Published as: *Proteins*. 2009 April ; 75(1): 147–167.

Improving NMR Protein Structure Quality by Rosetta Refinement: A Molecular Replacement Study

Theresa A. Ramelot[1], **Srivatsan Raman**[2], **Alexandre P. Kuzin**[3], **Rong Xiao**[4], **Li-Chung Ma**[4], **Thomas B. Acton**[4], **John F. Hunt**[3], **Gaetano T. Montelione**[4], **David Baker**[2], and **Michael A. Kennedy**[*,1]

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

[2] Department of Biochemistry, University of Washington, and Howard Hughes Medical Institute, Seattle, Washington

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

[4] Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers University, Piscataway, New Jersey

Abstract

The structure of human protein HSPC034 has been determined by both solution NMR spectroscopy and X-ray crystallography. Refinement of the NMR structure ensemble, using a Rosetta protocol in the absence of NMR restraints, resulted in significant improvements not only in structure quality, but also in molecular replacement (MR) performance with the raw X-ray diffraction data using MOLREP and Phaser. This method has recently been shown to be generally applicable with improved MR performance demonstrated for eight NMR structures refined using Rosetta.[1] Additionally, NMR structures of HSPC034 calculated by standard methods that include NMR restraints, have improvements in the RMSD to the crystal structure and MR performance in the order DYANA, CYANA, XPLOR-NIH, and CNS with explicit water refinement (CNSw). Further Rosetta refinement of the CNSw structures, perhaps due to more thorough conformational sampling and/or a superior force field, was capable of finding alternative low energy protein conformations that were equally consistent with the NMR data according to the RPF scores. Upon further examination, the additional MR-performance shortfall for NMR refined structures as compared to the X-ray structure MR performance were attributed, in part, to crystal-packing effects, real structural differences, and inferior hydrogen bonding in the NMR structures. A good correlation between a decrease in the number of buried unsatisfied hydrogen-bond donors and improved MR performance demonstrates the importance of hydrogen-bond terms in the force field for improving NMR structures. The superior hydrogen-bond network in Rosetta-refined structures, demonstrates that correct identification of hydrogen bonds should be a critical goal of NMR structure refinement. Inclusion of non-bivalent hydrogen bonds identified from Rosetta structures as additional restraints in the structure calculation results in NMR structures with improved MR performance

[*] Correspondence to: Michael A. Kennedy, Department of Chemistry and Biochemistry, 701 E. High Street, Miami University, Oxford OH, 45056. michael.kennedy@muohio.edu. Phone: 1-513-529-8267. Fax: 1-513-529-5715.

Supporting Information Available. Supplementary tables have molecular replacement statistics for: Table S1: HSPC034 20 structure ensembles using Phaser, Table S2: idealized averaged HSPC034 structures using Phaser, Table S3: 20 structure ensembles using MOLREP, and Table S4: idealized averaged structures using MOLREP. This material is available free of charge via the Internet at http://pubs.acs.org.

Keywords

NMR; X-ray; HSPC034; PP25; C1orf41; Northeast Structural Genomics Consortium; structural genomics; comparison of NMR and X-ray structures; Rosetta; NMR force field refinement; molecular replacement; hydrogen bonding; X-ray crystallography; refinement methods

