J|A|C|S
ARTICLES
Published on Web 11/08/2005

Comparisons of NMR Spectral Quality and Success in Crystallization Demonstrate that NMR and X-ray Crystallography Are Complementary Methods for Small Protein Structure Determination

David A. Snyder,[†] Yang Chen,[‡] Natalia G. Denissova,[†] Thomas Acton,[†] James M. Aramini,[†] Melissa Ciano,[†] Richard Karlin,[‡] Jinfeng Liu,[§] Philip Manor,[‡] P. A. Rajan,[†] Paolo Rossi,[†] G. V. T. Swapna,[†] Rong Xiao,[†] Burkhard Rost,[§] John Hunt,*[,‡] and Gaetano T. Montelione*[,†,||]

Contribution from the Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers University, Piscataway, New Jersey 08854, Department of Biological Sciences, and Northeast Structural Genomics Consortium, Columbia University, New York, New York 10027, Department of Biochemistry and Molecular Biophysics, and Northeast Structural Genomics Consortium, Columbia University, 650 West 168th Street BB217, New York, New York 10032, and Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Piscataway, New Jersey 08854

Received May 31, 2005; E-mail: guy@cabm.rutgers.edu; jfhunt@biology.columbia.edu

Abstract: X-ray crystallography and NMR spectroscopy provide the only sources of experimental data from which protein structures can be analyzed at high or even atomic resolution. The degree to which these methods complement each other as sources of structural knowledge is a matter of debate; it is often proposed that small proteins yielding high quality, readily analyzed NMR spectra are a subset of those that readily yield strongly diffracting crystals. We have examined the correlation between NMR spectral quality and success in structure determination by X-ray crystallography for 159 prokaryotic and eukaryotic proteins, prescreened to avoid proteins providing polydisperse and/or aggregated samples. This study demonstrates that, across this protein sample set, the quality of a protein's [^{15}N−^{1}H]-heteronuclear correlation (HSQC) spectrum recorded under conditions generally suitable for 3D structure determination by NMR, a key predictor of the ability to determine a structure by NMR, is not correlated with successful crystallization and structure determination by X-ray crystallography. These results, together with similar results of an independent study presented in the accompanying paper (Yee, et al., J. Am. Chem. Soc., accompanying paper), demonstrate that X-ray crystallography and NMR often provide complementary sources of structural data and that both methods are required in order to optimize success for as many targets as possible in large-scale structural proteomics efforts.

Introduction

The unprecedented success of genome-wide sequencing efforts has given us a wealth of data on the proteins which perform the vast majority of life processes. However, in characterizing and understanding the molecular function(s) of a protein, combined knowledge of sequence and 3D structure generally provides deeper insights than sequence information alone. Currently, X-ray diffraction studies of crystallized proteins and NMR studies of isotope-enriched proteins are the two primary experimental methods providing atomic-resolution protein structural information. While both methods can reliably produce high quality structures for a wide variety of proteins targeted in structural biology and structural proteomics projects, each has its own technical limits and barriers. In particular, NMR methods for determining high-resolution structures are generally limited to smaller (<30−40 KDa) proteins, while X-ray crystallography requires crystals that provide suitable quality diffraction data. Sample preparation for both methods generally requires that the protein of interest is homogeneous, stable, reasonably soluble, and not irreversibly aggregated at high concentrations. However, even small, soluble, and monodisperse protein samples do not necessarily yield NMR spectra of high enough quality for rapid structure determination. Thus, successful determination of a protein structure by either X-ray crystallography or NMR depends on many factors that vary from protein to protein, and understanding this dependence is an area of active research.

While the factors affecting NMR spectral quality are only beginning to be explored, some have suggested that proteins

[†] Rutgers University.
[‡] Department of Biological Sciences, Columbia University.
[§] Department of Biochemistry and Molecular Biophysics, Columbia University.
[||] Robert Wood Johnson Medical School.

10.1021/ja053564h CCC: $30.25

that provide easily analyzed NMR spectra also have properties that provide diffraction quality crystals suitable for X-ray crystallography. Indeed, a recent study by Page et al. indicates that proteins shown by 1D ^{1}H NMR to have clean spectra with dispersed methyl resonance chemical shifts and line shapes typical of folded proteins also tend to provide diffraction-quality crystals.[1] However, while this work shows that NMR spectroscopy can be used to screen for samples that will exhibit crystallization success, the NMR experiment utilized is not one that is typically used for structure determination by NMR. Even proteins with well-dispersed methyl proton shifts and relatively clean 1D spectra might have other chemical shifts, in particular the backbone amide proton and nitrogen chemical shifts so critical throughout the NMR-based structure determination process, that are poorly dispersed, broad, and/or otherwise difficult to analyze. On the other hand, a protein that does not provide diffraction quality crystals might have sufficiently high quality 2D and 3D spectra, (i.e., with sufficiently sharp line widths and chemical shift dispersion) so as to still provide data sufficient for reliable NMR-based structure determination.

