J Struct Funct Genomics (2010) 11:223–232
DOI 10.1007/s10969-010-9095-6

Fully automated high-quality NMR structure determination of small ^{2}H-enriched proteins

Yuefeng Tang · William M. Schneider · Yang Shen · Srivatsan Raman · Masayori Inouye · David Baker · Monica J. Roth · Gaetano T. Montelione

Received: 21 March 2010 / Accepted: 12 August 2010 / Published online: 24 August 2010

Abstract Determination of high-quality small protein structures by nuclear magnetic resonance (NMR) methods generally requires acquisition and analysis of an extensive set of structural constraints. The process generally demands extensive backbone and sidechain resonance assignments, and weeks or even months of data collection and interpretation. Here we demonstrate rapid and high-quality protein NMR structure generation using CS-Rosetta with a perdeuterated protein sample made at a significantly reduced cost using new bacterial culture condensation methods. Our strategy provides the basis for a high-throughput approach for routine, rapid, high-quality structure determination of small proteins. As an example, we demonstrate the determination of a high-quality 3D structure of a small 8 kDa protein, *E. coli* cold shock protein A (CspA), using <4 days of data collection and fully automated data analysis methods together with CS-Rosetta. The resulting CspA structure is highly converged and in excellent agreement with the published crystal structure, with a backbone RMSD value of 0.5 Å, an all atom RMSD value of 1.2 Å to the crystal structure for well-defined regions, and RMSD value of 1.1 Å to crystal structure for core, non-solvent exposed sidechain atoms. Cross validation of the structure with ^{15}N- and ^{13}C-edited NOESY data obtained with a perdeuterated ^{15}N, ^{13}C-enriched $^{13}CH_3$ methyl protonated CspA sample confirms that essentially all of these independently-interpreted NOE-based constraints are already satisfied in each of the 10 CS-Rosetta structures. By these criteria, the CS-Rosetta structure generated by fully automated analysis of data for a perdeuterated sample provides an accurate structure of CspA. This represents a general approach for rapid, automated structure determination of small proteins by NMR.

Keywords Cold shock protein · CS-Rosetta · Rapid NMR structure determination · Protein perdeuteration · cSPP expression system

Abbreviations

cSPP	Condensed-phase signal protein production
CspA	Cold shock protein A
PDB	Protein Data Bank
RMSD	Root-mean-square deviation
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid

Introduction

NMR spectroscopy is well suited for rapid and, in favorable cases, largely automated structure determination of

The authors Yuefeng Tang and William M. Schneider contributed equally to this work.

Y. Tang · G. T. Montelione (✉)
Department of Molecular Biology and Biochemistry, Center for Advanced Biotechnology and Medicine, Northeast Structural Genomics Consortium, Rutgers University,
Piscataway, NJ 08854, USA
e-mail: guy@cabm.rutgers.edu

W. M. Schneider · M. Inouye · M. J. Roth · G. T. Montelione
Department of Biochemistry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey,
Piscataway, NJ 08854, USA

Y. Shen
Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA

S. Raman · D. Baker
Department of Biochemistry, University of Washington,
Seattle, WA 98195, USA

small (<125 residues) proteins [22, 36]. While backbone assignments for such proteins are routinely obtained in a largely automated fashion [1, 17, 23, 39], assignment of sidechain resonances can often be a bottleneck for the process of structure determination. Automated sidechain assignment methods are, however, evolving and beginning to have an important impact on the field [1, 13]. Recently, we described an approach for solving protein NMR structures using Rosetta conformational energy calculations together with only a limited amount of experimental NMR data, including backbone resonance assignments, residual dipolar coupling data, and some manually-assigned long-range backbone-backbone NOEs [26]. This approach was demonstrated to provide accurate backbone structures (chain folds) for proteins of up to 25 kDa, with reasonably accurate core sidechain packing.

Several years ago, we described a strategy for rapid *automatic* determination of small (<100 residue) protein structures using only the sparse constraints that can be obtained using a perdeuterated protein [38]. Our strategy for rapid fold determination derives from ideas that were originally introduced for determining NMR structures of larger proteins [9, 10, 12], using [^{2}H,^{13}C,^{15}N]-enriched protein samples with protonated sidechain methyl groups ($^{13}CH_3$). Data collection includes acquiring NMR spectra for determining assignments of backbone and sidechain ^{15}N, H^N resonances, and sidechain $^{13}CH_3$ methyl resonances. Backbone resonance assignments and NOESY cross peaks are then determined automatically, and 3D structures generated using CNS [5, 20]. This strategy provides reliable backbone chain folds for small (<100 residue) proteins, which are useful for certain applications, and good starting points for further refinement to high precision and accuracy using additional NMR data.

This "sparse constraint" approach exploits the fact that perdeuteration generally improves spectral quality and interpretability even of smaller proteins. Although protein deuteration is not generally required for small protein structure determination, it is valuable for improving sensitivity of many amide or methyl proton-detected heteronuclear NMR experiments [2, 14] even for proteins in the 7–12 kDa range. As the gyromagnetic ratio of the ^{2}H is $\sim$6.5 fold smaller than that of ^{1}H, the dipolar interaction between ^{13}C or ^{15}N and the directly bound proton spin is greatly reduced. Therefore the transverse relaxation times T_2 of ^{13}C and ^{15}N nuclei are increased, providing sharper linewidths and higher signal-to-noise ratios (S/N). Constant-time NMR experiments which may have poor S/N with fully protonated proteins can be recorded with higher sensitivity due to the reduced transverse relaxation rates of ^{13}C and ^{15}N obtained for perdeuterated proteins. We also observe better performance of automated resonance assignment software for backbone resonance assignments (e.g. AutoAssign [39]) because of the improved resolution and sensitivity of amide H^N-detected triple resonance experiments on the perdeuterated protein samples. Another advantage of longer transverse relaxation times and the reduction in spin-diffusion pathways is that it permits the detection of weaker NOEs that may not otherwise be observed when longer NOESY mixing times are used. Some poor NMR signals resulting from exchange broadening and limited protein solubility can also be improved by perdeuteration. These advantages of deuterium incorporation are well-known for studies of larger (15–50 kDa) proteins, but also provide improved performance and improved S/N for smaller sized (<70–100 residues) proteins.