INTRODUCTION

The use of nuclear magnetic resonance (NMR) spectroscopy-derived protein models as templates for molecular replacement (MR) [2, 3] dates back to proof-of-principle in 1987 where an NMR model of the 46 amino acid protein, crambin, was used to phase crystallographic data from the same protein whose crystal structure was already known.[4] Use of a NMR-derived protein model to solve an unknown crystal structure by MR followed soon after when the NMR structure of interleukin 8 was used to solve its crystal structure in 1991.[5] Notwithstanding more than 30 examples of NMR models having been successfully used to solve protein crystal structures by MR since 1991,[6] it has been historically difficult to use NMR models for MR.[7, 8] This observation is interesting in light of the early demonstration that, at least in certain cases, it is possible to jointly refine a protein model against both NMR and X-ray diffraction data to yield crystallographic R factors and geometry of equal or better quality than obtained from conventional X-ray diffraction studies alone.[9] Difficulty in using NMR-derived protein models for MR can arise from real structural differences between solution and crystalline forms, structural differences caused by crystal packing effects, and/or lack of precision and accuracy the NMR model, caused by insufficient or misinterpreted NMR restraints.[7, 8] In order to obtain a correct MR solution, the generally accepted "rule of thumb" is that the root mean square deviation between the C^{α} backbone atoms of the model and the crystal structure must agree to within about 1.5 Å.[8] The inherent difficulty in using NMR models for MR has prompted recommendations and protocols on how to prepare NMR models to improve their MR performance.[7, 8] These suggestions have included replacing amino acids containing long side chains with alanine,[10] removing poorly defined regions of the NMR structure,[8] use of NMR ensembles,[11] and the assignment of distance-derived pseudo B factors to individual atoms.[12] However, the NMR spectroscopist can maximize the utility of NMR proteins models for MR by maximizing the quality and accuracy of the NMR model in folded regions of the protein.

Genuine differences between solution and crystal structures of proteins will always represent an upper limit in the use of NMR models for MR. However, since the first NMR-derived proteins structures were reported in the mid 1980's,[13, 14] the NMR community has collectively made substantial progress towards improving the quality and accuracy of protein structures derived from solution state NMR data. Part of this progress has been due to innovations and improvements in NMR methodologies for obtaining and utilizing experimental restraints.[15–28] The recent introduction of NMR RPF quality assessment scores used to assess "goodness of fit" between calculated structures and raw NMR data, which are similar in nature to X-ray R-factors, can also be used to guide and improve overall NMR-derived protein structure quality.[29]

It has also long been realized that, due to the sparseness of NMR restraints, the force field used for refinement can have a large impact on the quality, and possibly the accuracy, of NMR structures.[30] Accordingly, over the last decade, there has been substantial effort and progress towards improving the force fields and protocols used for refinement of protein NMR structures. A major step forward has been the treatment of solvent in structure calculations. Prior to the mid 1990's, NMR-derived protein structures were calculated in the highly unrealistic *in vacuo* environment. Since then, it has been shown that refinement using explicit water and ions could substantially improve the quality and precision of NMR models.[31–35] Other improvements have included more realistic treatment of non-bonded interactions [30] and inclusion of conformational database potentials.[36–38] Specifically, a new PARALLHDG force field was introduced in 1999 by Linge and Nilges for NMR protein structure refinement using covalent parameters based on the CSDX force field.[30] The final step in the implementation of this force field is a final short refinement in explicit solvent using the OPLS non-bonded parameters.[35, 39] The CSDX parameters,[40] derived from the Cambridge Structural Database,[41] have updated non-bonded interactions calculated from the PROLSQ program.[42] Importantly, the parameters from the CSDX force field are also used as the reference parameters for the commonly used structure validation programs such as WHATIF [43] and PROCHECK.[44] Furthermore, adopting the CSDX force field parameters has established some uniformity in the force fields used for both X-ray crystallography and NMR-based protein structure refinement, perhaps making it more meaningful when comparing NMR and X-ray structures of proteins.

Naturally, protein structures submitted to the protein data bank prior to introduction of refinement in explicit water and more sophisticated representation of non-bonded interactions have inferior quality scores when assessed by modern structure validation software packages. In spite of these improvement, NMR structures are still not subjected to a universally consistent refinement protocol, resulting in NMR protein models that vary considerably in structure quality. To address this problem, large numbers of NMR structures submitted to the protein data bank (http://www.rcsb.org) were “re-refined” using restraints deposited to the Biological Magnetic Resonance Bank (http://www.bmrb.wisc.edu) and the CNS water refinement protocol, resulting in the generation of the large RECOORD database of NMR protein structures refined in a uniform fashion.[45, 46]