Due to the differences between the 1D ^{1}H NMR spectroscopy used by Page et al. and the multidimensional heteronuclear experiments used in NMR-based structure determination, the demonstrated ability to use 1D NMR spectroscopy to screen for crystallization potential does not answer the question of whether NMR spectroscopy and X-ray crystallography are complementary methods of protein structure analysis. Recent work suggests, however, that NMR and crystallography are indeed complementary.[2,3] For instance, it has long been believed and recently shown in a genomic scale context that proteins with high pI values tend to be difficult to crystallize.[4] Yet, both traditional structural biology and structural proteomics research programs have been able to solve the structures of many basic proteins reliably and rapidly using NMR. Savchenko et al., in screening 23 pairs of homologous proteins, have reported good quality NMR spectra for many targets that do not readily yield crystals, demonstrating for a limited number of proteins the complementarity of NMR and X-ray crystallography in a structural proteomics project.[3] Tyler et al. have also observed a lack of correlation between NMR HSQC quality and crystallization success for a small set of eukaryotic protein targets produced using a cell-free wheat germ expression system.[2]

This paper, and the accompanying paper by Arrowsmith and co-workers (Yee, et al., J. Am. Chem. Soc., accompanying paper), present two extensive and independent studies both demonstrating that the quality of a protein's 2D [^{15}N-^{1}H]-heteronuclear single-quantum coherence (HSQC) spectrum, a key experiment seeding the process of structure determination by solution-state NMR and hence a key predictor of success in NMR-based structure determination, does not correlate in any significant way with successful crystallization and structure determination by crystallographic methods. These studies were carried out within the context of a five-year pilot project in structural proteomics by the Northeast Structural Genomics Consortium (NESG). Between the two studies, as of May 1, 2005 over 420 different proteins, 159 proteins in this study alone, have been purified and studied by both 2D HSQC and crystallization screening. These results demonstrate clearly and conclusively that NMR and X-ray crystallography indeed do represent complementary ways of obtaining structural data in structural proteomics projects, which together can provide a more complete coverage of protein structures than is possible using exclusively one method or the other.

Materials and Methods

Protein Target Selection. The NESG strategy and practice for target selection are outlined elsewhere.[5,6] Briefly, NESG efforts focus on eukaryotic protein domain families targeted from five target organisms, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Homo sapiens*, and *Arabidopsis thaliana*, including homologues within these domain families from a wide range of prokaryotic and archaeal organisms. Domain families include only those for which no member can be modeled based on the structural information present in the PDB at the time of protein target family selection. Due to anticipated deleterious effects on experimental progress, predicted helical membrane proteins (PHDhtm[7,8]), beta-membrane proteins in eukaryotes,[9] proteins dominated by coiled-coil regions (COILS[10]), low-complexity regions (SEG[11]), or long regions without regular secondary structure (NORS[12,13]) are not included in these domain families. NORS regions were predicted using default parameters.[12,13] We considered stretches of >70 consecutive residues of which $<12\%$ are predicted helix or strands, as NORS.[5] The NESG project has focused initial efforts on full length proteins shorter than 340 residues, so as to avoid some of the special problems of multidomain proteins; over 90% of the structural domains in SCOP[14] and PrISM[15] are shorter than 340 residues.[16] In addition, hypothetical proteins shorter than 50 residues were excluded. The resulting mix of eukaryotic, prokaryotic, and archeal proteins, each generally <340-residues, selected from these clusters[5,6] comprise the protein target list used in this study. Each protein target is assigned a unique NESG identifier, in the format X{x}Y#, where X{x} is an organism code, Y is the institution code, and # is the target number; e.g. HR41 is human protein target #41 produced at the Rutgers NESG Protein Production Facility.

Protein Target Cloning and Sample Preparation. The target cloning and sample preparation were carried out at the NESG Protein Production Facility at Rutgers University, using standardized protocols described in detail elsewhere.[17] NESG target proteins used in this study were all produced with short 8−10 residue N-terminal or C-terminal hexa-His purification tags that allowed for the implementation of high-throughput parallel methods. Protein coding sequences were amplified by PCR using genomic DNA, for prokaryotic reagent organisms, or a common cDNA pool generated from messenger RNA, for eukaryotic target organisms, as a template. Resulting individual DNA fragments were then cloned into one of several modified pET expression vectors.[17] The resulting constructs were used for transformation of the

(1) Page, R.; Peti, W.; Wilson, I.; Stevens, R.; Wüthrich, K. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1901−1905.
(2) Tyler, R. C.; et al. *Proteins: Struct., Funct., Bioinform.* **2005**, *59*, 633−643.
(3) Savchenko, A.; Yee, A.; Khachatryan, A.; Skarina, T.; Evdokimova, E.; Pavlova, M.; Semesi, A.; Northey, J.; Beasley, S.; Lan, N.; Das, R.; Gerstein, M.; Arrowsmith, C.; Edwards, A. *Proteins: Struct., Funct., Bioinform.* **2003**, *50*, 392−399.
(4) Canaves, J.; Page, R.; Wilson, I.; Stevens, R. *J. Mol. Biol.* **2004**, *344*, 977−991.
(5) Liu, J.; Hegyi, H.; Acton, T. B.; Montelione, G.; Rost, B. *Proteins: Struct., Funct., Bioinform.* **2004**, *56*, 188−200.
(6) Wunderlich, Z.; Acton, T. B.; Liu, J.; Kornhaber, G.; Everett, J.; Carter, P.; Lan, N.; Echols, N.; Gerstein, M.; Rost, B. *Proteins: Struct., Funct., Bioinform.* **2004**, *56*, 181−187.
(7) Rost, B. *Methods Enzymol.* **1996**, *266*, 525−539.
(8) Rost, B.; Casadio, R.; Fariselli, P. *Protein Sci.* **1996**, *5*, 1704−1718.
(9) Schulz, G. E. *Curr. Opin. Struct. Biol.* **2000**, *10*, 443−447.
(10) Lupas, A. *Methods Enzymol.* **1996**, *266*, 513−525.
(11) Wootton, J. C.; Federhen, S. *Methods Enzymol.* **1996**, *266*, 554−571.
(12) Liu, J.; Rost, B. *Nucleic Acids Res.* **2003**, *31*, 3833−3835.
(13) Liu, J.; Tan, H.; Rost, B. *J. Mol. Biol.* **2002**, *322*, 53−64.
(14) Lo Conte, L.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. *Nucleic. Acids. Res.* **2002**, *30* (1), 264−7.
(15) Yang, A. S.; Honig, B. *J. Mol. Biol.* **2000**, *301* (3), 679−689.
(16) Liu, J.; Rost, B. *Nucleic Acids Res.* **2004**, *32*, 569−571.
(17) Acton, T. B.; et al. *Methods Enzymol.* **2005**, *394*, 210−243.

