeffrey L. Mills,[5] Alexander F. Montelione,[1] Kari Pederson,[6] Robert Powers,[7] Theresa Ramelot,[8] Paolo Rossi,[1] Jayaraman Seetharaman,[3] David Snyder,[9] G. V. T. Swapna,[1] Sergey M. Vorobiev,[3] Yibing Wu,[5] Rong Xiao,[1] Yunhuang Yang,[8] Cheryl H. Arrowsmith,[10] John F. Hunt,[3] Michael A. Kennedy,[8] James H. Prestegard,[6] Thomas Szyperski,[5] Liang Tong,[3] and Gaetano T. Montelione[1,11]*

[1]Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854, USA

[2]Departamento De Química Física, Universidad De Valencia, Valencia, Spain

[3]Department of Biological Sciences and Northeast Structural Genomics Consortium, Columbia University, New York, NY 10027, USA

[4]Fundamental and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99354, USA

[5]Department of Chemistry, The State University of New York at Buffalo, and Northeast Structural Genomics Consortium, Buffalo, New York 14260, USA

[6]Complex Carbohydrate Research Center and Northeast Structural Genomics Consortium, University of Georgia, Athens, Georgia 30602, USA

[7]Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588, USA

[8]Department of Chemistry and Biochemistry, Northeast Structural Genomics Consortium, Miami University, Oxford, Ohio 45056, USA

[9]Department of Chemistry, College of Science and Health, William Paterson University of NJ, Wayne, New Jersey 07470, USA

[10]Cancer Genomics & Proteomics, Department of Medical Biophysics, Ontario Cancer Institute, and Northeast Structural Genomics Consortium, University of Toronto, Toronto, Ontario M5G 1L7, Canada

[11]Department of Biochemistry, Robert Wood Johnson Medical School, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854, USA

Received 22 June 2015; Accepted 17 August 2015
DOI: 10.1002/pro.2774
Published online 21 August 2015 proteinscience.org

Abstract: We have developed an online NMR / X-ray Structure Pair Data Repository. The NIGMS Protein Structure Initiative (PSI) has provided many valuable reagents, 3D structures, and technolo-

Abbreviations: 3D, three-dimensional; BioMagResDB, Biological Magnetic Resonance Data Bank; HSQC, heteronuclear single quantum coherence NMR spectroscopy; PDB, Protein Data Bank; RMSD, root-mean-squared distance between superimposed atomic coordinates; RMS_{ens}, the backbone (N, Cα, C′) RMSD of each conformer (for well-defined regions) to the representative (medoid) conformer of the ensemble of NMR conformers reported as the "NMR structure"; RMS_{Xtal}, the backbone (N, Cα, C′) RMSD between the representative (medoid) NMR conformer and the X-ray crystal structure, using the well-defined regions of the NMR structure and the corresponding reported coordinates of the X-ray crystal structure; Γ, RMS_{Xtal}/RMS_{ens}.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Institutes of Health Protein Structure Initiative; Grant number: U54-GM094597; Grant sponsor: Jerome and Lorraine Aresty Charitable Foundation.

*Correspondence to: Gaetano T. Montelione, CABM-Rutgers University, 679 Hoes Lane, Piscataway, NJ, USA. E-mail: gtm@rutgers.edu

gies for structural biology. The Northeast Structural Genomics Consortium was one of several PSI centers. NESG used both X-ray crystallography and NMR spectroscopy for protein structure determination. A key goal of the PSI was to provide experimental structures for at least one representative of each of hundreds of targeted protein domain families. In some cases, structures for identical (or nearly identical) constructs were determined by both NMR and X-ray crystallography. NMR spectroscopy and X-ray diffraction data for 41 of these "NMR / X-ray" structure pairs determined using conventional triple-resonance NMR methods with extensive sidechain resonance assignments have been organized in an online NMR / X-ray Structure Pair Data Repository. In addition, several NMR data sets for perdeuterated, methyl-protonated protein samples are included in this repository. As an example of the utility of this repository, these data were used to revisit questions about the precision and accuracy of protein NMR structures first outlined by Levy and coworkers several years ago (Andrec *et al*., Proteins 2007;69:449–465). These results demonstrate that the agreement between NMR and X-ray crystal structures is improved using modern methods of protein NMR spectroscopy. The NMR / X-ray Structure Pair Data Repository will provide a valuable resource for new computational NMR methods development.

Keywords: protein NMR spectroscopy; X-ray crystallography; structural bioinformatics; accuracy and precision of NMR structures

Introduction

The National Institutes of General Medical Sciences (NIGMS) Protein Structure Initiative was established in 2000 as advances in genomic sequencing, bioinformatics, and methods for rapid determination of protein 3D structures by X-ray crystallography and NMR converged to suggest the potential for "genomic-scale" protein structure determination.[1] The long-term goal of the Protein Structure Initiative (PSI) was to provide 3D structural information for most proteins in nature. The vision of the PSI was to make 3D protein structure information an integral part of biology research.

Over the past 15 years, the PSI program has provided more than 6800 new protein structures into the public domain (http://sbkb.org/). The primary mission of the PSI was to complement traditional structural biology research by determining 3D structures of proteins (or protein domains) selected primarily to provide extensive coverage of the largest protein domain families. In the final phase of PSI, PSI-Biology, these structure determination efforts were coupled with specific biomedical driving projects. These PSI protein structures are being used as templates for modeling tens of thousands of homologous proteins,[2] and provide a database of protein structures and biophysical properties (e.g., chemical shifts) that allow researchers to more accurately predict and design protein structures. The PSI program also developed extensive databases of protein sample production information, improved protocols for using existing technologies, and new technologies that are just beginning to have their most significant impact in structural biology research.[3–5]

The Northeast Structural Genomics Consortium (NESG) was one of four Large-Scale Centers for structure production. NESG scientists used broad biological, genomic, and bioinformatics criteria, together with protein targets selected from specific biological projects, to (i) provide significant structural coverage of a large number of protein sequences in nature, (ii) develop and disseminate novel and/or improved technologies for structural biology and bioinformatics, and (iii) make these structures, structure production data, and the associated reagents and technologies publicly available to the worldwide scientific community. A hallmark of the NESG was the combined use of both protein crystallography and NMR spectroscopy in high throughput protein structure determination. More than 1200 NMR and X-ray crystal structures were deposited in the PDB by NESG scientists.

The primary mission of the NESG consortium in its first 10 years was structural coverage of large protein domain families consisting of many sequences with unknown 3D structures.[1–3,6] The aim was to determine a representative structure for each of these domain families. Representative proteins were selected based on their "modeling leverage" which was a measure of how many additional sequences could be accurately modeled using the structure of the representative protein. In the interest of broad structural coverage, a process evolved in which having completed a 3D structure for one representative from a protein domain family, the family was scored as "covered", and additional work on the same domain family was deprioritized. In particular, if the domain family was "covered" by an X-ray crystal structure, work on the corresponding NMR structure was suspended, and vice versa. However, in a limited number of cases both NMR and X-ray crystal structures were produced for the same (or similar) protein construct. These pairs of NMR and X-ray crystal structures for identical (or nearly identical) sequences are referred to here as "NMR / X-ray structure pairs." These structure pairs were in fact a byproduct of the primary goal of the PSI program. However, together with the corresponding protein sample production and raw experimental data (structure factors, crystallization conditions, NMR resonance assignments, NOESY

Table I. *Data Sets for 41 NESG NMR / X-Ray Pairs Determined by Conventional Triple-Resonance NMR Methods on Fully Protonated Samples*