Still, new efficient conformational sampling algorithms are being developed that might find useful application in NMR structure refinement protocols, e.g. replica exchange molecular dynamics using a generalized Born implicit solvent representation [47] and template-based selection of fragments for model building that has also been used to “re-refine” NMR structures resulting in NMR models closer to the corresponding X-ray counterparts.[48] A promising new approach for NMR protein structure refinement is embodied in the Rosetta method that employs a novel force field and conformation sampling algorithm that has been highly successful in a number of applications, including *de novo* protein structure prediction in recent CASP competitions,[49] novel protein design,[50] generation of physically realistic homology models,[51] rapid protein fold determination using sparse NMR restraints,[52] and determination of protein backbone conformations using residual dipolar couplings.[53] The Rosetta all atom potential includes van der Waals interactions, an orientation-dependent hydrogen bonding term, an implicit solvent model, and neglects long-range electrostatics.[49, 50, 54] The sampling protocol is designed to optimize the three-dimensional jigsaw puzzle-like packing of sidechains and evaluate the refined structures using the Rosetta free energy function.[49, 54]

In the context of our effort in the Northeast Structural Genomics Consortium (NESG, http://www.nesg.org), we face the dual challenge of generating large numbers of NMR-derived protein structures using highly automated structured determination methods,[55–57] while striving to maintain or increase structure quality.[29] Generating high quality NMR protein models for structural genomics targets is important given that every experimental NMR structure will initially represent an entire family of protein sequences for homology modeling and for solving future crystal structures of homologous proteins by MR.[58, 59] In this paper, we report the structure of the human protein HSPC034 (NESG ID: HR1958) solved by both NMR and X-ray crystallography (PDB IDs 1XPW.pdb and 1TVG.pdb, respectively), as well as an analysis of Rosetta refined protein models in terms of overall structure quality and MR performance. The Rosetta-refined NMR structures, which were calculated in the absence of NMR restraints, exhibited improved overall structure quality and MR performance compared to conventional structure calculation methods, and have improved agreement with the X-ray structure in loop regions, surface exposed side chains, metal binding site, backbone and side chains geometry, and the hydrogen-bonding network.

RESULTS AND DISCUSSION

Comparison of NMR and X-ray Structures

Both the NMR structure (1XPW.pdb) and the X-ray structure (1TVG.pdb) of HSPC034 have been solved by the NESG Consortium. HSPC034 protein obtained from expression of the same construct was used for both studies and includes the 10 residue N-terminal His tag sequence, MGHHHHHHSH (not included in sequence numbering), followed by the native sequence ending at residue Ser143. This sequence of HSPC034, also known as placental protein 25, has a deletion of D109 compared to the sequence of UniProtKB/Swiss-Prot entry Q9Y547/PP24_Human. Residue D109 was not in the sequence of the cloned protein used in this study.