Table 1. NMR Screening and Aggregation Screening Buffers

NMR Screening Buffers

pH	buffer
pH 6.5 ± 0.1	20 mM MES, 100 mM NaCl, 5 mM $CaCl_2$, 10 mM DTT, 0.02% sodium azide, 5% D_2O
pH 5.5 ± 0.1	20 mM NaOAc, 100 mM NaCl, 5 mM $CaCl_2$, 10 mM DTT, 0.02% sodium azide, 5% D_2O
pH 4.5 ± 0.1	20 mM NaOAc, 100 mM NaCl, 5 mM $CaCl_2$, 10 mM DTT, 0.02% sodium azide, 5% D_2O

Aggregation Screening Buffers

pH	buffer
pH 7.5 ± 0.1	10 mM Tris-HCl, 5 mM DTT
pH 7.5 ± 0.1	10 mM Tris-HCl, 100 mM NaCl, 5 mM DTT

BL21(DE3)pMgK, a strain of *E. coli*, containing plasmid-derived genes for arginine and isoluecine tRNAs,[18−20] and tested for expression and solubility. The soluble targets were sequence verified and scaled up for preparative fermentation using a defined minimal media MJ9,[21] allowing for either uniform isotope-enrichment with ^{15}N or selenomethione (SeMet) labeling, as described elsewhere.[17] Cell pellets were resuspended in the buffer containing 10 mM imidazol, lysed by sonication, and insoluble components were removed by centrifugation. Proteins of interest were then purified using one-step affinity purification on nickel-charged HiTrap FPLC columns (Pharmacia) or Ni−NTA (nickel-nitrilotriacetic acid) agarose (Qiagen) open columns.

Sample Preparation for NMR Studies. NiNTA-purified, uniformly ^{15}N-enriched protein samples were subdivided into three fractions, with each fraction exchanged into one of three NMR sample buffers (pH 4.5, pH 5.5, and pH 6.5) listed in Table 1. Only buffers with pH values significantly different (>0.5 pH units) from the predicted pI of the protein were used. Samples were concentrated to 0.3 to 1 mM, transferred to 5 mm NMR tubes (Wilmad, 535PP) and stored at 4 °C for 1−10 days prior to NMR spectral analysis.

NMR Data Collection. NMR screening was performed using 500 or 600 MHz NMR spectrometers. Two-dimensional $^{15}N-^{1}H$ HSQC spectra,[22] with sweep widths sufficient for all backbone and amide side-chain amide correlation peaks to be observed without aliasing or folding, were recorded for each target protein in different NMR buffers. In addition, $^{15}N-^{1}H$ TROSY−HSQC[23] and/or one or two-dimensional $^{1}H-^{15}N$ heteronuclear NOE (HetNOE) spectra[24,25] were recorded for some samples. The experimental parameters for each experiment were adjusted based on sample characteristics to rapidly achieve optimal results; for most samples, fewer than 128 scans were required with 1028−4096 points in the direct dimension and 40−128 points in the indirect dimension. HetNOE spectra were recorded with 3.0 s NOE buildup time (3.08 s total recycle times) and water flipback pulses[24] to minimize solvent saturation transfer effects in hetNOE data.[25] HSQC and HetNOE spectral information, the raw free-induction decay (FID) data, and a representative 2D plot of the processed spectrum were archived into the SPINS database[26] and passed automatically to the project-wide SPINE data warehouse.[27]

Sample Preparation for Crystallization Studies. Prior to further purification, samples were screened for aggregation properties, and those that could not be prepared in monodisperse form were excluded from subsequent crystallization screening. NiNTA-purified SeMet-labeled protein samples were buffer-exchanged into different Aggregation Screening Buffers (Table 1), concentrated to ~10 mg/mL, and evaluated by analytical gel filtration followed by static light scattering as described previously.[17] Briefly, aliquots of up to 70 μL in volume were passed through a Shodex gel filtration column (KW-802.5) using an AKTA HPLC system running at 0.5 mL/min at 4 °C with a buffer consisting of 100 mM Tris (pH 7.5), 100 mM NaCl, and 250 ppm sodium azide. Room temperature measurements of refractive index and static light scattering at three angles (45°, 90°, and 135°) using an Optilab DSP Refractometer (Wyatt Technology) and a Dawn EOS Static Light Scatterer (Wyatt Technology) followed gel filtration. These measurements provide an estimate of the shape-independent weight-molecular mass (MW_w) distribution of biopolymers in the sample. The subset of proteins exhibiting monodispersive hydrodynamic properties (i.e., exclusively monomeric, exclusively dimeric, etc.) was then further purified by gel filtration chromatography in the particular Aggregation Screening Buffer found to provide monodisperse mass distributions in the analytical analysis. The resulting gel-filtration purified proteins were >98% homogeneous (based on SDS polyacrylamide gel electrophoresis with Commassie blue staining) and were verified with respect to molecular mass by MALDI-TOF mass spectrometry. These protein samples were then concentrated to ~10 mg/mL, flash frozen in 50 μL aliquots, and shipped on dry ice to the (i) Columbia NESG Crystallization Facility and (ii) High Throughput Protein Crystallization Laboratory of the Hauptman-Woodward Research Institute (HWRI). These frozen 50 μL aliquots were then used as stock solutions for crystallization screening.