While the idea of rapid, fully automated structure determination of small perdeuterated proteins is attractive and innovative, two drawbacks have hindered the routine application of this method for high-throughput NMR protein structure determination. First, producing perdeuterated proteins by conventional expression methods is expensive, and secondly, only backbone chain folds are reliably determined using sparse constraints and CNS refinement [38]; the details of the resulting structures are not particularly good.

Here, we combine the fully automated sparse constraint approach for small proteins, first outlined by Zheng et al. [38], with two recent innovations. First, we have adopted recently developed condensed-phase single protein production (cSPP) methods [29, 33–35] to allow bacterial expression in 10 to 40-fold condensed-phase fermentations without reduction in protein expression per cell, allowing significantly less expensive production of ^{2}H, ^{13}C, ^{15}N-enriched proteins. In the cSPP system, MazF, an mRNA interferase functioning as an ACA-specific endoribonuclease, is co-expressed with the target protein. The expression of MazF eliminates almost all cellular mRNAs containing ACA sequences. The target gene is selectively expressed by engineering it to contain no ACA sequences, without altering the amino acid sequence of the protein encoded by the resulting mRNA. Secondly, we replace CNS refinement with the recently introduced CS-Rosetta method [31] for small protein structure analysis. The CS-Rosetta program provides a powerful approach for NMR structure determination of small proteins using only ^{1}H, ^{13}C, and ^{15}N backbone and $^{13}C^{\beta}$ resonance assignments [31]. Exploiting these recent innovations, we have extended the approach originally described by Zheng et al. [38] to demonstrate, using a ^{2}H, ^{15}N, ^{13}C-enriched sample of the 86-residue *E. coli* cold shock protein (CspA) as an example of a general process for determining accurate small protein structures requiring only a few days of NMR data collection, a specific data collection protocol, and largely automated data analysis.

Methods and materials

Preparation of [^{1}H-^{13}C]-I(δ1)LV, ^{13}C, ^{15}N, ^{2}H—CspA for structural studies

Competent *E. coli* BL21(DE3) cells containing the pACYC*mazF* [35] plasmid were transformed with pColdI(SP-4) [33] plasmid (Takara Bioscience, Inc) containing ACA-less *cspA* gene. The resulting constructs include a 16-residue N-terminal tag, consisting of a translation enhancing element (TEE), a His$_6$ tag, and a Factor Xa cleavage site. Protein expression was performed essentially as described by Schneider et al. [29], with the following details: single colonies were selected and used to inoculate 2.5 ml LB medium at 37°C for 6 h. 2 ml of the LB culture was inoculated into 100 ml of MJ9 minimal medium at 37°C overnight. When OD$_{600}$ reached 1.8–2.0 units, the culture was centrifuged at 3,000 × *g* for 15 min at 4°C. The cell pellet was resuspended in 1 l of fresh MJ9 medium and cells were grown at 37°C until OD$_{600}$ reached 0.5. At this point the culture was chilled on ice for 5 min and shifted to 15°C for 45 min to acclimate the cells to cold shock conditions. Target protein (CspA) was then expressed along with MazF for 1.5 h by addition of 1 mM isopropyl-β-D-thiogalactoside (IPTG) prior to expression in isotope enriched medium. Cultures were then centrifuged at 3,000 × *g* for 15 min at 4°C, resuspended in 2.5% volume (40× condensed) in deuterated (^{2}H$_2$O) wash solution [7.0 g/l Na$_2$HPO$_4$; 3.0 g/l KH$_2$PO$_4$; 0.5 g/l NaCl; pH 7.4], centrifuged again, and resuspended in 25 ml of deuterated MJ9 minimal medium containing 1 g/l ^{15}NH$_4$Cl; 4 g/l ^{13}C, ^{2}H-glucose; 50 mg/l α-^{13}C-ketobutyric acid; 100 mg/l α-^{13}C-ketoisovaleric acid; and 1 mM IPTG. Protein expression continued at 15°C for 24 h. Cells were harvested by centrifugation as described above and cell pellets were stored at −80°C. All isotopes were purchased from Cambridge Isotope Laboratories.

CspA purification and concentration

Cell pellets were resuspended in 40 ml of lysis buffer [50 mM Na$_2$HPO$_4$-NaH$_2$PO$_4$; 300 mM NaCl; 5 mM imidazole; 5 mM 2-mercaptoethanol; with 1 EDTA-free protease inhibitor tablet (Roche Cat. # 11 873 580 001) per 50 ml at pH 8.0] and sonicated to lyse the cells. Lysed cells were then centrifuged at 4°C for 1 h at 16,000 rpm in a Sorvall SS-34 rotor. The protein was further purified by binding to Ni–NTA agarose at 40 ml of soluble extract per 1 ml of bed resin at 4°C overnight. Resin was washed twice with 25 ml of Wash Buffer [50 mM Na$_2$HPO$_4$-NaH$_2$PO$_4$; 300 mM NaCl; 25 mM imidazole; 5 mM 2-mercaptoethanol, pH 8.0], and protein was eluted in 8 ml of Elution Buffer [50 mM Na$_2$HPO$_4$-NaH$_2$PO$_4$; 300 mM NaCl; 300 mM imidazole; 5 mM 2-mercaptoethanol, pH 8.0]. The protein solution was then dialyzed overnight at 4°C into NMR Buffer containing 50 mM KH$_2$PO$_4$, 1 mM NaN$_3$, pH 6.0, containing 10% ^{2}H$_2$O, and concentrated to a final concentration of ~0.2 mM.