Uniprot id	Method	NESG target id	PDB id	BMRB id	No. of residues[a]	Solution oligomer state	NOESY peak lists	^{15}N–^{1}H RDC data
Q7VV99	NMR	BeR31	2K2E	15702	150	Monomer	No	Yes
	XRAY	BeR31	3CPK	n.a.[b]	150		n.a.	n.a.
Q9AAR9	NMR	CcR55	2JQN	15281	114	Monomer	Yes	No
	XRAY	CcR55	2O0Q	n.a.	114		n.a.	n.a.
Q481E4	NMR	CsR4	2JR2	15317	67	Dimer	Yes	Yes
	XRAY	CsR4	2OTA	n.a.	67		n.a.	n.a.
Q8KFZ1	NMR	CtR107	2KCU	16097	158	Monomer	Yes	Yes
	XRAY	CtR107	3E0H	n.a.	158		n.a.	n.a.
Q8KC80	NMR	CtR148A	2KO1	16486	79	Dimer	Yes	Yes
	XRAY	CtR148A	3IBW	n.a.	79		n.a.	n.a.
Q24NW5	NMR	DhR29B	2KPU	16570	89	Monomer	Yes	Yes
	XRAY	DhR29B	3LYW	n.a.	89		n.a.	n.a.
Q251Q8	NMR	DhR8C	2KYI	16961	63	Dimer	Yes	Yes
	XRAY	DhR8C	3IPF	n.a.	63		n.a.	n.a.
Q9RZE3	NMR	DrR147D	2KCZ	16100	146	Monomer	Yes	No
	XRAY	DrR147D	3GGN	n.a.	146		n.a.	n.a.
P65294	NMR	ER382A	2JN0	15079	52	Monomer	No	No
	XRAY	ER382A	3FIF	n.a.	52		n.a.	n.a.
Q39VC5	NMR	GmR137	2K5P	15844	70	Monomer	No	No
	XRAY	GmR137	3CWI	n.a.	70		n.a.	n.a.
Q9Y547	NMR	HR1958	1XPW	6344	144	Monomer	Yes	No
	XRAY	HR1958	1TVG	n.a.	144		n.a.	n.a.
P62195	NMR	HR3102A	2KRK	16640	76	Monomer	Yes	No
	XRAY	HR3102A	3KW6	n.a.	76		n.a.	n.a.
Q15811	NMR	HR3646E	2KHN	16250	111	Monomer	Yes	Yes
	XRAY	HR3646E	3FIA	n.a.	111		n.a.	n.a.
Q9Y3C8	NMR	HR41	2K07	6546	167	Monomer	Yes	No
	XRAY	HR41	3EVX	n.a.	167		n.a.	n.a.
Q01826	NMR	HR4435B	2L1P	17092	72	Monomer	Yes	Yes
	XRAY	HR4435B	3NZL	n.a.	72		n.a.	n.a.
Q12906	NMR	HR4527E	2L33	17169	80	Monomer	Yes	Yes
	XRAY	HR4527E	3P1X	n.a.	74		n.a.	n.a.
P15056	NMR	HR4694F	2L05	17030	84	Monomer	Yes	Yes
	XRAY	HR4694F	3NY5	n.a.	85		n.a.	n.a.
P20700	NMR	HR5546A	2KPW	16572	111	Monomer	Yes	Yes
	XRAY	HR5546A	3JT0	n.a.	111		n.a.	n.a.
Q5FJ43	NMR	LaR80A	2LFI	17754	113	Monomer	Yes	Yes
	XRAY	LaR80A	3Q69	n.a.	113		n.a.	n.a.
E3YVT8	NMR	LkR112	2KPP	16563	105	Monomer	Yes	Yes
	XRAY	LkR112	3LD7	n.a.	92		n.a.	n.a.
Q7U294	NMR	MbR242E	2KKO	16368	101	Dimer	Yes	Yes
	XRAY	MbR242E	3GW2	n.a.	101		n.a.	n.a.
Q8KNE9	NMR	MiR12	2LUZ	18547	182	Dimer	Yes	Yes
	XRAY	MiR12	4FPW	n.a.	182		n.a.	n.a.
Q6LYF9	NMR	MrR110B	2K5V	15849	95	Monomer	Yes	No
	XRAY	MrR110B	3E0E	n.a.	95		n.a.	n.a.
P03495	NMR	OR8C	2KKZ	16376	131	Monomer	Yes	No
	XRAY	OR8C	2RHK	n.a.	131		n.a.	n.a.
Q8U1U6	NMR	PfR193A	2KL6	16385	105	Monomer	Yes	Yes
	XRAY	PfR193A	3IDU	n.a.	118		n.a.	n.a.
Q880Y4	NMR	PsR293	2KFP	16186	117	Monomer	Yes	No
	XRAY	PsR293	3H9X	n.a.	117		n.a.	n.a.
Q6N882	NMR	RpR324	2KW2	16805	93	Monomer	Yes	Yes
	XRAY	RpR324	3LMO	n.a.	93		n.a.	n.a.
Q55544	NMR	SgR209C	2L06	17031	148	Dimer	Yes	Yes
	XRAY	SgR209C	3OSJ	n.a.	148		n.a.	n.a.
P74795	NMR	SgR42	2JZ2	15604	58	Monomer	Yes	No
	XRAY	SgR42	3C4S	n.a.	58		n.a.	n.a.
Q8EF26	NMR	SoR77	2JUW	15456	72	Monomer	Yes	No
	XRAY	SoR77	2QTI	n.a.	72		n.a.	n.a.
Q97RM2	NMR	SpR104	2L3A	17175	73	Dimer	Yes	No
	XRAY	SpR104	3OBH	n.a.	73		n.a.	n.a.
P50833	NMR	SR213	2HFI	16113	123	Monomer	Yes	Yes

Table I. *Continued*

Uniprot id	Method	NESG target id	PDB id	BMRB id	No. of residues[a]	Solution oligomer state	NOESY peak lists	^{15}N–^{1}H RDC data
	XRAY	SR213	2IM8	n.a.	123		n.a.	n.a.
O31818	NMR	SR384	2JVD	15476	46	Monomer	Yes	Yes
	XRAY	SR384	3BHP	n.a.	52		n.a.	n.a.
P71066	NMR	SR478	2JS1	15350	72	Dimer	Yes	Yes
	XRAY	SR478	2GSV	n.a.	72		n.a.	n.a.
Q2S6C5	NMR	SrR115C	2KCV	16084	91	Monomer	Yes	Yes
	XRAY	SrR115C	3MA5	n.a.	91		n.a.	n.a.
P95883	NMR	SsR10	2JPU	15265	121	Monomer	Yes	No
	XRAY	SsR10	2Q00	n.a.	121		n.a.	n.a.
Q8ZRJ2	NMR	StR65	2JN8	15089	107	Monomer	Yes	No
	XRAY	StR65	2ES9	n.a.	107		n.a.	n.a.
E7UZA7	NMR	StR70	2JZT	7178	134	Monomer	No	No
	XRAY	StR70	2ES7	n.a.	134		n.a.	n.a.
B2D8H3	NMR	UuR17A	2KRT	16648	112	Monomer	Yes	No
	XRAY	UuR17A	3K63	n.a.	117		n.a.	n.a.
Q8P6W3	NMR	XcR50	1XPV	6363	78	Monomer	Yes	No
	XRAY	XcR50	1TTZ	n.a.	78		n.a.	n.a.
Q99U58	NMR	ZR18	1PQX	5844	83	Monomer	No	No
	XRAY	ZR18	2FFM	n.a.	83		n.a.	n.a.

[a] The reported construct length excludes small (~8 residue) purification tags present in some constructs, unless these purification tags provided well-defined atomic coordinates.
[b] n.a. - not applicable

spectra and peak lists, residual dipolar coupling (RDC) data, etc.), the NMR / X-ray pairs are particularly useful for testing new methods for protein NMR structure determination and structure validation.[7–9] These data sets are a unique and valuable resource for the broader scientific community.

As an example of the unique value of NMR / X-ray pairs in methods development for structural bioinformatics, in 2007 Levy and coworkers[10] explored fundamental questions of the precision and accuracy of NMR structure ensembles[10,11] using 148 NMR / X-ray pairs culled from the Protein Data Bank. This study reported that for every one of these 148 protein structure pairs, the backbone root-mean-square distance (RMSD) over core atoms of the crystal structure to the mean NMR structure is larger than the average RMSD of the members of the NMR ensemble to the mean NMR coordinates. Three-quarters of these structure pairs were reported to have backbone RMSDs between the X-ray crystal structure and mean NMR structure of more than twice the average RMSD within the NMR ensemble.[10] The authors concluded that this difference is real, but could not determine the underlying biophysical or statistical basis for this difference in precision (the RMSD of atomic coordinates within the NMR ensemble) and accuracy (the RMSD between the mean NMR structure coordinates and the X-ray crystal structure). This landmark article presents an open question to the structural biology community, which nearly a decade later still has not been adequately addressed.

In this article, we present an organized data repository containing both raw and processed data for 41 NESG NMR / X-ray pairs for which the NMR structure was determined using fully protonated samples. These 3D structures are for pairs of identical (or very similar) protein constructs. The repository (http://spine.nesg.org/nmrdata) includes raw and processed NMR data, NMR resonance assignments, and X-ray crystallography structure factor files. For many of these proteins, we also provide NOESY time domain data, NOESY peak lists, and RDC data mapped to the corresponding resonance assignments. We also provide data sets for seven additional perdeuterated, methyl protonated protein samples,[9] three of which have X-ray crystal structures available.

Results

NMR / X-ray structure pair data repository

Forty-one (41) protein structures have been determined in the NESG project by both NMR and X-ray crystallography using conventional triple resonance NMR methods on fully protonated protein samples (Table I). NMR data and structure factors for these 41 structures are collected together in a single data repository, the NESG NMR / X-ray Structure Pair Data Repository (http://spine.nesg.org/nmrdata). These structures have also all been deposited in the Protein Data Bank (Supporting Information Table S1). These proteins and protein domains range in size from 46 to 182 residues. The NMR structures include both monomers and homodimers. Some of these proteins form higher-order oligomers in the crystal structure. A few of these structures have been published as independent papers,[9,12] while others are currently being used for follow-up structure–function studies.

Table II. *Data Sets for $^{2}H,^{13}C,^{15}N$-ILV(CH_3) Protein Samples*

UniProt id	Method	NESG target id	PDB id	BMRB id	No. of residues[e]	Solution oligomer state	NOESY peak lists	^{15}N–^{1}H RDC data
P74712	NMR	SgR145	2KW5[a]	16806	194	Monomer	Yes	Yes
	XRAY	SgR145	3MER[a]	n.a.	194	Monomer	n.a.	n.a.
Q92786	NMR	HR4460B	2LMD[a]	18112	163	Monomer	Yes	Yes
P54155	NMR	SR10	2KZN[a]	17008	143	Monomer	Yes	Yes
	XRAY	SR10	3E0O[b]	n.a.	144	Monomer	n.a	n.a
Q93573	NMR	WR73	2LOY[a]	16833	181	Monomer	Yes	Yes
Q5V502	NMR	HmR11	2LNU[a]	18180	182	Monomer	Yes	Yes
Q9HRE7	NMR	HsR50	2LOK[a]	18215	189	Monomer	Yes	Yes
P0AEX9	NMR	ER690	2MV0[a]	25237	370	Monomer[c]	Yes	Yes
	XRAY	ER690	1DMB[d]	n.a.	370	Monomer[c]	n.a.	n.a

[a] Lange *et al*.[9]
[b] Not determined by NESG consortium.[23]
[c] Complex bound to β-cyclodextrin.
[d] Not determined by NESG consortium.[24]
[e] The reported construct length excludes small (8 residue) purification tags present in some constructs, unless these purification tags provided well-defined atomic coordinates

The X-ray crystal structures are solved to resolutions of 1.20–2.80 Å. Both structure factor data and successful crystallization conditions (Supporting Information Table S2) are available for all of the X-ray crystal structures, and extensive chemical shift data are available for all of the NMR structures. NMR data available for the fully-protonated protein targets include NMR restraints (41 targets), NOESY peak lists (36 targets), 3D NOESY time domain data (29 targets), and backbone ^{15}N–^{1}H residual dipolar coupling data mapped to the corresponding resonance assignments (23 targets, with 17 RDC data sets recorded using at least 2 alignment media) (Table I). These data can be downloaded as a complete set from the NMR / X-ray Structure Pair Repository (http://spine.nesg.org/nmrdata/).