Crystallization Screening and X-ray Crystallography. For all samples, crystallization conditions were first screened at HWRI using a 1536-well high-throughput robotic format.[28] These data were then transferred to the Columbia University group and used to design finer crystallization screens in a 96-well format. Prior to this second stage of crystallization trials, a literature search was carried out, based on knowledge of the target protein's function, to identify potential ligands that might facilitate protein crystallization, and the crystallization trials were conducted in the presence of suitable ligand(s). Leads from these screens were optimized using a wide variety of traditional approaches facilitated by a Tecan Genesis 200 liquid handling robot. Crystal structures from the NESG consortium generally are determined by NESG scientists at Columbia University, using data collected primarily at the National Synchrotron Light Source at Brookhaven National Laboratories. To date, a subset of proteins screened in this way has yielded diffraction quality crystals and 3D crystal structures, which have been deposited in the Protein Data Bank.

Plots and Statistical Calculations. Microsoft Excel was used to generate plots of crystallization success as a function of spectral quality. The Cochran−Armitage trend test was performed using the SAS software package and the Kolmogorov−Smirnov test using the "Javascript E-labs Learning Objects for Decision Making" website.

Results

Protein samples are produced and characterized for the Northeast Structural Genomics Consortium at two sites, the Rutgers University NESG Protein Production Facility and the University of Toronto NESG Protein Production Facility. A statistical analysis of crystallization success rates and NMR spectral qualities for samples produced at the Toronto facility is presented in the accompanying paper (Yee, et al., J. Am.

(18) Ikemura, T. *Mol. Biol. Evol.* **1985**, *2*, 13−34.
(19) Sorensen, M. A.; Kurland, C. G.; Pedersen, S. *J. Mol. Biol.* **1989**, *207*, 365−377.
(20) Chen, G. F.; Inouye, M. *Nucleic Acids Res.* **1990**, *18*, 1465−1473.
(21) Jansson, M.; Li, Y.-C.; Jendeberg, L.; Anderson, S.; Montelione, G.; Nilsson, B. *J. Biomol. NMR* **1996**, *7*, 131−141.
(22) Kay, L.; Keifer, P.; Saarinen, T. *J. Am. Chem. Soc.* **1992**, *114*, 10663−10665.
(23) Pervushin, K.; Riek, R.; Wider, G.; Wüthrich, K. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 12366−12371.
(24) Grzesiek, S.; Bax, A. *J. Am. Chem. Soc.* **1993**, *115*, 12593−12594.
(25) Li, Y.-C.; Montelione, G. *J. Magn. Reson.* **1994**, *B 105*, 45−51.
(26) Baran, M. C.; Haung, Y. J.; Moseley, H. N.; Montelione, G. T. *Chemical Reviews* **2004**, *104*, 3451−3555.
(27) Goh, C.-S.; Lan, N.; Echols, N.; Douglas, S.; Milburn, D.; Bertone, P.; Xiao, R.; Ma, L.-C.; Zheng, D.; Wunderlich, Z.; Acton, T. B.; Montelione, G.; Gerstein, M. *Nucleic Acids Res.* **2003**, *31*, 2833−2838.
(28) Luft, J.; Collins, R.; Fehrman, N.; Lauricella, A.; Veatch, C.; DeTitta, G. *J. Struct. Biol.* **2003**, *142*, 170−179.