The X-ray structure model at 1.6 Å resolution is in good agreement with the experimental data and expected geometric parameters (Table 1). The small difference between R-free and the standard crystallographic R factor (2.9%) indicates that the model is well refined. Electron density was only observed for residues 4–139, and not for the N-terminal 3 or C-terminal 4 residues. There is one molecule in the asymmetric unit, which is consistent with the protein being monomeric in solution. The structure was solved using a combination of SeMet data and data collected on a Sm derivative (SeMet + Sm). The data processing and refinement statistics are given in Table 1. Two heavy atom sites were identified: the Se atom of SeMet44 and a Sm^{+3} ion. The side chain for Met44 is well ordered (B-factor for Se atom is 8.38 $Å^2$) despite being located in a surface loop. The loop consists of hydrophobic residues (Gly43, Met44, Phe45, Pro46) that form hydrophobic interactions with a symmetry related molecule. The single heavy atom Sm^{+3} is located close to the crystallographic axis and is bound to the carboxylate side chain of Asp92 in two symmetry related molecules of HSPC034. A hepta-coordinate Ca^{+2} ion is bound in a loop. Four coordinate covalent bonds with lengths 2.37 Å, 2.46 Å, 2.57 A, and 2.64 Å are formed with carbonyl oxygens of Asn29, Asn34, Thr37 and His129. The side-chain oxygen atoms of Thr37 and Asp32 are located 2.64 Å from the Ca^{+2} ion. A well-ordered water molecule (temperature factor of 6 $Å^2$) is located 2.62 Å from the Ca^{+2} ion. Structure quality scores obtained from the protein validation server, PSVS,[60] are also given in Table 1. The Z-scores reported for PROCHECK, Verify3D, ProsaII, and the MolProbity clashscore, are all within two standard deviations of the average for the high-resolution crystal structure database used to calibrate the score. These scores are not as close to zero as would be expected for protein with this high of resolution.[60] However, it's important to remember that Z-scores are normality scores, rather than quality scores, and are defined as the number of standard deviations away from the mean of the database. In general, mainly β-sheet proteins, like HSPC034, have more negative Z-scores than α-helical proteins (this has been reported for a representative NMR data set [46]). Two other proteins with a similar fold, the galactose-binding domain of *Micromonospora viridifaciens* sialidase, 1EUT,[61] and human anaphase-promoting complex subunit 10, 1JHJ,[62] also have PROCHECK and MolProbity clashscores that are more negative than average for their resolution (average values in reference.[60])

The NMR structure of HSPC034 was solved by standard triple-resonance protocols. Chemical shifts assignment for ^1H, ^{13}C, and ^{15}N atoms were 97.6% and 92.5% complete for routinely assignable backbone and side chain resonances for residues 1–143 and were deposited in the BioMagResBank (BMRB ID 6344). ^1H-^{15}N HSQC crosspeaks were missing for the 10 N-terminal His-tag residues as well as Ile17, Phe38, Ser66, and His129. NMR structures were calculated with 923 NOE, 93 hydrogen bond, and 210 dihedral angle restraints using a standard Xplor simulated annealing protocol followed by refinement in explicit water in CNS, CNSw. ICP-MS analysis confirmed that stoichiometric (1:1) calcium ion was bound in the NMR structure. Statisitics for the NMR assignments and calculated structures are in Table 2, including structure validation statistics calculated by PSVS. The Z-scores reported for PROCHECK, Verify3D, ProsaII, and MolProbity clashscore, are all better than −3. Z-scores, with the exception of ProsaII, are better than average for NMR structures in the PDB.[60] The ProsaII Z-score, which models a reduced-representation energy of pair-wise interactions from the spatial separation of residues, is similar to that observed for the X-ray structure and is therefore a typical score for this structure. The Z-scores that are most sensitive to X-ray structure resolution, PROCHECK G-factor (phi-psi), PROCHECK G-factor (all dihedral) and MolProbity clashscore, are −2.2, −2.8, and −2.4, respectively (PROCHECK values are for ordered residues). These scores are comparable to averages for low-resolution crystal structures in the PDB (2.5–3.5 Å resolution). In all cases, the X-ray structure of HSPC034 has Z-scores closer to zero than the NMR structure. The X-ray and NMR structures are similar with average backbone (N,C^{α},C′) and heavy atom RMSD values of 1.24 ± 0.18 Å and 2.01 ± 0.17 Å, respectively (residues 6–138). The RPF R/P/DP scores (Table 2) indicate that the NMR structure has a global “good fit” with the NMR data.[29]

The core structure of HSPC034 is a β-sandwich with a jelly-roll topology (Figure 1AB). Two beta sheets make up the sandwich structure: a five-stranded antiparallel β sheet (strands β2, β3, β7, β4, β5) and a three-stranded antiparallel beta sheet (β6, β3, β8). The short β strands β2′ and β1 may be included in the smaller sheet. Strand β2′ is antiparallel to β8, however, β1 is parallel to β8 in the X-ray structure and antiparallel in the NMR structure. In the X-ray structure, a hepta-coordinate Ca^{+2} ion is bound in the calcium-binding loop, residues 27–38, which includes the 3^{10} helix H2. In the NMR structure this loop is not well restrained by NOE data even though ICP-MS analysis indicated bound stoichiometric (1:1) Ca^{+2} ion. The major differences between the X-ray and NMR structures of HSPC034 (Figure 1C) are located in the calcium binding loop, loop 42–48 between β2′ and β3, loop 93–100 between β5 and β6, the C-terminal end of strand β8, and the N-terminal strand β1. In the X-ray structure the electron density is only observed for residues 4–139, whereas in the NMR structure residues 2–3 are extended and have NOEs to strand β8, and residues 139–141 have NOEs that extend strand β8 and connect it with β3.