Determining larger protein structures (20–70 kDa) by NMR is challenging but highly feasible.[13–15] For such larger proteins, deuteration becomes necessary to circumvent the efficient spin relaxation properties resulting from their slow rotational correlation times.[13,16,17] Backbone and sidechain amide hydrogens (H^N) can be exchanged back into the protein sample, providing backbone and sidechain H^N–H^N NOE data. However, removing protons also eliminates long-range NOESY information from sidechains except for selectively protonated side-chain moieties. The difficulty in determining accurate structures with no or limited side-chain information (i.e., sparse NMR data) is a major bottleneck that currently prevents routine application of NMR to larger systems. Several NMR data sets have been generated in NESG projects on perdeuterated proteins[9,18–22] which are labeled with ^{13}C–^{1}H methyl groups of Ile(δ1), Leu, Val, and/or Ala residues. These data, along with several other "sparse NMR data" sets generated for proteins ranging in size from 143 to 370 residues (Table II), are also included in the NMR / X-ray Structure Pair Data Repository.

Many of these same NMR and X-ray crystallography data sets have already been used for new NMR technology development. Some of these structures have been used to assess the value of NMR structures for phasing the corresponding X-ray diffraction data using the method of molecular replacement.[7] A large number of these structures have also been used to explore refinement of NMR structures using NMR-data-restrained Rosetta calculations,[8,25] and a few have been used to explore *de novo* structure generation using restrained CS-Rosetta.[9,25] Coordinates and input restraint files for 39 restrained-Rosetta refined NMR structures generated by Mao *et al*.,[8] are also available in the NMR / X-ray Structure Pair Data Repository.

Precision and accuracy of NMR structures

Using the NMR / X-ray structure pairs deposited by NESG in the PDB, we re-examined questions raised by Snyder *et al*.[11] and Andrec *et al*.[10] on the relationships between the precision and accuracy of NMR structures. NMR structures are typically presented as an ensemble of 10–20 conformers, each of which is an approximately equally good fit to the experimental NMR restraints. As a convention, the wwPDB NMR-VTF has recommended that the "representative NMR structure" is defined as the single conformer in the ensemble that is most similar to all the others[26]; that is, the medoid of the conformer distribution. The NMR-VTF has also recommended that ill-defined regions (i.e., regions of the polypeptide structure that are not converged in the NMR ensemble) are excluded in computing RMSDs of atomic coordinates used to define the medoid conformer. These calculations were done using the PDBStat software package,[25] where well-defined vs. ill-defined regions are determined using a variance matrix analysis provide by the FindCore2[27] algorithms of PDBStat.

In this study, the convergence within the ensemble was characterized by the average (and standard

Table III. *Comparisons of RMS_{ens} and RMS_{Xtal} for Different Sets of NMR / X-Ray Pairs*

Data set	Number of pairs used in analysis	RMS_{ens}			RMS_{Xtal}		
		Mean (Å)	S.D. (Å)	Median (Å)	Mean (Å)	S.D. (Å)	Median (Å)
PDB	145	0.76	0.25	0.76	1.60	0.68	1.49
NESG	41	0.95	0.37	0.87	1.41	0.55	1.29
NESG/R3	39	0.83	0.40	0.76	1.19	0.41	1.11

Structures with $RMS_{Xtal} > 3.5$ Å were excluded from analysis.

deviation) RMSD between each member of the NMR ensemble and the representative structure (i.e., the medoid conformer). This metric, which we call RMS_{ens}, is a proxy for the precision of the NMR structures. The "accuracy" of the structure was assessed as the RMS_{Xtal}, the backbone RMSD between the medoid NMR conformer and the X-ray crystal structure, again excluding regions of the NMR structure that are ill-defined in the NMR ensemble, as well as atoms of the X-ray crystal structure for which atomic positions are not defined. Using RMS_{Xtal} as a measure of structural accuracy assumes that the "true" structure is the corresponding X-ray crystal structure, which may not always be a correct assumption (see Discussion section).

In the following statistical analyses, three sets of NMR / X-ray pairs were considered. The "NESG" set (Table I) includes 41 NESG/NMR pairs (Table I of NMR / X-ray Structure Pair Data Repository http://spine.nesg.org/nmrdata). The "NESG/R3" set is a subset of the NESG structure pairs which have been energy-refined using restrained Rosetta refinement with version Rosetta.v3, as described by Mao *et al.*[8] Of the 40 Rosetta-refined structures described in the original publication (Table III of NMR / X-ray Structure Pair Data Repository http://spine.nesg.org/nmrdata/), NESG target DrR147D was excluded from NMR / X-ray structure comparisons because its solution NMR structure is a monomer solved at pH 4.5, while its X-ray structure is a dimer solved at pH 6.0, and NMR studies demonstrate that there is a significant structural change over this pH range.[8] The "PDB" set of NMR / X-ray pairs includes 145 of the 148 pairs used by Andrec *et al.*,[10] chosen so as to provide comparison with these benchmark results. Two of the original pairs (NESG targets HR1958 and ZR18) were excluded from the PDB set because they are in the NESG set, and one pair was excluded because of inconsistencies in atom naming conventions. The resulting PDB set does not include any NESG pairs.

The RMS_{ens} and RMS_{Xtal} values for each pair from the various groups of NMR / X-ray pairs (Supporting Information Table S3) are shown as step plots (0.2 Å bins) in Figure 1. Mean, standard deviations, and median values are summarized in Table III. For most of the structures in all three data sets, the structural variability within the NMR ensemble is smaller than the difference between the NMR and X-ray crystal structures. This RMSD difference is smaller for the NESG NMR / X-ray pairs (median $RMS_{ens} = 0.87$ Å vs. median $RMS_{Xtal} = 1.29$ Å) than for the larger set of non-NESG structure pairs selected from the PDB (median $RMS_{ens} = 0.76$ Å vs. median $RMS_{Xtal} = 1.49$ Å). The difference between median RMS_{ens} and RMS_{Xtal} is smaller still for the restrained Rosetta refined NESG/R3 structures (median $RMS_{ens} = 0.76$ Å vs median $RMS_{Xtal} = 1.11$

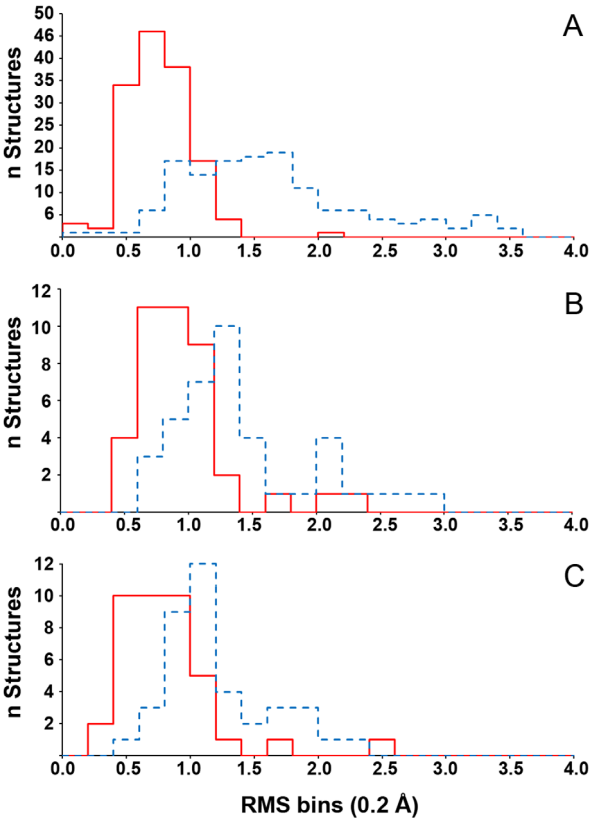

Figure 1. Backbone RMSDs to the medoid conformer within NMR ensembles are generally smaller than RMSDs between the medoid conformer and the corresponding X-ray crystal structure. Histogram plots of backbone RMSDs within each NMR ensemble (RMS_{ens}—red solid lines), and between the medoid NMR conformer and the X-ray crystal structure (RMS_{Xtal}—blue dashed lines), showing the numbers of structure (*n*) in each 0.2 Å bin. Data are presented for NMR / X-ray pairs of A: PDB set of 145 pairs, B: NESG set of 41 pairs, and C: NESG/R3 set of 39 pairs, a subset of the NESG set that have been energy minimized using the restrained Rosetta protocol.[8] These sets of NMR / X-ray pairs are defined in the text.