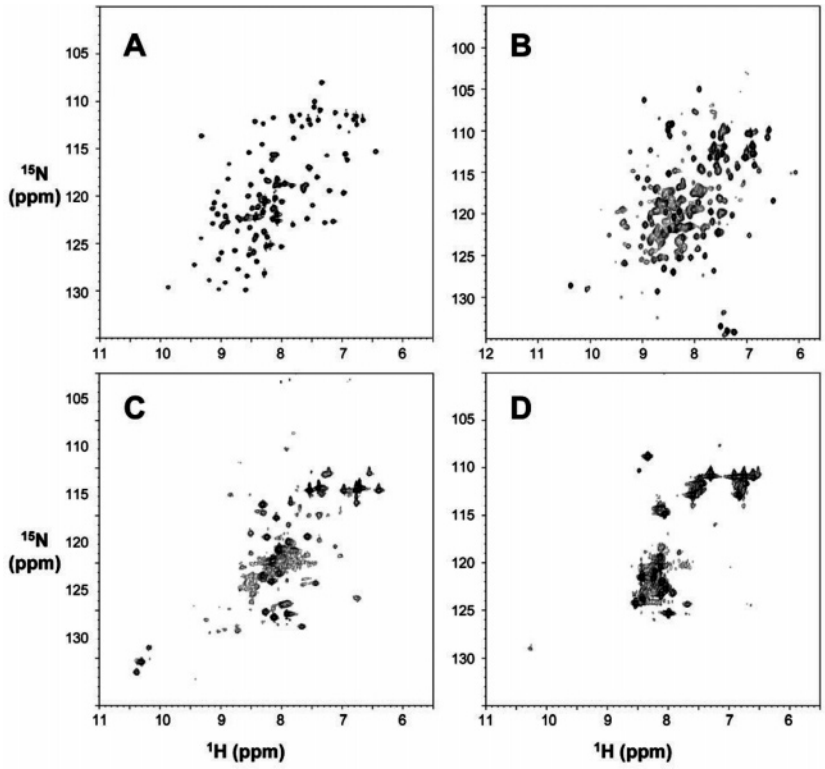

Figure 1. Representative examples of HSQC spectra classified as Excellent, Good, Promising, and Poor/Unfolded. (A) "Excellent" spectrum from NESG target MrR16, (B) "Good" spectrum from target SeR24, (C) "Promising" spectrum from target NeR5, and (D) "Poor/Unfolded" spectrum from target SR212. These spectra were recorded at pH 6.5 and 20 °C.

Chem. Soc., accompanying paper). Here we report results for 159 NESG target proteins produced, purified, and screened for NMR spectral quality at the Rutgers Protein Production Facility and also screened for crystallization and diffraction by the Columbia NESG Protein Crystallization Facility. In this study, successful crystallization/diffraction is defined as diffraction data collection, crystallographic structure determination, and deposition of the structure into the PDB.

Samples for NMR screening were prepared in one, two, or all three of the NMR Buffers listed in Table 1. Potential ligands were not included in these NMR samples. NMR spectral quality was assessed by recording conventional 2D HSQC spectra at 500 or 600 MHz, with the sample probe equilibrated at 20 °C. Proteins exhibiting poor solubility and/or extensive precipitation in all three of these buffers were excluded from further analysis. The resulting NMR spectra were subjectively scored by protein NMR experts as (i) Excellent, (ii) Good, (iii) Promising, and (iv) Poor and/or Unfolded, and stored in the SPINS NMR spectral Laboratory Information Management System.[29] Examples of spectra assigned to each of these Spectral Quality score classes are presented in Figure 1. These subjective assessments were based primarily on spectral dispersion, line widths, and numbers of resolved peaks observed compared to the number expected from the amino acid sequence. Grades of Excellent, Good, and Promising primarily reflect the completeness and quality of the spectrum in hand and generally predict how feasible it will be to determine a three-dimensional structure from NMR data.

Intrinsically unfolded proteins generally have a characteristic pattern of peaks in an HSQC spectrum (e.g., all side-chain amide ^{15}N−^{1}H correlation peaks are degenerate). However, in our experience HSQC information alone is sometimes not sufficient to differentiate among a folded protein with highly degenerate proton chemical shifts, a protein in which folded and unfolded states are in equilibrium, a molten globule which may be induced to fold under the proper circumstances, and a largely disordered unfolded protein.[30−33] For this reason, 1D or 2D heteronuclear ^{1}H−^{15}N NOESY (HetNOE) spectra[24,25] were also recorded for many of these protein samples and could be used in some cases to distinguish among samples providing "Poor" spectra those that appear to be intrinsically "Unfolded" under the conditions of these NMR measurements. Although it is possible to use such data to characterize intrinsically unfolded proteins, the distinction between different degrees of disorder complicates this analysis, and for the purposes of this study the two categories of "Poor" and "Unfolded" proteins were combined in the statistical analyses.

The overall NMR Spectral Quality score for each protein target is reported as the best quality score observed for conventional HSQC spectra in the one to three buffer conditions screened. A summary of these Quality Scores for each of 159 proteins analyzed by both NMR and crystallization screening, together with the PDB IDs for those proteins which have provided crystal and/or NMR structures, is presented in Table S1 in the Supporting Information. Thirty-three (i.e., 21%) of these proteins have provided crystal structures that are deposited in the PDB.

(29) Baran, M. C.; Moseley, H. N. B.; Sahota, G.; Montelione, G. *J. Biomol. NMR* **2002**, *24*, 113−121.
(30) Hennig, M.; Bermel, W.; Spencer, A.; Dobson, C.; Smith, L.; Schwalbe, H. *J. Mol. Biol.* **1999**, *288*, 705−723.
(31) Logan, T.; Theriault, Y.; Fesick, S. *J. Mol. Biol.* **1994**, *236*, 637−648.
(32) Penkett, C.; Redfield, C.; Jones, J.; Dodd, I.; Hubbard, J.; Smith, R.; Smith, L.; Dobson, C. *Biochemistry* **1998**, *37*, 17054−17067.
(33) Schwalbe, H.; Fiebig, K.; Buck, M.; Jones, J.; Grimshaw, S.; Spencer, A.; Glaser, S.; Smith, L.; Dobson, C. *Biochemistry* **1997**, *36*, 8977−8991.