NMR spectroscopy

Backbone resonance assignments for [^{1}H-^{13}C]-I(δ1)LV, ^{13}C, ^{15}N, ^{2}H-enriched CspA were determined using conventional triple resonance NMR experiments [37] including HNCO and deuterium-decoupled pulse sequences HN(ca)CO; HNCA; HN(co)CA; HNCACB and HN(co)-CACB. The carrier positions were set to 118.0 ppm for ^{15}N, 176 ppm for ^{13}CO, 54 ppm for ^{13}C$^{\alpha}$ and 39 ppm for ^{13}C$^{\alpha}$/^{13}C$^{\beta}$. Key parameters of data collection are summarized in Table 1. The total data collection time for all of these triple resonance experiments was about 3.5 days.

In addition, 3D ^{13}C-edited NOESY (mixing time of 350 ms) and ^{15}N-edited NOESY (mixing time of 175 ms) were collected on a 600 MHz Bruker spectrometer with TXI probe. The matrix sizes of these spectra were

Table 1 800 MHz triple resonance data used for determining backbone resonance assignments

	^{15}N-HSQC	HNcoCA	HNCO	HNCA	HNCACB	HNcoCACB	HNcaCO
No. of points							
Collected	1024, 256	1024, 40, 50	1024, 40, 40	1024, 40, 50	1024, 64, 100	1024, 64, 100	1024, 40, 40
After LP	1024, 512	1024, 72, 82	1024, 72, 72	1024, 72, 82	1024, 96, 164	1024, 96, 164	1024, 72, 72
After zero filling	1024, 512	1024, 128, 128	1024, 128, 128	1024, 128, 128	1024, 128, 256	1024, 128, 256	1024, 128, 128
No. of scans	8	4	4	4	16	16	16
Spectral width (ω_1, ω_2, ω_3; ppm)	14, 28	14, 23, 32	14, 23, 24	14, 23, 32	14, 28, 72	14, 28, 72	14, 23, 24
Recycle delay (s)	1	1	1	1	1	1	1
Collection time (h)	0.6	2.2	2.0	2.2	33.2	33.6	8.6

1,024 × 32 × 220 total data points for ^{13}C-edited NOESY, and 1,024 × 64 × 256 total data points for ^{15}N-edited NOESY. For ^{13}C-edited NOESY, the spectrum widths in ^{1}H, ^{13}C and indirect detected ^{1}H dimensions were set to 14, 16 and 12 ppm respectively and the carrier positions were set 4.7 ppm for ^{1}H and 16 ppm for ^{13}C dimension. For ^{15}N-edited NOESY, the spectrum widths in ^{1}H, ^{15}N and indirect detected ^{1}H dimensions were set to 14, 28 and 11.5 ppm respectively and the carrier positions were set 4.7 ppm for ^{1}H and 118 ppm for ^{15}N dimensions. The total data collection time for ^{13}C-edited and ^{15}N-edited NOESY spectra was approximately 2.5 days.

In all NMR experiments, FIDs were processed with linear prediction and zero filling, and weighted by sine bell function in all direct and indirect detected dimensions. All NMR spectra were processed and examined with NMRPipe and NMRDraw software packages [7]. The program SPARKY [11] was used for data visualization and analysis. Chemical shifts of proton were referenced to external DSS. ^{13}C and ^{15}N chemical shifts were referenced indirectly based on the proton referencing.

Analysis of resonance assignments

AutoAssign [23] software was used for automated analysis of backbone and side chain ^{13}C$^{\beta}$ resonance assignments for CspA. Peak list of [^{15}N-^{1}H^{N}]-HSQC, and peak lists from the triple resonance experiments, including 3D HNCO; HN(ca)CO; HNCA; HN(co)CA; HNCACB and HN(co)-CACB, were peak picked automatically using the 'restrictive peak picking' function of the SPARKY [11] software; in order to improve the performance of AutoAssign for backbone assignments, these peak lists were manually refined prior to input into AutoAssign [23, 39] for automated analysis of backbone resonance assignments. Cleaning up the peak lists only required 2–3 h. Sidechain ^{13}C and ^{1}H methyl resonances of Leu, Val and Ile (δ1) were determined subsequently by interactive spectral analysis using [^{13}C–^{1}H]-HSQC, 3D ^{13}C-edited NOESY, and 3D ^{15}N-edited NOESY spectra. These methyl sidechain assignments were used in the "conventional 3D structure calculations", but not in the CS-Rosetta calculations.

Sparse-constraint 3D structure calculations

Sparse-constraint 3D structure calculations were performed using the AutoStructure [15, 16] software ver. 2.2.1-CND for automated analysis of NOESY cross peak assignments, implemented together with the program CYANA ver. 2.1 for structure generation. The input for AutoStructure analysis consisted of (1) a list of backbone and ^{13}C-^{1}H methyl sidechain assignments; (2) manually edited NOESY peak lists, including chemical shift and peak heights, generated from ^{13}C-edited and ^{15}N-edited NOESY spectra; (3) sites of slowly exchanging amide hydrogens based on published amide ^{1}H/^{2}H exchange data for CspA [8, 24]; (4) broad ϕ, ψ angle constraints (±40° and ±50°, respectively) derived from chemical shift data (after correction of ^{2}H isotope-shift effect) using the program TALOS [6]. The best 10 of 56 structures (lowest energy) from the final cycle of AutoStructure were refined by restrained molecular dynamics in an explicit water bath using CNS 1.1 [5, 20].