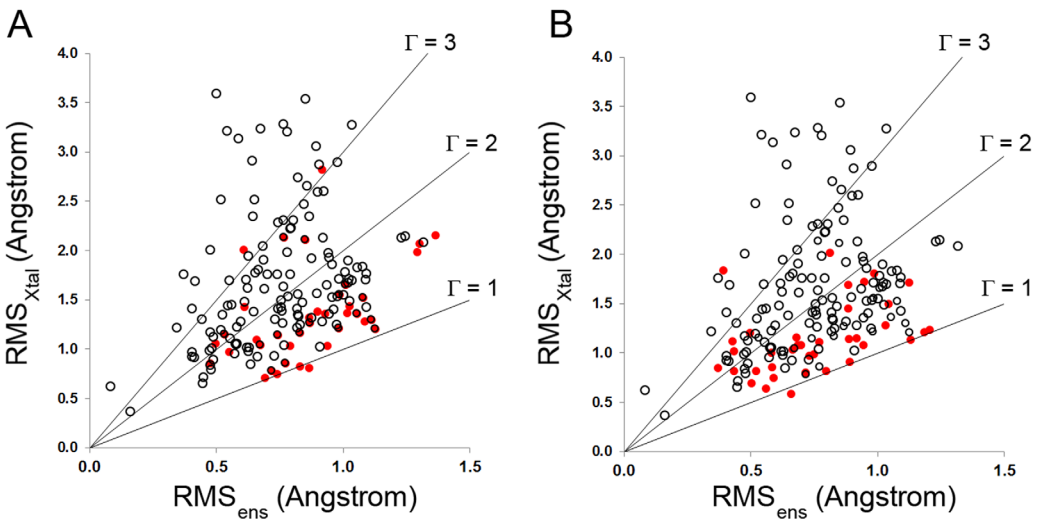

Figure 2. NESG NMR structures are more like corresponding X-ray crystal structures than the PDB NMR structures. Scatter plots between RMS_{ens}, providing an estimate of the convergence of the NMR ensemble, and RMS_{Xtal}, comparing the medoid NMR conformer with the corresponding X-ray crystal structure. The three solid lines in each plot indicate the points for which RMS_{Xtal}/RMS_{ens} ratios are 1:1, 2:1, and 3:1, respectively. A: Comparison of PDB set (open black circles) and NESG set (closed red circles) NMR / X-ray structure pairs. B: Comparison of PDB set (open black circles) and NESG/R3 set (closed red circles) of NMR / X-ray structure pairs.

Å). This trend is also obvious from the mean values of RMS_{Xtal}: PDB set (mean $RMS_{Xtal} = 1.60 \pm 0.68$ Å), NESG set (mean = 1.41 ± 0.55 Å), and the NESG/R3 set (mean $RMS_{Xtal} = 1.19 \pm 0.41$ Å). The methods used by the NESG consortium generated NMR structures that are more like the corresponding X-ray crystal structures than those used to generate the PDB pairs analyzed by Andrec *et al*. However, for all three distributions of Figure 1, Kolmogorov–Smirnov statistical tests[28] show that the RMS_{Xtal} distributions are distinct (i.e., significantly higher) from the corresponding RMS_{ens} distributions ($P < 0.001$).

Andrec *et al*. observed that for every one of the 148 NMR / X-ray pairs selected in their study, $RMS_{ens} < RMS_{Xtal}$, with 76% of the structure pairs having an RMSD of the crystal structure to the mean NMR structure more than a factor of two larger than the average RMSD of the NMR ensemble.[10] Using the essentially same set of 145 PDB NMR / X-ray pairs, with the methods of superimposition and RMSD analysis described in the Methods and Materials section, we see this same trend (Fig. 2). For convenience, we define for each NMR / X-ray pair the parameter $\Gamma = RMS_{Xtal}/RMS_{ens}$. While most PDB NMR / X-ray pairs have $\Gamma > 2$, most NESG and NESG/R3 pairs have $\Gamma < 2$ (Fig. 2). For the PDB, NESG, and NESG/R3 pairs, the mean values of Γ are 2.29 ± 1.06, 1.58 ± 0.56, and 1.61 ± 0.69, respectively (Table IV). In fact, many of the NESG and NESG/R3 pairs have values of Γ close to unity (Fig. 2). More specifically, the percentage of pairs with RMS_{Xtal} within two standard deviations of the corresponding mean RMS_{ens} are 7.6%, 29.2%, and 46.1% for the PDB, NESG, and NESG/R3 NMR / X-ray pairs, respectively (Table IV). However, ~70% of NESG pairs (and ~55% of NESG/R3 pairs) have RMS_{Xtal} significantly greater than the corresponding RMS_{ens}, as originally observed by Andrec *et al*.[10] for most of the NMR / X-ray pairs in the PDB data set.

Examination of discrepancies between precision and accuracy

We also examined the backbone structures of NMR ensembles and corresponding X-ray crystal structures using the PyMOL molecular graphics program.[29] The nine NESG or NESG/R3 structure pairs with smallest and largest values of Γ are shown graphically in Figures 3 and 4, respectively. Examination of these images suggests that at least some of the structures

Table IV. *Comparisons of RMS_{ens} and RMS_{Xtal} for Different Data Sets of NMR / X-Ray Pairs*

Data set	Number of pairs used in analysis	$\Gamma = RMS_{Xtal}/RMS_{ens}$ Mean (Å)	$\Gamma = RMS_{Xtal}/RMS_{ens}$ S.D. (Å)	% RMS_{Xtal} < RMS_{ens} + 2 S.D.
PDB	145	2.29	1.06	7.6%
NESG	41	1.58	0.56	29.2%
NESG/R3	39	1.61	0.69	46.1%

Structures with $RMS_{Xtal} > 3.5$ Å were excluded from this analysis.

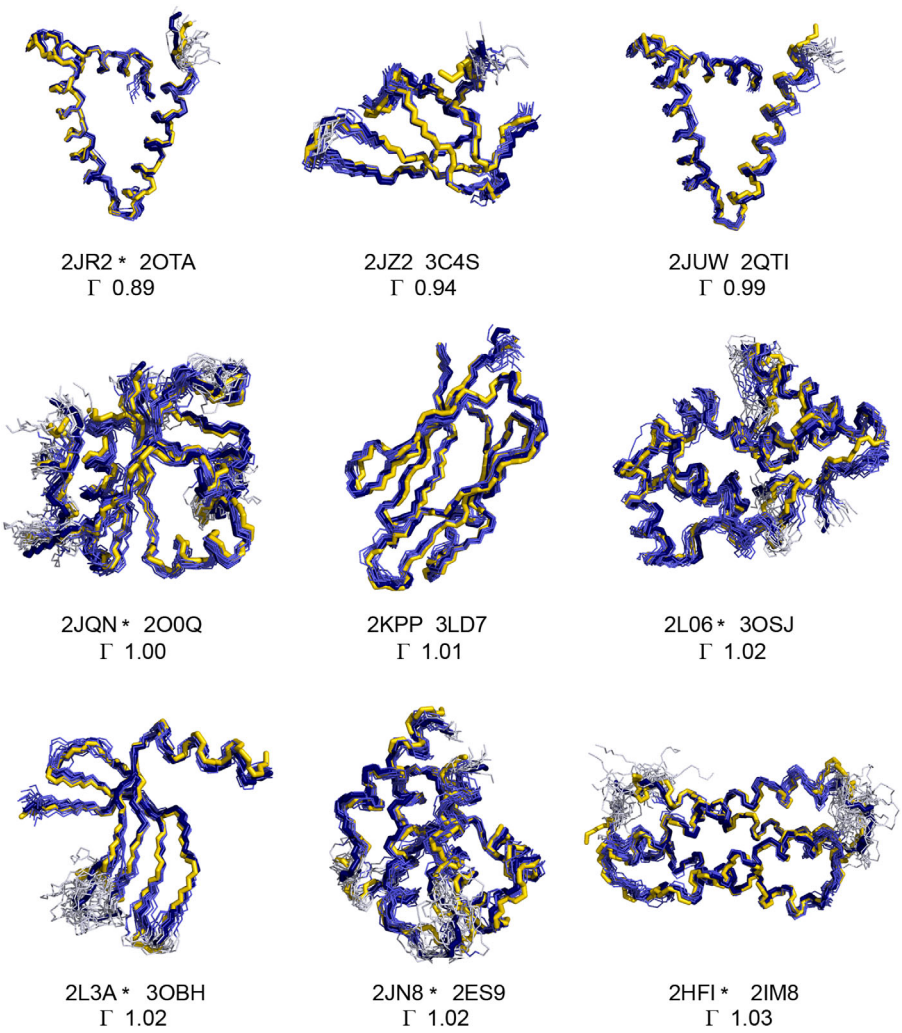

Figure 3. Examples of NESG NMR / X-ray pairs with low $\Gamma = RMS_{Xtal}/RMS_{ens}$. Nine NMR / X-ray pairs, from the NESG or NESG/R3 sets, with lowest values of Γ. For each ensemble, the superimposed backbone (N, Cα, C′) trace of the X-ray crystal structure (gold) and representative NMR conformer (dark blue) are shown, together with the converged well-defined (blue) and non-converged ill-defined (gray) backbone structures of the ensemble of the NMR conformers. Ill-defined N- and C-terminal segments are excluded from these images, but ill-defined internal loops are included. The PDB id of the NMR structure is followed by the PDB id of the X-ray structure. NMR structures refined with restrained Rosetta are indicated with asterisk following the PDB id.

with large values of Γ also have low values of RMS_{ens}; that is, these ensembles tend to be very tight bundles. Plots of Γ vs RMS_{ens} (Supporting Information Fig. S1) support this impression, although the correlation between Γ vs. RMS_{ens} is only modest.