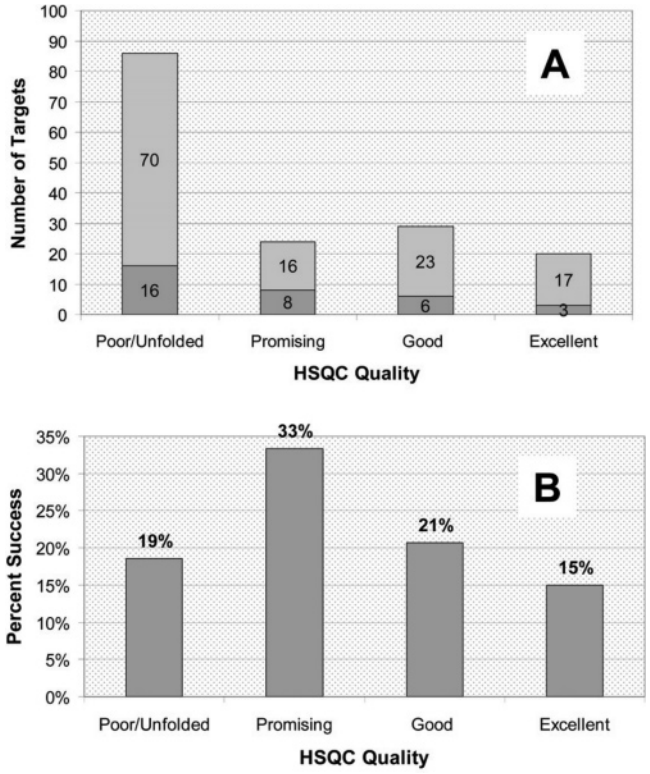

Figure 2. Successful structure determination by X-ray crystallography vs HSQC Quality. (A) Stacked histogram plot of the number of proteins for which NESG X-ray crystallographic studies resulted in high-resolution protein structures (dark gray) and those for which no structures have been obtained (light gray). (B) Percentage of proteins in each HSQC quality class yielding structures by X-ray crystallography.

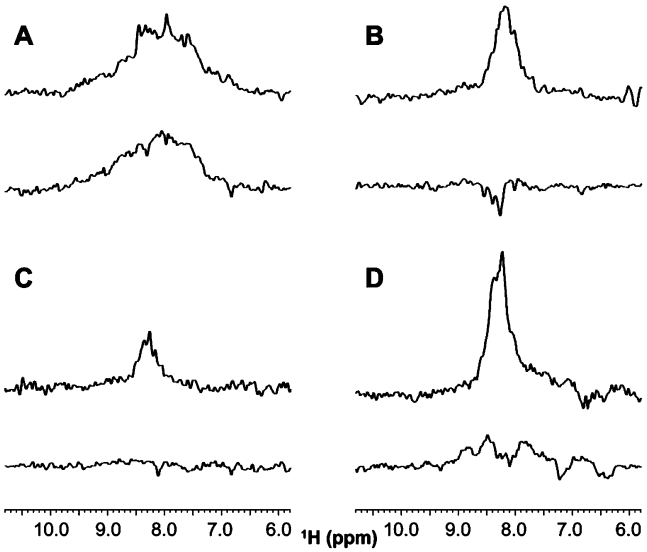

Figure 3. 1D HetNOE NMR spectra of folded and unfolded proteins. In each pair of spectra, the bottom spectrum is recorded with 3 s presaturation of amide protons, and the top spectrum is without proton presaturation. Spectra are recorded at optimized pH for the target and 20 °C. (A) HetNOE spectra of a typical folded protein (NESG Target ID BtR7 at pH 6.5) displaying amide proton chemical shift dispersion and relative similarity in intensities between spectra with and without saturation. (B) HetNOE spectra of a typical unfolded protein (NESG target VpR30 at pH 5.5) with the expected narrow range of amide proton chemical shifts and negative signal intensity when the spectrum is recorded following proton presaturation. (C−D) HetNOE spectra for two proteins, NESG targets (C) CaR26 and (D) SR128, with HSQC spectra characterized as poor/unfolded which nonetheless yielded structures from our crystallization pipeline. These spectra, recorded at 20 °C and a pH of 6.5, display a lack of chemical shift dispersion characteristic of unfolded proteins but display some weak positive signal in HetNOE spectra obtained with amide proton presaturation. These data are interpreted as indicating an equilibrium between unfolded and folded proteins or some partial folded character in the solution state.

An analysis of success in crystal structure determination, binned by the HSQC Spectral Qualities Scores, for 159 protein targets for which crystallization was attempted is presented in Figure 2A. Figure 2B shows rates of successful crystallization (i.e., the percent of targets for which crystallization efforts have yielded a crystal structure) in each HSQC spectral quality bin. The rate of crystallization success decreased slightly but not significantly with respect to increasing HSQC quality (Cochran-Armitage Trend Test $Z = -0.0837$, one-tailed $p = 0.47$); proteins exhibiting "Promising" spectra, with fewer than expected resonance peaks, exhibited the highest percentage of crystal structure successes (∼33%). Even the "Poor and/or Unfolded" class of proteins exhibited an ∼19% success rate in crystal structure determination. Furthermore, the Kolmogorov−Smirnov test fails to distinguish, at even a borderline level of significance, the HSQC quality distribution among targets yielding crystal structures from the distribution of HSQC quality evaluations of proteins which have not yielded crystal structures. The quality of the HSQC spectrum for a given protein target in our data set thus does not itself predict the crystallization success of a protein in any statistically measurable way.