Chemical-shift based protein structure prediction by ROSETTA (CS-ROSETTA)

Chemical shift information, including backbone ^{13}C$^{\alpha}$, ^{15}N, ^{13}C', ^{1}H^{N} and sidechain ^{13}C$^{\beta}$ assignments, were used as input for CS-ROSETTA. Details of the process of generating the CS-ROSETTA protein structure are described in Shen et al. [31] Three key steps are involved. First, based on the chemical shift values (which did not include backbone ^{1}H$^{\alpha}$ shifts) and protein sequences, peptide fragments were selected from a protein structure database using the MFR module [7, 18] of the NMRPipe software package. All proteins with PSI-BLAST E-val score <0.05 with *E. coli* CspA were removed from the database. Second, a standard ROSETTA [27] protocol was used for *de novo* structure generation. Third, ROSETTA all-atom models resulting from the above procedure were evaluated based on how well backbone chemical shifts predicted for the models using SPARTA [30] agree with the experimental chemical shifts. If the lowest energy models cluster within less than ~2 Å from the model with the lowest energy, the structure prediction is considered successful and lowest energy models are converged. A total of 10,000 all-atom Rosetta models were generated from the MFR-selected peptide fragments, using a cluster of 20 CPUs. These CS-Rosetta runs required approximate 3 days. The 1,000 lowest-energy models were chosen and their all-atom ROSETTA energies were recalculated in terms of the fitness with respect to the experimental chemical shift values. The lowest energy models are converged based on the fact that C$^{\alpha}$ RMSD values are less than ~2 Å relative the lowest energy model. 10 lowest energy models were selected as a representation of the 3D structure of CspA. The CS-ROSETTA package used in this work may be downloaded from http://spin.niddk.nih.gov/bax/software/CSROSETTA/indes.html.

Structure quality assessment

Global structure quality factors for the ensemble of CspA structures generated using sparse NMR constraints with conventional data analysis methods or by CS-Rosetta were determined using Protein Structure Validation Suite (PSVS) software package [3]. The output of the PSVS includes raw scores and normalized statistical *Z*-scores [3] for metrics assessed by the Verify 3D [4], Prosa II [32], PROCHECK [19], and MolProbity [21] software packages.

Results

Rapid resonance assignments with perdeuterated CspA sample prepared by cSPP

An 86-residue construct of *E. coli* CspA was produced in 40-fold condensed ^{2}H, ^{13}C, ^{15}N-enriched media using the cSPP system [29]. A 0.2 mM sample of ILV-perdeuterated CspA was used for collection of deuterium-decoupled triple resonance experiments. The complete list of the experiments includes HNCO, HN(ca)CO, HNCA, HN(co)CA, HNCACB, and HN(co)CACB experiments, collected over a period of 3.5 days at 800 MHz. The experiments executed and corresponding key parameters of the data collation are summarized in Table 1. Following automatic peak picking and some manual refinement of the peak lists with the SPARKY program [11], the program AutoAssign [39] was used for automatic analysis of backbone H^{N}, ^{15}N, $^{13}C^{\alpha}$, ^{13}C’, and sidechain $^{13}C^{\beta}$ resonance assignments. The resulting $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ connectivity map, documented in Fig. 1, is essentially complete, indicating high reliability of the assignments. The reliability of the automatically-determined backbone resonance assignments was subsequently validated by manual analysis of these same data by interactive spectral analysis with SPARKY [11]. Except for the N-terminal tag (in the region of 6 consecutive His residues), a complete set of backbone and $^{13}C^{\beta}$ assignments were obtained for CspA. The automated backbone resonance assignments are consistent with the published assignments for CspA (BMRB accession number 4296), which have been validated by self-consistent analysis of NOESY data and 3D structure calculations. Perdeuterated protein samples produced with the cSPP system thus provide high-quality NMR spectra suitable for rapid automated analysis of backbone resonance assignments.

Protein structure determination using sparse NMR constraints

As a further example of the novel utility of such perdeuterated samples produced with the cSPP system, we next demonstrated rapid analysis of the 3D structure of [^{1}H-^{13}C]-I(δ1)LV, ^{13}C, ^{15}N, ^{2}H-enriched CspA using conventional sparse NOESY-based methods. Additional 3D ^{15}N-edited NOESY and 3D ^{13}C-edited NOESY data were acquired and used to assign side-chain methyl resonances, and NOESY cross peaks were assigned in order to determine the 3D structure by conventional automated