We further examined several of the PDB NMR / X-ray pairs which have unusually high RMS_{Xtal}. Only three NESG pairs (targets CtR107, HR4435B, and UuR17A) have $RMS_{Xtal} > 2.5$ Å (Fig. 5), while none of the NESG/R3 pairs have $RMS_{Xtal} > \sim 2.0$ Å. Restrained Rosetta refinement makes the NESG structures more like the corresponding X-ray crystal structures. On the other hand, many of the 145 PDB pairs have $RMS_{Xtal} > \sim 2.5$ Å, and several have $RMS_{Xtal} > 3.5$ Å in "well defined" regions (Fig. 6). These pairs may be useful for assessing methods of protein NMR structure validation. Two of the most significantly different PDB NMR / X-ray pairs (not illustrated in Fig. 6) are PDB ID's 2EZN/3EZM ($RMS_{Xtal} = 16.3$ Å) and 1QLZ/1I4M ($RMS_{Xtal} = 19.2$ Å), which upon examination appear to be domain-swapped dimers in the crystal structures, but simple monomers or dimers in the corresponding NMR structures.

Discussion

Resource for community

The NMR / X-ray Structure Pair Data Repository was developed from experimental data sets generated in the NESG program as a resource for the broader structural bioinformatics community. Although most of these same data are available in the Protein Data Bank and BioMagResDB, it is valuable to have the NMR / X-ray structure pairs

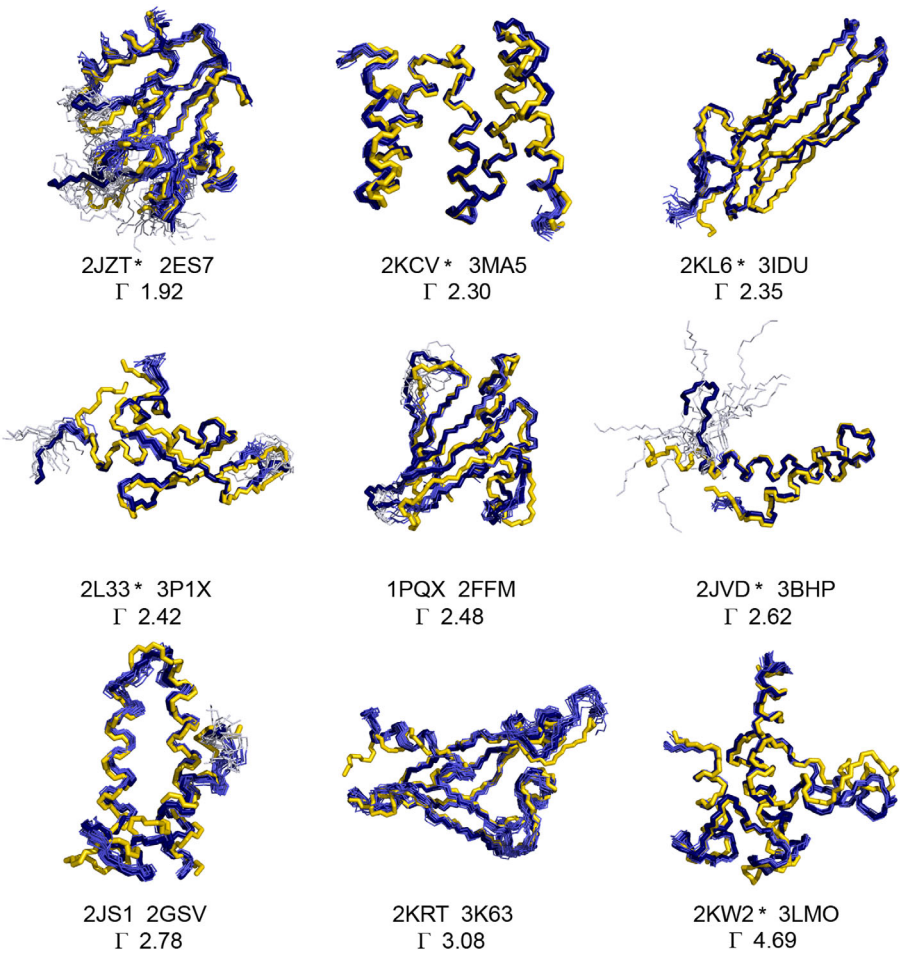

Figure 4. Examples of NESG NMR / X-ray pairs with high RMS_{Xtal} / RMS_{ens}. Nine NMR / X-ray pairs, from the NESG or NESG/R3 sets, with highest values of Γ. The color coding of backbone traces is the same as in the legend of Figure 3. The PDB id of the NMR structure is followed by the PDB id of the X-ray structure. NMR structures refined with restrained Rosetta are indicated with asterisk following the PDB id.

collected together on one site and characterized as a consistent data set for new methods development. These structures include all α, all β, α + β and α /β structures (Supporting Information Table S4 and Fig. S2). The NMR / X-ray Structure Pair Data Repository complements more comprehensive databases such at the PDB[30] and BioMagResDB,[31] as a specialized site for NMR / X-ray structure pairs. Of particular value are unprocessed NOESY time domain free-induction decay (FID) and peak list data for proteins for which both NMR and X-ray crystal structures have been determined. Another asset of this data set are the crystallization conditions listed in Supporting Information Table S2 which were not previously available to the community. Protein expression vector plasmids for these

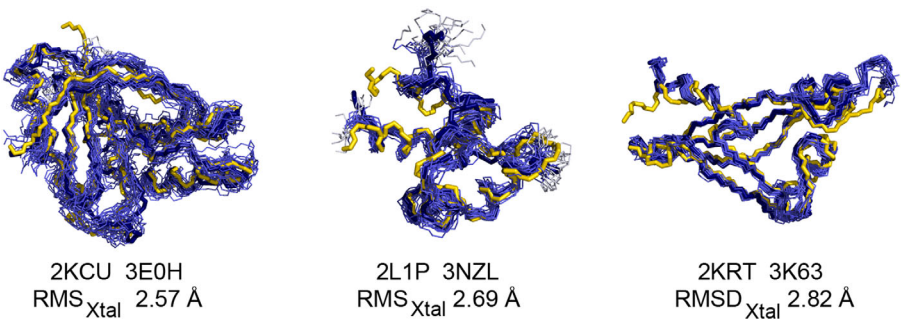

Figure 5. Examples of NMR / X-ray pairs with lower structural similarity. The three NMR / X-ray pairs from the NESG set with $RMS_{Xtal} > 2.5$ Å. The color coding of backbone traces is the same as in the legend of Figure 3. The PDB id of the NMR structure is followed by the PDB id of the X-ray structure.

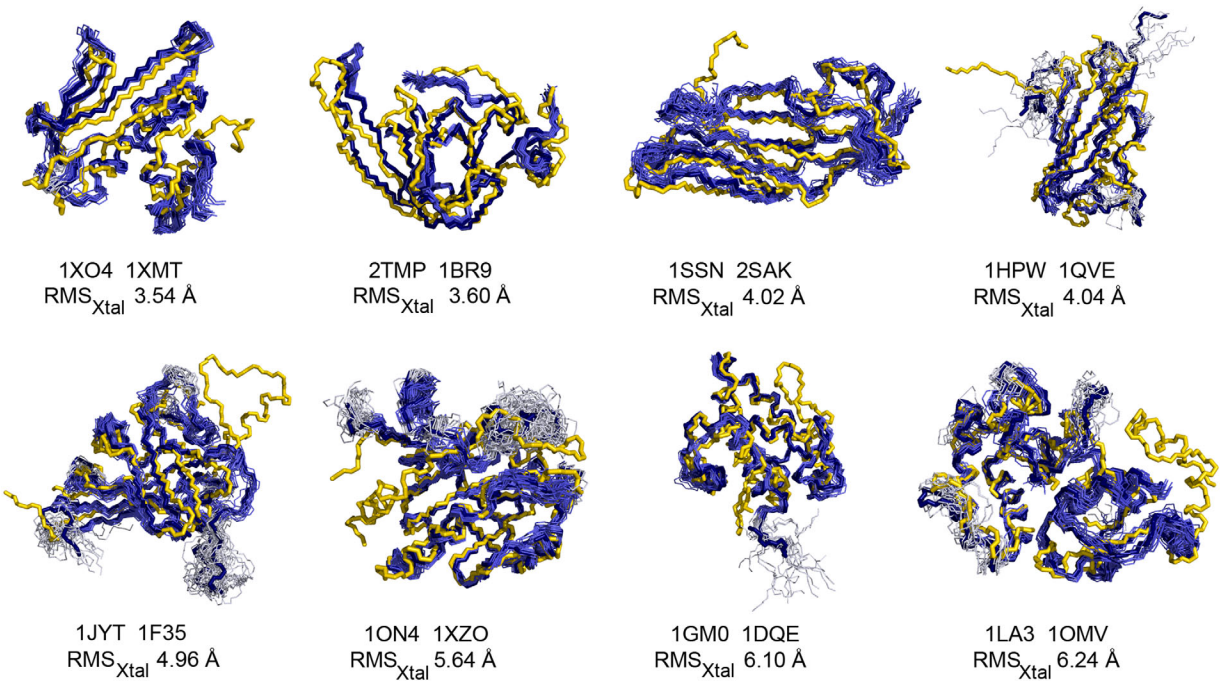

Figure 6. Examples of NMR / X-ray pairs with very low structural similarity. Eight NMR / X-ray pairs from the PDB set with low structural similarity ($RMS_{Xtal} > 3.5$ Å). The color coding of backbone traces is the same as in the legend of Figure 3. The PDB id of the NMR structure is followed by the PDB id of the X-ray structure. These "outliers" were excluded from the statistical analyses.

same protein constructs are also available from the PSI Materials Repository (PSI-MR) (http://psimr.asu.edu), which will enable further studies of these proteins and/or engineered variants. It is anticipated that this curated set of NMR / X-ray structure pairs will continue to grow as new NMR / X-ray structure pairs are completed.