In the set of proteins summarized in Figure 2, the NESG crystallization pipeline has yielded 16 crystal structures from proteins that exhibited "Poor and/or Unfolded" HSQC spectra in the NMR screening process. ^{1}H−^{15}N HetNOE measurements, which provide a more accurate assessment of the degree of structural order than 2D HSQC data alone,[31−33] revealed that some of the proteins in this HSQC class are indeed largely disordered under the conditions of the NMR screening. One-dimensional HetNOE data for two of proteins characterized as "Poor and/or Unfolded" and providing crystal structures (PDB IDs 1XX6 and 1RTY), shown in Figure 3, suggest that under the conditions of the NMR screen the backbone structures in these proteins exist in a dynamic equilibrium between fully folded and unfolded states (i.e., the ^{1}H−^{15}N HetNOE is approximately zero, rather than positive or negative) or an intermediate between the folded and unfolded states, e.g., a molten globule. Presumably, the crystallization conditions and/or the crystallization process itself can drive this equilibrium toward the folded state, allowing X-ray crystallographic analysis even for some proteins that are largely disordered in solution under conditions of NMR screening.

Discussion

Using data on 159 proteins produced, all produced and analyzed in a consistent way as part of a standardized sample production pipeline of the Northeast Structural Genomics Consortium, we observe a lack of correlation between HSQC spectral quality and success obtaining diffraction-quality crystals sufficient for structure determination by X-ray crystallography. This conclusion is completely consistent with the previously published smaller, limited studies by Savchenko et al.,[3] and Tyler et al.,[2] as well as the expanded study from the Toronto group on 264 proteins presented in the accompanying paper (Yee et al., J. Am. Chem. Soc., accompanying paper). Although these three studies were carried out by collaborating structural

proteomics centers, the processes of target selection, protein construct design, affinity-purification tags utilized, aggregation screening, crystallization screening, and HSQC quality scoring are different across these three studies, demonstrating that the results presented here are not specific to the particular process in place in the Rutgers/Columbia components of the NESG Consortium. The combined studies on over 420 purified proteins clearly demonstrate nonoverlapping success in obtaining good quality 2D HSQC spectra and crystallization success.

Even for smaller proteins, crystallography is sometimes able to provide structures where NMR spectral quality is insufficient to allow for structure determination by NMR. On the other hand, many targets that give excellent HSQC spectra and whose structures can be solved by NMR do not yield diffraction quality crystals. Targets which do not readily yield diffraction quality crystals but which yield excellent HSQC spectra and whose structures have subsequently been determined using NMR include such biologically important targets as ribosomal proteins,[34,35] proteins involved in iron−sulfur cluster synthesis,[36] and many enzymes. Currently, NMR-derived structures are available for 10 of the 18 targets yielding excellent HSQC spectra. Only three of these "Excellent" targets have yielded structures from the crystallography pipeline; of these three, two (HR1958 and XcR50) have provided atomic-resolution structures by both NMR and crystallography. NMR also provided structures for 3 of the 23 proteins with "Good" HSQC spectra that have not yielded structures via X-ray crystallography. An additional 30 proteins, produced at the Rutgers NESG Protein Production facility prior to initiating this side-by-side comparison of NMR data and success in crystal structure determination, have also provided 3D structures by NMR. Some of these exhibited significant aggregation in gel filtration chromatography and light scattering measurements and were deprioritized for crystallization analysis, while others were solved by NMR before SeMet-labeled samples were available for crystallization screening. As these protein samples were not analyzed by crystallization screening, they are not included in this study.

Considering the extensive crystallization screening process used by the HWRI and Columbia groups, it is unlikely that any more of the "Good"/"Excellent" samples from this series will yield structures via X-ray crystallography without considerable effort being put into construct optimization and/or crystallization. Some of these remaining "Good"/"Excellent" samples which have not yet been solved by X-ray crystallography were initially deprioritized for NMR data collection as they have molecular masses of 20−40 kDa, at the upper end of the range for routine NMR structure analysis with current NESG technologies. However, having attempted unsuccessfully to obtain diffraction quality crystals for these samples, some of these may be reprioritized for NMR structure analysis. Moreover, crystallization screening results could retrospectively suggest solution conditions which would improve the quality of the HSQC spectrum.

Addition of ligands in NMR screening buffers can improve a protein's HSQC spectrum and/or promote protein folding. Appropriate ligands could potentially improve spectral quality of some of the proteins providing poor HSQC. However, such ligand screening studies are not feasible for the 159 different protein samples used in this work. Nonetheless, a key conclusion of this paper is that there are many proteins whose structures can be studied by NMR, as evidenced by yielding good or excellent HSQC spectra using a small number of conditions (i.e., three pH values) and without added ligands, but which do not readily yield diffraction-quality crystals even when screened using hundreds of conditions, including addition of potential ligands.

On the face of it, these studies appear to contradict recent results by Page et al.,[1] showing that 1D NMR spectral quality is a predictor of crystallization. However, our study focuses not on the use of NMR as a predictor of crystallization but rather on the correlation between 2D HSQC spectra, indicating potential suitability for complete 3D structure analysis, and success in crystallization. It is well established that a well-resolved and nearly complete HSQC spectrum is prerequisite for NMR structure determination, while the correlations between 1D methyl dispersion and success in determining resonance assignments and 3D structures by NMR are not well established. On the other hand, 1D NMR screening has advantages over 2D HSQC, as it is very rapid and does not require isotope enrichment. The results of Page et al.[1] suggest that such 1D NMR spectra are predictive of crystallization success. Indeed, proteins that do not yield high-quality HSQC spectra may have well-defined structures and/or exhibit upfield-shifted methyl resonances. However, proteins with upfield-shifted methyl resonances will not necessarily provide 2D HSQC spectra of sufficient quality for 3D structure analysis by NMR.