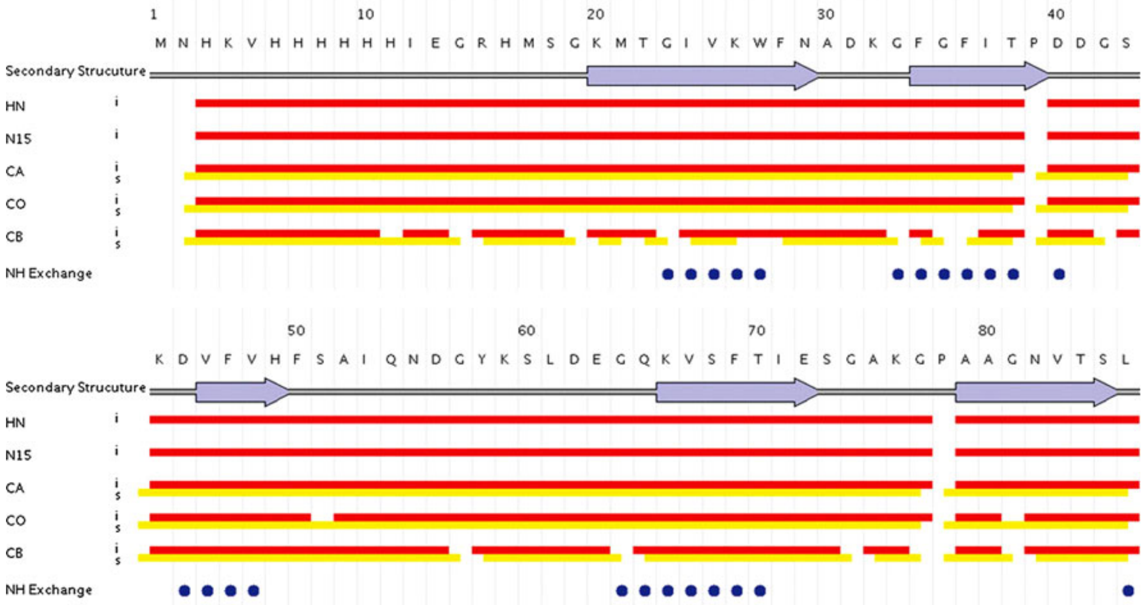

Fig. 1 Summary of backbone and $^{13}C^{\beta}$ resonances assignments for CspA derived from triple resonance NMR experiments. Red bars and yellow bars underneath the amino acid sequence represent the connectivity established between intra and sequential residues respectively. These data were obtained by analyzing six 2D and 3D NMR spectra, summarized in Table 1. Slowly exchanging backbone amides, used in the conventional structure analysis but not in the CS-Rosetta analysis, identified by ^{1}H/^{2}H exchange measurements, are represented by *filled circles*. Secondary structures of the β-barrel found in the final structure are indicated by *arrows* along the amino acid sequence

methods with energy refinement. AutoStructure [15, 16] was used to determine NOESY cross peak assignments and to generate distance constraints, structure generation calculations were carried out using these constraints as input to CYANA, and CNS refinement was done by restrained molecular dynamics in explicit water [5, 20]. Table 2 summarizes the NOE-based distance constraints, hydrogen bonds, and dihedral angle constraints, identified by AutoStructure. AutoStructure identified a total of 131 distance constraints, including 61 long-range constraints. Based on characteristic NOE-based contact patterns and slow amide hydrogen/deuterium (H/D) exchange data, AutoStructure also identified a total of 22 hydrogen bond upper/lower constraints (11 hydrogen bonds); 20 of which are long-range hydrogen bond constraints. Identification of hydrogen bonds by AutoStructure is critical for proper registration of β-strands and folding β-sheet structures derived from sparse constraint data. In each of these calculations, 56 structures were generated from extended conformations, and 10 with lowest values of the target function were selected to represent the solution NMR structure of CspA. The resulting ensembles of these sparse-constraint CspA structures, and comparison with the crystal structure (PDB ID: 1mjc) [28], are shown in Fig. 2. In the remainder of the text, we use PDB id 1mjc to designate the crystal structure of CspA. Structural statistics of the minimal-constraint structures are also listed in Table 2. These structures exhibit good structural convergence and few residual constraint violations. The averaged backbone RMSD in the ordered residue regions is 1.2 Å. For the well-defined core residues, the averaged backbone RMSD relative to 1mjc is ~1.6 Å. These results show that the backbone structure generated with these sparse constraint automated analysis methods can be reasonably accurate, as discussed in detail by Zheng et al. [38].

CS-Rosetta structure generation for perdeuterated CspA

The recently introduced CS-Rosetta method [31] provides an alternative approach for small protein structure analysis using *only backbone and* $^{13}C^{\beta}$ *chemical shift data*. CS-Rosetta calculations were carried out using these resonance assignments determined with <4 days of data collection using the perdeuterated CspA sample, produced with the cSPP system [29]. In this work, we tested the performance of CS-Rosetta with and without hydrogen–deuterium isotope chemical shift corrections on ^{13}C chemical shift values. The isotope chemical shift corrections for backbone ^{15}N and $^{13}C^{\alpha}$ nuclei and sidechain $^{13}C^{\alpha}$ nuclei were made using values proposed by Gardner et al. [10]. In our experience, the isotope chemical shift corrections did not impact the quality of the resulting CS-Rosetta structure. The resulting ensemble of 10 structures generated using no isotope shifts correction, shown in Fig. 3, exhibits excellent structure quality scores (Table 2). The CS-Rosetta structure is also in excellent agreement with the 1mjc crystal structure [28], with backbone RMSD of 0.5 Å, all atom RMSD of 1.2 Å for well-converged regions, and 1.1 Å RMSD for core, non-solvent-exposed sidechain atoms, relative to 1mjc. Additional key structural statistics for the CS-Rosetta structure are listed in Table 2. Also included in Table 2 are structural statistics for the conventional NMR structure of CspA (PDB ID: 3mef) determined several years ago with extensive analysis of sidechain assignments and NOEs [8]. In the remainder of the text, we use 3mef to designate the conventionally-determined NMR structure with full sidechain assignment. The Ramachandran statistics and global quality scores for CS-Rosetta structure are significantly better than those for the 3 mef or for the sparse-constraint structure.