The major point of this work is to organize the NESG NMR data (e.g., NOESY peak lists and RDC data) for NMR / X-ray pairs, and make them more accessible to the community for methods development. As a representative example of the kind of analysis that can be done, we compared NMR and X-ray structures determined by the NESG consortium. These structures were determined using tools like *Talos+* and *Rosetta* which use information from X-ray crystal structures available in the Protein Data Bank. In earlier work,[8] we assessed and discussed the impact of Rosetta refinement; the resulting structures are indeed closer to the corresponding X-ray crystal structures. The Rosetta refined structures also have stronger phasing power in molecular replacement studies. This does not appear to be an artifact, but rather an improvement in the NMR structures resulting from using low energy conformations of fragments from the Protein Data Bank.[8]

Analysis of "well-defined" and "ill-defined" regions of the protein structure

One of the most commonly used and generally accepted methods for distinguishing 'well-defined' (i.e., converged) from "ill-defined" (i.e., not well-converged) residue backbones is the dihedral angle order parameter (DAOP).[32] This method underlies the *Cyrange* program,[33] the recommended convention for distinguishing well-defined and ill-defined regions of solution protein NMR structures.[26] As discussed elsewhere,[11,25,27] the DAOP method has the advantage of being fast, simple, and widely used by the protein NMR community. However, it has some significant shortcomings. The DAOP cannot distinguish local from long-range order; for example, 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.[34] 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." In this work, we used the "*FindCore2*" variance matrix algorithm[27] 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. Comparisons of these two methods, DAOP and *FindCore2*, have been presented elsewhere.[25,27] The resulting well-defined and ill-defined residue ranges identified by *FindCore2* are similar to those indicated by *Cyrange*. However, these earlier studies[25,27] also demonstrate special value of atom specific designators of structural precision over the

current standard convention of defining only residue ranges of the well-defined regions of the protein NMR structure.

For the NESG and NESG/R3 structures pairs, there is generally good agreement between NMR and X-ray coordinates in regions that are well-defined in the corresponding NMR structures. However for some of the structure pairs from the PDB set shown in Figure 6, even well-converged regions of the NMR structure may diverge significantly from the corresponding X-ray crystal structure. There is no simple metric to predict where well-defined regions of the NMR structures diverge from the corresponding X-ray crystal structures.

We also assessed the question of whether missing X-ray electron density relates to particular characteristics of the NMR ensemble. There is generally a good correlation between well-defined atoms in the NMR structures and atoms for which positions could be determined from the electron density; most residues that are well-defined in the NMR structures also have reported atomic coordinates and electron density. The fraction of well-defined residues in the NMR structures that do not have electron density is $< 1\%$ (i.e., 0.39%, 0.53%, and 0.09% for the NESG, NESG/R3, and PDB structure pair sets, respectively). The few well-defined residues in the NMR structures lacking electron density are tabulated in Supporting Information Table S5.

Ground truth structures and protein dynamics

Throughout this analysis of structural similarities between NMR and corresponding X-ray crystal structures, it was implicitly assumed that the X-ray crystal structure is the "ground truth structure." This is clearly an oversimplification. The protein structure in solution samples is a Boltzmann distribution of states, in dynamic equilibrium. The conformation(s) selected for in the crystallization process need not correspond to the most populated conformer in solution, as lattice packing effects may stabilize excited conformational states of the protein structure. These intrinsic dynamics also undermine a basic assumption made by most NMR-based structure modeling methods: that every conformer in the ensemble should best-fit all of the experimental distance, chemical shift, and residual dipolar coupling data. In fact, a Boltzmann distribution of conformers should be modeled to fit these ensemble-averaged NMR data. The X-ray crystal structures themselves also have uncertainties in atomic positions which were not accounted for in our analyses. In addition, these X-ray crystal structures were all solved under cryogenic conditions ($\sim$70 K), while the NMR structures were determined at 298–303 K.[35] In some cases, crystallization required conditions of pH and buffers that are significantly different than those used in the solution NMR structures. Hence, these NMR / X-ray pairs have intrinsic shortcomings for assessing the "accuracy" of the NMR structures. Indeed, there is no "single ground truth structure." Rather, the solution structure of the protein is a condition-dependent, Boltzmann-weighted distribution of conformers, and ideally should be modeled as such based on the experimental NMR data. The NMR / X-ray Structure Pair Data Repository provides useful data for developing such methods.

RMS_{Xtal} and RMS_{ens}

Andrec *et al.*[10] pointed out that most NMR structures are more different from the corresponding X-ray crystal structure than would be anticipated from the uncertainty in atomic coordinates indicated by RMS_{ens}. Some 93% of the PDB NMR / X-ray structure pairs used in these earlier studies have RMS_{Xtal} larger than two standard deviations of RMS_{ens} above the corresponding mean value of RMS_{ens}. On the other hand, 30% of NESG NMR / X-ray structure pairs have values of Γ close to unity, with RMS_{Xtal} within two standard deviations of the corresponding mean value of RMS_{ens}. The restrained-Rosetta protocol increases this to about 45% of the NMR / X-ray pairs. Many of these NESG structures were refined using backbone ^{15}N–^{1}H residual dipolar coupling data and restrained energy minimization in explicit water using the program CNS. Visual examination of NESG and NESG/R3 pairs with $RMS_{Xtal} \gg RMS_{ens}$ reveals that many of these have especially low values of RMS_{ens}. This suggests that some methods used to generate NMR structure ensembles may underestimate the uncertainty of NMR structure atomic coordinates,[11,36–38] and illustrates the urgent need to develop more statistically reliable methods for estimating such uncertainty, such as Bayesian methods that can propagate uncertainties in experimental measurements to uncertainties in atomic positions.[39,40]

Despite advances over the past two decades in the analysis of protein structures and dynamics from NMR data, the full potential of NMR data for modeling the dynamic structures of proteins has not yet been realized. Our analysis of NESG NMR / X-ray pairs presented in this study opens as many questions as it answers. The work demonstrates the need to develop improved computational methods for modeling dynamic protein structures as Boltzmann ensembles,[41–45] the shortcomings of the current NMR ensembles in estimating the uncertainty and precision of NMR structure models,[39,40] and the challenges in using X-ray crystal structures as ground truth descriptions of solution-state protein structures.[7,8,10] Hopefully, the NMR / X-ray Structure Pair Data Repository (http://spine.nesg.org/nmrdata), together with access to the expression plasmids and crystallization conditions, will challenge and enable the broader structural bioinformatics modeling community to develop new methods and algorithms to address these important technological issues that define our understanding of protein structure.

Methods and Materials

Protein sample production

Proteins were produced using standard protocols for protein sample production in *E. coli* expression hosts. Genes were cloned from genomic DNA, cDNA libraries, or from synthetic genes, into modified pET expression systems,[46] expressed in BL21(DE3) *E. coli*, and purified following standard protocols of the NESG consortium.[47–49] [*U*-^{15}N, 5%^{13}C]-, [*U*-^{15}N, *U*-^{13}C]-, and [*U*-^{2}H,^{13}C,^{15}N; ^{1}H-Ile-δ1,Leu-δ,Val-γ Ala-β protonated]-enriched proteins were expressed using MJ9 minimal media.[50] Selenomethione (SeMet) was incorporated using MJ9 media supplemented with SeMet.[51] The [*U*-^{15}N, 5%^{13}C]-enriched protein samples were generated for stereo-specific assignments of isopropyl methyl groups of valines and leucines[52] and for RDC measurements.[53] The final purified protein samples generally include a short N-terminal tag, with sequence MGH$_6$SHM or a short C-terminal Hexa-His tag. Samples were characterized by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (typically > 98% homogeneous) and Matrix Assisted Laser Desorption/Ionization - Time of Flight (MALDI-TOF) mass spectrometry (generally > 98% isotope enriched). Buffer optimization using [^{15}N-^{1}H]- heteronuclear single quantum coherence (HSQC) spectra recorded using a robotic sample changer and 600 MHz 1.7-mm micro cryogenic probe was often carried out, as described previously.[54] Oligomerization states were determined using analytical gel filtration with static light scattering detection.[47–49] The purified proteins were dissolved at 0.2–1.5 m*M* concentrations in 90% ^{1}H$_2$O/10% ^{2}H$_2$O NMR buffers,[41,42] which are defined in the PDB header files. Oligomerization states under NMR conditions were generally verified in the NMR tube by 1D ^{15}N T_1/T_2 measurements.[54]

NMR data collection and structure determination

NMR data were recorded at Miami University, Pacific Northwest National Laboratory, Rutgers, The State University of New Jersey, The State University of New York at Buffalo, and The University of Georgia, using Varian or Bruker AVANCE 600, 750, 800, and/or 850 MHz spectrometers. NMR data were generally processed using the program *NMRPipe*[55] and visualized using the program *SPARKY3* (T. D. Goddard and D. G. Kneller, University of California, San Francisco). Protein NMR spectra were generally obtained at 298 K, except where noted differently in the PDB header file, and chemical shifts were referenced to internal 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS). Resonance assignments and structures were determined by protocols that have been outlined in detail elsewhere.[56–58] Additional details of data collection and processing are provided on the NESG Wiki site (http://www.nmr2.buffalo.edu/nesg.wiki/Main_Page). Sequence-specific resonance assignments were determined by automated methods using the programs *AutoAssign*[59,60] and/or *PINE*,[61] followed by manual analysis and refinement. Backbone dihedral angle restraints were derived from chemical shift data using the program *TALOS+*.[62] Residual dipolar coupling measurements were made at University of Georgia on a 600 MHz Varian INOVA spectrometer. Samples were aligned with phage, polyacrylamide gel, or polyethylene-glycol-alkyl bicelles, and RDCs were collected using either interleaved HSQC-TROSY or *J*-modulation sequences, as described elsewhere.[63,64] 3D structures were generally determined using initial automated analysis with the programs *CYANA*,[65] and/or *AutoStructure*,[66] followed by manual refinement of NOESY peak list data. NMR structures were further refined by restrained energy minimization using the program *CNS*[67] with explicit water, *Xplor-NIH*,[68] *CNSw*,[69] and/or restrained *Rosetta*,[8] as outlined on the NESG Wiki site (http://www.nmr2.buffalo.edu/nesg.wiki/Main_Page).