A second significant difference between our study and that published by Page et al.[1] involves our process of Aggregation Screening to exclude polydisperse protein samples from the crystallization screening process. These proteins generally have broadened NMR resonances and yield poor quality 1D or 2D NMR spectra. In our target set, some 50% of protein samples (40% of prokaryotic proteins and 70% of eukaryotic proteins) are excluded from crystallization screening by the Aggregation Screening process because they are polydisperse and aggregated; inclusion of large numbers of aggregated proteins in the analysis reported by Page et al.[1] may contribute to their observed correlations between spectral quality and crystallization success.

Our aim in this work was to compare success in obtaining good-quality conventional HSQC data, an excellent predictor of success in full 3D structure analysis by NMR, with success in obtaining diffraction quality crystals. Another type of information that can be obtained by NMR involves quantitative and/or qualitative assessments of internal dynamics using nuclear relaxation studies and/or HetNOE data. Interestingly, even proteins exhibiting extensive internal dynamics under our standard NMR screening conditions, manifested in HetNOE data, sometimes will provide diffraction quality crystals in conditions used for crystallization screening. Although beyond the scope of this study, it is possible that under the lower

(34) Liu, G.; Xiao, R.; Parish, D.; Ma, L.; Sukumaran, D.; Acton, T. B.; Montelione, G.; Szyperski, T. Solution NMR Structure of Methanosarcina mazei Protein Rps24E: The Northeast Structural Genomics Consortium Target MaR11. Protein Data Bank id 1XN9.
(35) Yang, C.; Acton, T. B.; Shen, Y.; Ma, L.; Liu, G.; Xiao, R.; Montelione, G.; Szyperski, T. Solution NMR Structure of Methanococcus maripaludis Protein Mmp0443: The Northeast Structural Genomics Consortium Target MrR16. Protein Data Bank id 1YWX.
(36) Ramelot, T.; Cort, J.; Goldsmith-Fischman, S.; Kornhaber, G.; Xiao, R.; Shastry, R.; Acton, T. B.; Honig, B.; Montelione, G.; Kennedy, M. *J. Mol. Biol.* **2004**, *344*, 567−583.

solubility conditions used for crystallization these proteins are more ordered and then crystallize. Alternatively, the crystallization process may drive some proteins into folded conformations by mass action effects. It is also possible that there are indeed correlations between the amplitudes and extent of such internal dynamics and crystallization success. In any case, it is clear from this study that some proteins providing poor HSQC spectra (and even highly dynamic structures) under the limited conditions used for NMR screening may indeed provide diffraction quality crystals, while many proteins providing overall well-ordered structures and excellent HSQC spectra do not crystallize using a standard array of crystallization conditions.

Taken together, our results comparing NMR spectral quality to crystallization success suggest that NMR and crystallization have complementary roles to play in structural proteomics projects. Although demonstrated in the context of a specific structural proteomics sample preparation and screening process, these results also have important implications for structural biology in general. Unlike what some have long suspected, in the context of the platform for protein sample production used by the NESG Consortium, the majority of proteins with high quality HSQC do *not* rapidly and readily yield diffraction quality crystals. Indeed, even in the Wüthrich study,[1] only about 20% of proteins yielding "A" quality 1D ^{1}H NMR spectra, with well-dispersed methyl resonances, provided diffraction quality crystals and 3D structures by X-ray crystallography.[1] On the other hand, many proteins with even "Poor" quality HSQC spectra yield diffraction quality crystals and have provided crystal structures through the NESG crystallization pipeline. As it is relatively quick and easy to record 1D NMR spectra with the same samples used for 2D HSQC and 1D/2D HetNOE screening, both kinds of experiments should be recorded in future screening efforts, so that the correlations between these different spectroscopic probes can be assessed. NMR data can also be valuable in optimizing construct design, and in specific cases it is possible to use HSQC and other NMR spectral quality to identify and remove disordered regions of proteins, and thus to enhance crystallization success. Conversely, crystallization screening results could retrospectively suggest solution conditions that would improve the quality of the HSQC spectrum. However, the results of the extensive and careful study presented here, and supported by an independent study presented in the accompanying paper by Yee, et al. (J. Am. Chem. Soc., accompanying paper), clearly demonstrate that for a particular construct design X-ray crystallography and NMR often provide complementary sources of structural data. Both methods are required in order to optimize success for as many targets as possible in large-scale structural proteomics efforts.

Acknowledgment. This work was supported by NIH Protein Structure Initiative Grant P50 GM62413. The authors would like to thank Prof. C. Arrowsmith for sharing drafts of her related study, Drs. George DeTitta and Joe Luft for providing initial crystallization screening data, Jessica Lau, Michael Baran, and Michael Wilson for their assistance in collating the data used in this study, and Profs. T. Szyperski and M. Gerstein for useful discussions.

Supporting Information Available: Complete refs 2 and 17. Supplementary Table S1: NESG Targets Analyzed by Both Crystallization and NMR Screening. This material is available free of charge via the Internet at http://pubs.acs.org.

JA053564H