A comparison of the CS-Rosetta structure of Fig. 3 with the NOESY constraint list used to generate the sparse-constraint NMR structure of Fig. 2 (i.e. the data obtained for ^{2}H, ^{15}N, ^{13}C-enriched $^{13}CH_3$ methyl protonated CspA) is also summarized statistically in Table 2. This analysis reveals only a few distance violations >0.5 Å (the largest being 1.7 Å) across the ensemble of 10 CS-Rosetta structures, cross-validating the high accuracy of the CS-Rosetta structure. Comparison with the more extensive NOESY constraint list used to determine the 3mef [8] reveals some additional constraint violations by the CspA structure; however this work was performed using a different CspA construct, and the overall structure quality scores (Table 2) for this published “full blown” NMR structure 3mef are much poorer than either the CS-Rosetta structure or the 1mjc. Indeed, structure quality scores for the published NMR structure (Table 2), particularly the Procheck (all dihedral) and Molprobity Clash scores, are well below the threshold ($Z = -5$) considered to be acceptable for a good quality NMR structure [3]. Based on its closer agreement with 1mjc, particularly for core sidechain atom positions, and better overall structure quality scores, it appears that the CS-Rosetta NMR structure of CspA (Fig. 3) is in fact more accurate than the previously published “full blown” NMR structure 3mef [8].

Discussion

Our results demonstrate a general, rapid, and simple approach for determining high quality 3D structures of small (<10 kDa) proteins, in fully automated fashion, with accuracies rivaling structures determined using more extensive NMR methods. In particular, the core sidechain packing, determined by the Rosetta potential energy function, is quite accurate based on comparison with the

Table 2 Summary of structural statistics for *E. coli* CspA NMR structures

	Sparse-constraint NMR structure[a]	3mef[b]	Sparse-constraint CS-Rosetta structure[c]
Conformationally-restricting constraints[d]			
Distance constraints			
Total	131		131
Intra-residue ($i = j$)	17		17
Sequential ($\lvert i - j \rvert = 1$)	45		45
Medium range ($1 < \lvert i - j \rvert \leq 5$)	8		8
Long range ($\lvert i - j \rvert > 5$)	61		61
Distance constraints per residue	2.0		2.0
Dihedral angle constraints	68		68
Hydrogen bond constraints			
Total	22		22
Long range ($\lvert i - j \rvert > 5$)	20		20
Number of constraints per residue	3.3		3.3
Number of long range constraints per residue	1.2		1.2
Residual constraint violations[d]			
Average number of distance violations per structure			
0.1–0.2 Å	1.4		0.9
0.2–0.5 Å	0		1.9
>0.5 Å	0		3.7
Average RMS distance violation/constraint (Å)	0.02		0.17
Maximum distance violation (Å)	0.18		1.74
Average number of dihedral angle violations per residue			
1–10°	3.6		3
>10°	0		0.8
Average RMS dihedral angle violation/constraint (°)	0.45		1.73
Maximum dihedral angle violation (°)	3.4		16.70
RMSD from average coordinates (Å)[d,e]			
Backbone atoms	1.2 ± 0.2	0.5 ± 0.1	0.8 ± 0.2
Heavy atoms	1.7 ± 0.2	1.1 ± 0.1	1.2 ± 0.2
RMSD from X-ray structure (Å)[d,f]			
Backbone atoms	1.58 ± 0.38	0.95 ± 0.11	0.52 ± 0.12
Heavy atoms	2.24 ± 0.34	1.63 ± 0.16	1.17 ± 0.11
Sidechain RMSD from X-ray structure (Å)[d,g]			
Heavy atoms	1.75 ± 0.20	1.59 ± 0.15	0.86 ± 0.11
Heavy sidechain atoms	1.81 ± 0.23	1.93 ± 0.22	1.14 ± 0.12
Ramachandran statistics[d,e]			
Most favored regions (%)	92.0	78.3	93.7
Additional allowed regions (%)	8.0	21.7	6.3
Generously allowed (%)	0.0	0.0	0.0
Disallowed regions (%)	0.0	0.0	0.0

Global quality Scores[d]	Raw/Z-score	Raw/Z-score	Raw/Z-score
Verify 3D	0.33/−2.09	0.43/−0.48	0.45/−0.16
ProsaII	0.61/−0.17	0.77/0.50	0.85/0.83
Procheck (phi-psi)	−0.49/−1.61	−1.37/−5.07	−0.28/−0.79

Table 2 continued

Global quality Scores[d]	Raw/Z-score	Raw/Z-score	Raw/Z-score
Procheck (all dihedrals)	−0.42/−2.48	−1.47/−8.69	0.00/0.00
Molprobity clash score	15.22/-1.09	64.74/-9.58	5.58/0.57

Analysis for residues 1–70, excluding disordered N-terminal expression tag

[a] Structure obtained from sparse NMR constraints

[b] NMR structure determined by conventional methods (PDB id 3mef)

[c] Structure obtained from CS-Rosetta structure generation, compared with constraints; note that these distance constraints were not used in generating the CS-Rosetta structure

[d] Generated using PSVS 1.3 program. Average distance violations were calculated using the sum over r^{-6}. Note that the conformational constraints were not used in CS-Rosetta calculations except to validate the structure by providing the statistics listed in this table

[e] Order residue ranges [S(phi) + S(psi) > 1.8]. NMR structure using minimum constraints: 4–24, 30–33, 35–36, 45–46, 51–55, 63–64, 67–69; Conventionally-determined NMR structure: 4–10, 20–23, 30–32, 48–51, 53–54, 68–69; CS-Rosetta generated structure: 4–27, 29–37, 40–60, 62–66