In the NMR structure of HSPC034, strand β1 is antiparallel to β8 and there is no indication of a contribution from a population of parallel β1. NOE cross peaks that would be present for the parallel population based on short distances in the X-ray structure were not observed, even at baseline threshold. In the X-ray structure the parallel strand is well defined and supported by the position of the strand in a difference electron density map; i.e. the positions of all of residue D5 and the carbonyl group of I4 are clearly defined. In the refined x-ray structure, the B-factors for I4 and D5 are about twice as large compared to the average B-factor for the overall structure. Based on Rosetta calculations (below), we find that both the parallel and antiparallel conformations of β1 are observed in low-energy structures, suggesting that differences in the NMR solution and X-ray crystal environment, such as temperature, pH, and salt, contribute to the favorability of one over the other in each system.

Comparison of X-ray structure to NMR derived restraints

In order to pinpoint regions where the X-ray structure is not consistent with the NMR derived restraints, the program PSVS was used to report violations of dihedral, NOE, and hydrogen bond restraints by the X-ray coordinates of HSPC034. There were thirteen dihedral angle violations > 1°, ninety-five NOE violations > 0.1 Å, and no hydrogen bond violations > 0.1 Å. The largest dihedral angle violations (> 50°) were for φ and Ψ of D5 and G43 that are located in two regions where the X-ray and NMR structure differ: the N-terminal six residues and the G43 loop. There were thirteen NOE violations involving residues 2–3 and 140–141 for which the X-ray structure did not include coordinates. Aside from these, the largest NOE restraint violations are found in N-terminal strand β1 (I4-L6) and three C-terminal residues (S139-L141). Additionally, NOE violations > 2 Å were found for H54, which is in a different rotameric state in the X-ray and NMR structures. NOEs from $H^{\delta 2}$ of H54 to S13 and E14 define the orientation of the H54 side chain in the NMR structure. NOE violations > 0.5 Å are shown in Figure 1D with violations > 2 Å in red, 1–2 Å in orange, and 0.5–1 Å in yellow. In general, violations between 1–2 Å are found mostly in solvent exposed side chains in loops or involve His, Ile, or Leu residues. Four His residues, H54, H95, H108, and H124, have NOE violations in this range, possibly due to differences in the His ring protonation states between the NMR and X-ray structures that could lead to differences in structure. In general, most violations between 0.5 and 1 Å involved buried side chain methyl groups and could indicate restraints that were improperly treated for cross peaks effected by spin diffusion in the NOESY data. No NOE violations > 0.1 Å were found in the calcium ion-binding loop. Therefore, although the NMR structure has few restraints and is poorly converged for this loop, it is consistent with the geometry seen in the X-ray structure. In fact, NMR structure calculations with Ca^{+2} ion restrained with the same coordination as observed in the X-ray structure resulted in structures that were consistent with all observed NOEs. This doesn’t mean that the Ca^{+2} binding site and coordination is the same as in the X-ray structure, but just that the NMR data is insufficient to characterize the structure of this loop.