Validation of the experimental NMR structures

NMR data statistics, structural statistics, and global structure quality factors including *Verify3D*,[70] *ProsaII*,[71] *PROCHECK*,[72] and *MolProbity*[73] raw and statistical *Z*-scores were computed using the *Protein Structure Validation Software (PSVS)* server.[74] Resonance assignments were validated using the *Assignment Validation Software (AVS)* server.[75] The global goodness-of-fit of the final structure ensembles with the NOESY peak list data was determined using the *RPF*/*DP* analysis program[76,77] which assesses how well structural models fit with the NOESY peak list and chemical shift assignment data. The NESG standards for global structure-quality scores [*Z* scores computed using *PSVS* ver1.4 for knowledge-based structure quality metrics *Verify3D*, *ProsaII*, *PROCHECK* (backbone and all dihedral angles), and *MolProbity* clashscore that are more positive than $Z = -3$, and DP score-> 0.73] were used on more recent (i.e., deposited after 2007) NMR structures to evaluate the quality of each structure. In addition, a closer examination of the local structure quality using graphical *RPF*/*DP* analysis tools was performed to identify potential problem areas in the structural models. If either global or local structural problems were identified, then this information was reported back to the researcher performing the structure determination, and the researcher was asked to carefully re-examine the experimental data to resolve these issues. Chemical shift and other NMR data were deposited in the Biological Magnetic Resonance Bank[31] and coordinates in the Protein Data Bank.[30]

X-ray crystallography and structure determination

Most initial crystallization conditions were identified using a robotic 1536-well microbatch-under-oil screen in which plates were incubated for 1 week at 4°C before transfer to 18°C for continued observation for 4–6 weeks.[78,79] After optimization, crystals useful for structure determination were generally grown in drops composed of 1.0 μL of protein and 1.0 μL of precipitant solution under paraffin oil. The crystals were generally cryoprotected with ethylene glycol or glycerol prior to flash-freezing in liquid nitrogen for data collection. In most cases, selenomethionyl single- or multiple-wavelength anomalous diffraction (SAD or MAD) data sets[80] were collected at the peak wavelength of the selenium K edge using beamline X4A at the National Synchrotron Light Source ($\lambda = 0.97903$ Å). The diffraction data were processed with the *HKL2000* package.[81] Some structures were solved by molecular replacement[82] using the 3D structures of homologs as templates (Supporting Information Table S2). The programs *Shelxe*/*d*[83] or *Resolve*[84] were used to locate a selenium site and to calculate phases. The models were completed using iterative cycles of manual rebuilding in *Coot*[85] and then refined using the programs *Phenix*[86] or *CNS*.[67] The quality of the final structure was assessed using the *PSVS*[74] server, including *PROCHECK*.[72] The atomic coordinates and structure factors for all structures were deposited in the Protein Data Bank.[30]

Calculation of structural superimpositions and RMS distances

For each NMR and X-ray structure pair, the amino acid sequences for which coordinates were available were aligned and the coordinate files were edited so that both the NMR and X-ray structure included identical subsets of the amino acid sequence. For oligomeric structures, discrepancies arise since the oligomer state, and even the protein–protein interface, can be different between the NMR structure and the biological unit reported for the crystal structure. For this reason, only chain A of multiple-chain NMR and/or X-ray structures was used for structural comparisons. NMR ensembles were analyzed using the *FindCore2*[27] software module of the *PDBStat* software[25] to identify well-defined and ill-defined backbone coordinates. Following the recommendations of the world-wide PDB NMR Structure Validation Task Force, the well-defined regions were then used to determine the medoid conformer[26]; that is, the representative conformer from the NMR ensemble most similar to all of the other conformers in the ensemble. The RMSD of each conformer (for well-defined regions) to this representative conformer was then computed. The average value of this RMSD (RMS_{ens}) provides a measure of the precision of the NMR ensemble. This representative structure was then superimposed on the X-ray crystal structure using residues common to the well-defined regions of the NMR structure and the reported coordinates of the X-ray crystal structure (i.e., excluding atoms which are not observed in the X-ray crystal structure). These superimposed pairs were then used to determine the RMSD between NMR and crystal structures, RMS_{Xtal}. $<RMS_{Xtal}>$ was calculated by computing the backbone RMSD between each conformer of the NMR ensemble and the X-ray crystal structure, for well-defined regions of the NMR structure ensemble, and averaging these values. As shown in Supporting Information Figure S3, RMS_{Xtal} and $<RMS_{Xtal}>$ are essentially the same when calculated using the well-defined regions of the protein NMR structure.

Statistical methods

Structure superimpositions and RMSD values were computed with the *PDBStat* program.[25] Binned RMSD distributions were compared using the two-sample Kolmogorov–Smirnov test[28] in order to determine whether they were significantly different from one another.

Acknowledgments

Authors thank all of the members of the Northeast Structural Genomics Consortium who generated and archived NMR data, particularly scientists in the laboratories of C. Arrowsmith, J. Hunt, M. Kennedy, G.T. Montelione, R. Powers, T. Szyperski, L. Tong, and J. Prestegard. Some of the NMR data used in this work was acquired in the Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory, Richland WA.

References

1. Montelione GT, Anderson S (1999) Structural genomics: keystone for a Human Proteome Project. Nat Struct Biol 6:11–12.
2. Liu J, Montelione GT, Rost B (2007) Novel leverage of structural genomics. Nat Biotechnol 25:849–851.
3. Burley SK, Joachimiak A, Montelione GT, Wilson IA (2008) Contributions to the NIH-NIGMS protein structure initiative from the PSI production centers. Structure 16:5–11.
4. Montelione GT (2012) The protein structure initiative: achievements and visions for the future. F1000 Biol Rep 4:7.
5. Graslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, Ming J, dhe-Paganon S, Park HW, Savchenko A, Yee A, Edwards A, Vincentelli R, Cambillau C, Kim R, Kim SH, Rao Z, Shi Y, Terwilliger TC, Kim CY, Hung LW, Waldo GS, Peleg Y, Albeck S, Unger T, Dym

O, Prilusky J, Sussman JL, Stevens RC, Lesley SA, Wilson IA, Joachimiak A, Collart F, Dementieva I, Donnelly MI, Eschenfeldt WH, Kim Y, Stols L, Wu R, Zhou M, Burley SK, Emtage JS, Sauder JM, Thompson D, Bain K, Luz J, Gheyi T, Zhang F, Atwell S, Almo SC, Bonanno JB, Fiser A, Swaminathan S, Studier FW, Chance MR, Sali A, Acton TB, Xiao R, Zhao L, Ma LC, Hunt JF, Tong L, Cunningham K, Inouye M, Anderson S, Janjua H, Shastry R, Ho CK, Wang D, Wang H, Jiang M, Montelione GT, Stuart DI, Owens RJ, Daenke S, Schutz A, Heinemann U, Yokoyama S, Bussow K, Gunsalus KC (2008) Protein production and purification. Nat Methods 5:135–146.

6. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C (2009) PSI-2: structural genomics to cover protein domain family space. Structure 17:869–881.
7. Mao B, Guan R, Montelione GT (2011) Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure 19: 757–766.
8. Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. J Am Chem Soc 136:1893–1906.
9. 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.
10. Andrec M, Snyder DA, Zhou Z, Young J, Montelione GT, Levy RM (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. Proteins 69:449–465.
11. Snyder DA, Bhattacharya A, Huang YJ, Montelione GT (2005) Assessing precision and accuracy of protein structures derived from NMR data. Proteins 59: 655–661.
12. Elshahawi SI, Ramelot TA, Seetharaman J, Chen J, Singh S, Yang Y, Pederson K, Kharel MK, Xiao R, Lew S, Yennamalli RM, Miller MD, Wang F, Tong L, Montelione GT, Kennedy MA, Bingman CA, Zhu H, Phillips GN Jr, Thorson JS (2014) Structure-guided functional characterization of enediyne self-sacrifice resistance proteins, CalU16 and CalU19. ACS Chem Biol 9:2347–2358.
13. Mueller GA, Choy WY, Yang D, Forman-Kay JD, Venters RA, Kay LE (2000) Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. J Mol Biol 300:197–212.
14. Tugarinov V, Choy WY, Orekhov VY, Kay LE (2005) Solution NMR-derived global fold of a monomeric 82-kDa enzyme. Proc Natl Acad Sci USA 102:622–627.
15. Hiller S, Garces RG, Malia TJ, Orekhov VY, Colombini M, Wagner G (2008) Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. Science 321:1206–1210.
16. Gardner KH, Rosen MK, Kay LE (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. Biochemistry 36:1389–1401.
17. Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, Kay LE (1996) Selective methyl group protonation of perdeuterated proteins. J Mol Biol 263: 627–636.
18. Zheng D, Huang YJ, Moseley HN, Xiao R, Aramini J, Swapna GV, Montelione GT (2003) Automated protein fold determination using a minimal NMR constraint strategy. Protein Sci 12:1232–1246.
19. Sgourakis NG, Lange OF, DiMaio F, Andre I, Fitzkee NC, Rossi P, Montelione GT, Bax A, Baker D (2011) Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. J Am Chem Soc 133:6288–6298.
20. Thompson JM, Sgourakis NG, Liu G, Rossi P, Tang Y, Mills JL, Szyperski T, Montelione GT, Baker D (2012) Accurate protein structure modeling using sparse NMR data and homologous structure information. Proc Natl Acad Sci USA 109:9875–9880.
21. Rossi P, Shi L, Liu G, Barbieri CM, Lee HW, Grant TD, Luft JR, Xiao R, Acton TB, Snell EH, Montelione GT, Baker D, Lange OF, Sgourakis NG (2015) A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta. Proteins 83: 309–317.
22. 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.
23. Kim YK, Shin YJ, Lee WH, Kim HY, Hwang KY (2009) Structural and kinetic analysis of an MsrA-MsrB fusion protein from *Streptococcus pneumoniae*. Mol Microbiol 72:699–709.
24. Sharff AJ, Rodseth LE, Quiocho FA (1993) Refined 1.8-A structure reveals the mode of binding of beta-cyclodextrin to the maltodextrin binding protein. Biochemistry 32:10553–10559.
25. Tejero R1, Snyder D, Mao B, Aramini JM, Montelione GT (2013) PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. J Biomol NMR 56:337–351.
26. Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL (2013) Recommendations of the wwPDB NMR Validation Task Force. Structure 21:1563–1570.
27. Snyder DA, Grullon J, Huang YJ, Tejero R, Montelione GT (2014) The expanded FindCore method for identification of a core atom set for assessment of protein structure prediction. Proteins 82 (Suppl 2):219–230.
28. Clapham C, Nicholson J. 2009. The concise Oxford dictionary of mathematics. Oxford, UK: Oxford University Press.
29. The PyMOL Molecular Graphics System, Version 1.4.1 Schrödinger, LLC. New York, NY.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242.
31. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408.
32. 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.
33. Kirchner DK, Guntert P (2011) Objective identification of residue ranges for the superposition of protein structures. BMC Bioinformatics 12:170.
34. 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.
35. Fenwick RB, van den Bedem H, Fraser JS, Wright PE (2014) Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. Proc Natl Acad Sci USA 111:E445–E454.
36. Spronk CA, Nabuurs SB, Bonvin AM, Krieger E, Vuister GW, Vriend G (2003) The precision of NMR structure ensembles revisited. J Biomol NMR 25:225–234.
37. Laughton CA, Orozco M, Vranken W (2009) COCO: a simple tool to enrich the representation of conformational variability in NMR structures. Proteins 75:206–216.
38. Buchner L, Guntert P (2015) Increased reliability of nuclear magnetic resonance protein structures by consensus structure bundles. Structure 23:425–434.
39. Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. Science 309:303–306.
40. Rieping W, Nilges M, Habeck M (2008) ISD: a software package for Bayesian NMR structure calculation. Bioinformatics 24:1104–1105.
41. Showalter SA, Bruschweiler R (2007) Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: application to the AMBER99SB force field. J Chem Theory Comput 3: 961–975.
42. Showalter SA, Bruschweiler R (2007) Quantitative molecular ensemble interpretation of NMR dipolar couplings without restraints. J Am Chem Soc 129: 4158–4159.
43. Li DW, Brüschweiler R (2010) Certifiication of molecular dynamics trajectories with NMR chemical shifts. J Phys Chem Letts 2:246–248.
44. Robustelli P, Stafford KA Palmer AG III (2012) Interpreting protein structural dynamics from NMR chemical shifts. J Am Chem Soc 134:6365–6374.
45. Palmer AG III (2015) Enzyme dynamics from NMR spectroscopy. Acc Chem Res 48:457–465.
46. Studier FW, Moffatt BA (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189:113–130.
47. Acton TB, Xiao R, Anderson S, Aramini J, Buchwald WA, Ciccosanti C, Conover K, Everett J, Hamilton K, Huang YJ, Janjua H, Kornhaber G, Lau J, Lee DY, Liu G, Maglaqui M, Ma L, Mao L, Patel D, Rossi P, Sahdev S, Shastry R, Swapna GV, Tang Y, Tong S, Wang D, Wang H, Zhao L, Montelione GT (2011) Preparation of protein samples for NMR structure, function, and small-molecule screening studies. Methods Enzymol 493:21–60.
48. Xiao R, Anderson S, Aramini J, Belote R, Buchwald WA, Ciccosanti C, Conover K, Everett JK, Hamilton K, Huang YJ, Janjua H, Jiang M, Kornhaber GJ, Lee DY, Locke JY, Ma LC, Maglaqui M, Mao L, Mitra S, Patel D, Rossi P, Sahdev S, Sharma S, Shastry R, Swapna GVT, Tong SN, Wang DY, Wang HA, Zhao L, Montelione GT, Acton TB (2010) The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. J Struct Biol 172:21–33.
49. Acton TB, Gunsalus KC, Xiao R, Ma LC, Aramini J, Baran MC, Chiang Y-W, Climent T, Cooper B, Denissova NG, Douglas SM, Everett JK, Ho CK, Macapagal D, Paranji RK, Shastry R, Shih LY, Swapna GVT, Wilson M, Wu M, Gerstein M, Inouye M, Hunt JF, Montelione GT (2005) Robotic cloning and protein production platform of the Northeast Structural Genomics Consortium. Meth Enzymol 394:210–243.
50. Jansson M, Li YC, Jendeberg L, Anderson S, Montelione GT, Nilsson B (1996) High-level production of uniformly N-15- and C-13-enriched fusion proteins in Escherichia coli. J Biomol NMR 7:131–141.
51. Doublie S, Kapp U, Aberg A, Brown K, Strub K, Cusack S (1996) Crystallization and preliminary X-ray analysis of the 9 kDa protein of the mouse signal recognition particle and the selenomethionyl-SRP9. FEBS Lett 384:219–221.
52. Neri D, Szyperski T, Otting G, Senn H, Wuthrich K (1989) Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional 13C labeling. Biochemistry 28:7510–7516.
53. Prestegard JH, Bougault CM, Kishore AI (2004) Residual dipolar couplings in structure determination of biomolecules. Chem Rev 104:3519–3540.
54. Rossi P, Swapna GV, Huang YJ, Aramini JM, Anklin C, Conover K, Hamilton K, Xiao R, Acton TB, Ertekin A, Everett JK, Montelione GT (2010) A microscale protein NMR sample screening pipeline. J Biomol NMR 46:11–22.
55. 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.
56. 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.
57. Huang YJ, Moseley HN, Baran MC, Arrowsmith C, Powers R, Tejero R, Szyperski T, Montelione GT (2005) An integrated platform for automated analysis of protein NMR structures. Methods Enzymol 394:111–141.
58. Baran MC, Huang YJ, Moseley HN, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104:3541–3556.
59. 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.
60. 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.
61. 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.
62. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223.
63. Lee HW, Wylie G, Bansal S, Wang X, Barb AW, Macnaughtan MA, Ertekin A, Montelione GT, Prestegard JH (2010) Three-dimensional structure of the weakly associated protein homodimer SeR13 using RDCs and paramagnetic surface mapping. Protein Sci 19:1673–1685.
64. Eletsky A, Jeong MY, Kim H, Lee HW, Xiao R, Pagliarini DJ, Prestegard JH, Winge DR, Montelione GT, Szyperski T (2012) Solution NMR structure of yeast succinate dehydrogenase flavinylation factor Sdh5 reveals a putative Sdh1 binding site. Biochemistry 51:8475–8477.

65. Herrmann T, Guntert P, Wuthrich 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.
66. Huang YJ, Tejero R, Powers R, Montelione GT (2006) AutoStructure: a topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins 62:587–603.
67. 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 & NMR system: a new software suite for macromolecular structure determination. Acta Cryst 54:905–921.
68. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73.
69. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M (2003) Refinement of protein structures in explicit solvent. Proteins 50:496–506.
70. Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356:83–85.
71. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17: 355–362.
72. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26: 283–291.
73. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Cryst D 66:12–21.
74. Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. Proteins 66:778–795.
75. 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.
76. Huang YJ, Powers R, Montelione GT (2005) Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 127:1665–1674.
77. Huang YJ, Rosato A, Singh G, Montelione GT (2012) RPF: a quality assessment tool for protein NMR structures. Nucleic Acids Res 40:W542–W546.
78. Chayen NE, Stewart PDS, Maeder DL, Blow DM (1990) An automated-system for microbatch protein vrystallization and dcreening. J Appl Cryst 23: 297–302.
79. Luft JR, Collins RJ, Fehrman NA, Lauricella AM, Veatch CK, DeTitta GT (2003) A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J Struct Biol 142:170–179.
80. Hendrickson WA (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science 254:51–58.
81. Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 276:307–326.
82. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Cryst 40:658–674.
83. Schneider TR, Sheldrick GM (2002) Substructure solution with SHELXD. Acta Cryst D 58:1772–1779.
84. Terwilliger TC (2003) SOLVE and RESOLVE: automated structure solution and density modification. Methods Enzymol 374:22–37.
85. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Cryst D 60:2126–2132.
86. Adams PD, Grosse-Kunstleve RW, Hung L-W, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002) PHENIX: building a new software for automated crystallographic structure determination. Acta Cryst D 58:1948–1954.