[f] Well-defined core region: 5–9, 19–22, 50–56, 63–69

[g] Buried hydrophobic residues: V9, I21, V30, V32, I37, L45, V51, F53, A64, V67

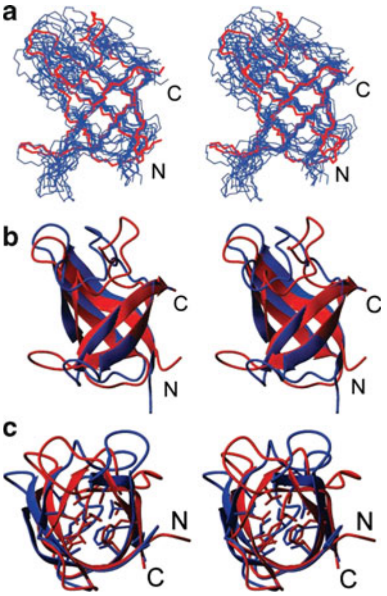

Fig. 2 Stereoview of the superimposition of AutoStructure-CNS structure for [^{1}H-^{13}C]-I(δ1)LV, ^{13}C, ^{15}N, ^{2}H-enriched CspA determined by conventional automated analysis methods (*blue*) with 1mjc (*red*). **a** Backbone line representations of the 10 lowest energy conformers obtained from AutoStructure-CNS structure compared with 1mjc. **b** Ribbon diagram of the lowest energy conformer of AutoStructure-CNS structure versus 1mjc. **c** The packing of the hydrophobic residues (viz, V9, I21, V30, V32, I37, L45, V51, F53, A64, and V67) for the lowest energy conformer of AutoStructure-CNS structure versus 1mjc. The disordered N-terminal hexaHis expression tag is excluded from the analysis

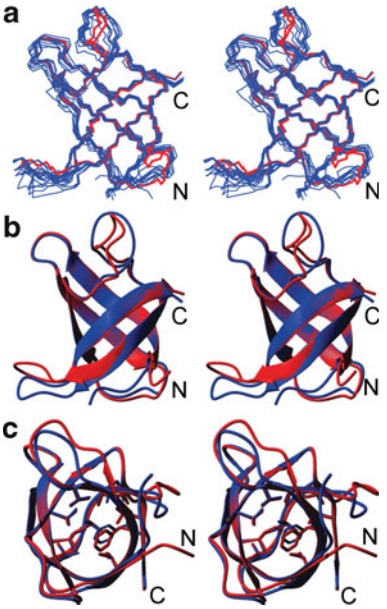

Fig. 3 Stereoview of the superimposition of the CS-Rosetta structure for ^{2}H,^{13}C,^{15}N-enriched CspA (*blue*) with 1mjc (*red*). **a** Backbone line representations of the 10 lowest energy conformers obtained from CS-Rosetta structure compared with 1mjc. **b** Ribbon diagram of the lowest energy conformer of CS-Rosetta structure versus 1mjc. **c** The packing of the core hydrophobic residues (viz, V9, I21, V30, V32, I37, L45, V51, F53, A64, and V67) for the lowest energy conformer of CS-Rosetta structure versus 1mjc. The disordered N-terminal expression tag is excluded from the analysis

crystal structure, despite the fact that no sidechain constraints are used in these calculations. Similar results were observed in CS-RDC-Rosetta calculations with larger proteins [26].

The time spent on CS-Rosetta runs depends on the number of Rosetta models generated and the number of CPUs used for the CS-Rosetta structural generation. In our study, we generate 10,000 models initially and we use 20 CPUs for the calculation. The process takes about 3 days. The time saved for structure determination using our proposed methods relative to conventional methods includes the time required for collection of spectra required for determining side-chain assignments and NOESYs, time required to process and analyze these spectra, as well as the time required for structure calculations and refinement which are the time-limiting steps for NMR structure determination. Our proposed approach only requires triple resonance NMR experiments for backbone assignments followed by automated analysis of backbone resonance assignments. Once most of the backbone resonance assignments are determined, these chemical shift data are submitted to CS-Rosetta. This approach, which is largely automated, not only saves time in data collection and analysis, but can generate a high-quality protein structure.

NOESY data and protein ILV methyl protonation are not required in the strategy proposed in this paper for small protein structure determination. NOESY data on the ILV-labeled sample was only used for cross validation of the CS-Rosetta structure. However, CS-Rosetta calculations do not always converge, even for small protein structures, and NOESY data for the perdeuterated ILV methyl protonated protein sample can be used if necessary together with CS-Rosetta if the chemical shift data alone do not provide a converged structure.

Our work further demonstrates that ^{2}H, ^{13}C,^{15}N-enriched protein samples made by the cSPP system at a drastically reduced cost and purified with a single-step Ni–NTA affinity chromatography, allow data collection and automated analysis of backbone ^{1}H^{N}, ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$'$, as well as sidechain ^{13}C$^{\beta}$, assignments in only a few days. In related work, we have demonstrated the combined use of CS-Rosetta and automated NOESY analysis to provide more accurate NOESY cross peak assignments, beginning with extensive backbone and sidechain assignments [25], and the use of CS-RDC-Rosetta with manually assigned NOESY-based constraints to generate good quality structures of larger (10–25 kDa proteins) [26]. The present study is the first example of applying CS-Rosetta for rapid fully-automated NMR structure determination of small proteins, a unique application that provides a new and general approach for obtaining 3D structures of small proteins. The CspA structure obtained rivals the best NMR structures available to date for CspA using conventional methods, even those utilizing extensive sidechain proton assignments [8]. This approach has tremendous value in preparing protein samples and generating assignments and structural information for small molecule screening studies, as well as in high-throughput structural and functional genomics studies.