Rosetta refined NMR structures

The CNSw NMR structures were refined by Rosetta using a recently described protocol that performs random backbone perturbations and sampling of discrete side-chain conformations followed by full atom refinement.[1] Each of the 20 CNSw NMR structures was used as a seed to generate 1000 Rosetta refined structures using the high-resolution perturbation sampling protocol. The 20 lowest-scoring structures of the final 20,000 were selected for further analysis. Visually, the improved agreement with the X-ray structure can be seen in Figure 2. A backbone overlay with the X-ray structure for both the CNSw and Rosetta structures is shown in Figure 2A. After refinement, the backbone conformations of N- and C-termini did not converge with the X-ray structure, but remained similar to the starting CNSw models. The calcium-binding loop, became more similar to the X-ray structure after refinement (Figure 2B) even though the Rosetta refinement did not include parameters for the calcium ion. A marked improvement in agreement with the X-ray structures can be seen for the charged side chains. In Figure 2C, the side chains for Asp, Asn, Glu, and Gln are shown. These side chains are predominantly located on the surface of the protein and have sparse restraints. For comparison, the side chains for Trp, Tyr, and Phe, which are relatively rich in NOE restraints, did not have as large an improvement in agreement with the X-ray structure (Figure 2D).

Comparison of Rosetta refined structures to NMR derived restraints

After Rosetta refinement, the 20 lowest scoring structures were compared to the NMR restraints to identify violations. Although these structures were calculated without NMR restraints, the resulting structures had only a few large NOE violations with none greater than 5 Å and an average of only 17.3 violations > 1 Å. Only two NOE restraints were violated in all 20 structures by > 0.5 Å and only nine by > 0.1 Å. The only dihedral angle restraint that was violated in every structure by > 2° was for φ of G109. This restraint has a minimum violation of only 12°, and is probably too tightly restrained since the phi angle is within 20° for the X-ray and NMR structures. Only two dihedral restraints were violated in every structure by > 1°. Taken together, the low number of restraint violations and similar RPF-DP scores (Table 3) indicate consistency with the raw NMR data. This demonstrates that low energy structures with alternate conformations generated by conformational sampling with the Rosetta force field can be found that are consistent with the NMR data and correspond to lower energy minima in the global free energy landscape.

On average, there were less NOE and dihedral restraint violations than were found for the X-ray structure. An average of 12.7 dihedral angle violations > 1°, 67.8 NOE violations > 0.1 Å, and 0.4 hydrogen bond violations > 0.1 Å per structure was calculated using PSVS. The NOE violations correspond to an average of 7.3% violations of the total number of NOE restraints per model (compared to 10.5% for the X-ray structure). There is < 50% agreement between the NOE restraints violated by the Rosetta structures compared to the X-ray structure (by > 0.1 Å). Comparing the 62 Rosetta violations to the 83 violations for the X-ray structure (excluding restraints for residues 2–3 and 140–141), only 47% are violated in both cases.

A total of 216 NOE restraints were violated by > 0.1 Å in any of the 20 models. Of the 67 NOE restraints violated by > 0.1 Å in 10/20 structures, the majority, 64%, contain an ILV proton in the restraint and 10% involve a His proton (data not shown). Most of the violations involve side chains that are in the hydrophobic core of the protein. NOE restraints violated by Rosetta structures may be incorrectly assigned restraints or restraints derived from cross-peaks affected by spin diffusion. Alternatively, certain restraints could accurately reflect the NMR structure in solution, and be violated in some Rosetta structures that represent alternative low energy conformations for a certain region. Likely, all of these possibilities contribute to the number of NOE restraints violated by the Rosetta refined structures. It may be advantageous to use the NOE restraint violations identified after Rosetta refinement as a guide to identify incorrect NOE assignments and/or spin diffusion affected cross peaks to obtain more accurate NMR structures.

Working backwards – Rosetta refinement of the X-ray structure

Examination of the Rosetta refined X-ray structures of HSPC034 shows us the best that Rosetta refinement of the NMR structure will be able to do with the current algorithm, given sufficient sampling, perturbations and minimizations to fully sample conformational space. In addition, areas where Rosetta-refined structures deviate from the starting X-ray structure can indicate regions of the X-ray structure that involve interactions that are not taken into account during the refinement such as crystal-packing and metal-binding or regions that have similar Rosetta energies and therefore may represent multiple low-energy conformations for a part of the structure. Lastly, deviation from the X-ray structure may indicate areas where Rosetta parameterization needs adjustments.