Acknowledgments We thank Drs. Ad Bax, Rajeswari Mani, and Rong Xiao for helpful discussions and comments on the manuscript. The CS-Rosetta structure of CspA has been submitted to the Protein Data Bank as entry 2L15. This work was supported by the National Institutes of General Medical Science Protein Structure Initiative program, grants U54 GM074958 (to G.T.M and M.I.) and U54 GM75026 (to G.T.M. and M.I.), and by grant RO1 GM070837 and GM088808 (to M.J.R.). W.M.S. was supported by NIH training grants T32 GM08360 and T32 A1007403. This work was also supported by the Intramural Research Program of the National Institutes of Health National Institute of Diabetes and Digestive and Kidney Diseases (to Y.S.).

References

1. 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
2. Bax A, Grzesiek S (1993) Methodological advances in protein NMR. Acc Chem Res 26:131–138
3. Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. Proteins 66:778–795
4. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170
5. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS et al (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921
6. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302
7. 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
8. Feng W, Tejero R, Zimmerman DE, Inouye M, Montelione GT (1998) Solution NMR structure and backbone dynamics of the major cold-shock protein (CspA) from *Escherichia coli*: evidence for conformational dynamics in the single-stranded RNA-binding site. Biochemistry 37:10881–10896
9. Gardner KH, Kay LE (1997) Production and incorporation of ^{15}N, ^{13}C, ^{2}H (^{1}H-δ1 Methyl) isoleucine into proteins for multidimensional NMR studies. J Am Chem Soc 119:7599–7600
10. Gardner KH, Rosen MK, Kay LE (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. Biochemistry 36:1389–1401
11. Goddard TD, Kneller DG (2008) SPARKY 3. University of California, San Francisco
12. Goto NK, Kay LE (2000) New developments in isotope labeling strategies for protein solution NMR spectroscopy. Curr Opin Struct Biol 10:585–592
13. Grishaev A, Steren CA, Wu B, Pineda-Lucena A, Arrowsmith C, Llinas M (2005) ABACUS, a direct method for protein NMR structure computation via assembly of fragments. Proteins 61:36–43
14. Grzesiek S, Anglister J, Ren H, Bax A (1993) ^{13}C line narrowing by ^{2}H decoupling in ^{2}H/^{13}C/^{15}N enriched proteins. Application to triple resonance 4D J connectivity of sequential amides. J Am Chem Soc 115:4369–4370
15. 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

16. Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins 62:587–603
17. Jung YS, Zweckstetter M (2004) Mars–robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23
18. Kontaxis G, Delaglio F, Bax A (2005) Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. Methods Enzymol 394:42–78
19. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486
20. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M (2003) Refinement of protein structures in explicit solvent. Proteins 50:496–506
21. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Calpha geometry: phi, psi and Cbeta deviation. Proteins 50:437–450
22. Montelione GT, Zheng D, Huang YJ, Gunsalus KC, Szyperski T (2000) Protein NMR spectroscopy in structural genomics. Nat Struct Biol 7 Suppl:982–985
23. 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
24. Newkirk K, Feng W, Jiang W, Tejero R, Emerson SD, Inouye M, Montelione GT (1994) Solution NMR structure of the major cold shock protein (CspA) from Escherichia coli: identification of a binding epitope for DNA. Proc Natl Acad Sci USA 91:5114–5118
25. Raman S, Huang YJ, Mao B, Rossi P, Aramini JM, Liu G, Montelione GT, Baker D (2010) Accurate automated protein NMR structure determination using unassigned NOESY data. J Am Chem Soc 132:202–207
26. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot T, Eletsky A, Szyperski T et al (2010) NMR structure determination for larger proteins using backbone-only data. Science 327:1014–1018
27. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93
28. Schindelin H, Jiang W, Inouye M, Heinemann U (1994) Crystal structure of CspA, the major cold shock protein of *Escherichia coli*. Proc Natl Acad Sci USA 91:5119–5123
29. Schneider WM, Tang Y, Vaiphei T, Mao L, Inouye M, Roth MJ, Montelione GT (2010) Efficient condensed-phase production of perdeuterated soluble and membrane proteins. J Struct Funct Genomics 11:143–154
30. Shan X, Gardner KH, Muhandiram DR, Rao NS, Arrowsmith CH, Kay LE (1996) Assignment of ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, and H^{N} resonances in an ^{15}N, ^{13}C, ^{2}H Labeled 64 kDa Trp repressor-operator complex using triple-resonance NMR spectroscopy and 2H-decoupling. J Am Chem Soc 118:6570–6579
31. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A et al (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690
32. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362
33. Suzuki M, Mao L, Inouye M (2007) Single protein production (SPP) system in *Escherichia coli*. Nat Protoc 2:1802–1810
34. Suzuki M, Roy R, Zheng H, Woychik N, Inouye M (2006) Bacterial bioreactors for high yield production of recombinant protein. J Biol Chem 281:37559–37565
35. Suzuki M, Zhang J, Liu M, Woychik NA, Inouye M (2005) Single protein production in living cells facilitated by an mRNA interferase. Mol Cell 18:253–261
36. Wider G, Wuthrich K (1999) NMR spectroscopy of large molecules and multimolecular assemblies in solution. Curr Opin Struct Biol 9:594–601
37. Yamazaki T, Lee W, Arrowsmith CH, Muhandiram DR, Kay LE (1994) A suite of triple resonance NMR experiments for the backbone assignment of ^{15}N, ^{13}C, ^{2}H labeled proteins with high sensitivity. J Am Chem Soc 116:11655–11666
38. Zheng D, Huang YJ, Moseley HNB, Xiao R, Aramini J, Swapna GVT, Montelione GT (2003) Automated protein fold determination using a minimal NMR constraint strategy. Protein Sci 12:1232–1246
39. 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