Rosetta refinement of the HSPC034 X-ray structure was used to calculate 1000 structures. Interestingly, 763 of the refined structures have β1 in the parallel conformation like the starting X-ray structure (X-Ros-para), whereas 237 are antiparallel, like the NMR structure (X-Ros-anti). The backbone RMSD for these structures and their Rosetta energies are shown in Figure 3 along with the 20,000 Rosetta refined NMR structures. Looking more closely at the 20 lowest scoring structures with the parallel β1 strand and the 20 lowest scoring structures with the antiparallel β1, we see several regions of the structure, mostly in loops, have moved away from the starting X-ray structures (Figure 4A–C). There are several reasons that the Rosetta refinement could cause divergence from the starting X-ray structure. Since the backbone region with the largest RMSD from the X-ray after refinement is the calcium-binding loop (res. 28–38), the deviation is likely due to the missing calcium (or samarium) ion that was not included in the Rosetta refinement. Additionally, HSPC034 makes crystal contacts with six other protein molecules in the unit cell, resulting in packing interactions that are not represented by Rosetta during the refinement of the monomeric X-ray coordinates. Residues that have any atomic distances < 5 Å with the symmetry related molecules are indicated in Figure 4A. These are present for 30% of the residues and are spread out across the sequence and structure. There are six backbone-backbone hydrogen bonds between the protein and the symmetry related molecules: F45H^N ---F45C=O, D84H^N --- V106C=O, S140H^N --- T42C=O, and three reciprocal hydrogen bonds. Interestingly, the residues F45, D84, and S140, which are involved in inter-protein hydrogen bonds, have large divergence from the X-ray structure after Rosetta refinement (Figure 4B,C). In addition, the G43 loop, which had a large difference between the NMR and X-ray structures, has many close contacts to symmetry related molecules in the crystal (Figure 4A). Rosetta refinement caused the backbone dihedral angles for G43 to become more similar to the NMR structures (recall that this was the largest NMR dihedral angle violation for the X-ray structure compared to the NMR restraints).

Another reason for variations in the structure after Rosetta refinement is that there may be several conformations that are similarly favorable, especially under different conditions, and crystallization may trap out just one of them. This is a commonly accepted idea since crystals of the same protein in different isoforms can have different conformations, and even in the same single crystal, multiple copies of the same protein in the asymmetric unit can have small structural differences. Additionally, in very high-resolution crystal structures it becomes evident that side-chains can have two or more populated states that are each partially occupied. In HSPC034 we see several examples where specific side chains adopt different rotomeric conformations in the NMR structure and the X-ray structures and the Rosetta-refined X-ray structures contain conformations that are representative of each. The antiparallel and parallel conformations of strand β1 is just one example. The conformation of this strand depends on the conditions experienced by the sample in the study, and both conformations, one similar to the X-ray structure and one similar to the NMR structure, are observed in low-energy Rosetta refined structures.

Structures obtained from Rosetta refinement of the CNSw NMR structures have worse agreement with the X-ray structure (higher RMSD) than structures calculated by refinement of the X-ray structure (Fig. 3). In addition, the energies of the 20 lowest Rosetta energy structures are much higher than those calculated in the X-ray refinement (compared to the 20 lowest energy structures for β1 parallel and anti-parallel) and the structural ensemble has more variability (Fig. 3 and Fig. 4B,C,D). All 20,000 structures have β1 in the antiparallel conformation, similar to the starting NMR structures (Figure 3). However, with enough sampling and loop rebuilding, the lowest energy structures from Rosetta refinement of the NMR structure should be able to match that from the X-ray refinement. Because of the number of backbone perturbation possibilities, in addition to the sampling of discrete side chain rotamers, the number of structure calculations needed to sample all the possible structural perturbations becomes large. Even though the sampling of the protein free energy landscape using Rosetta cannot be exhaustive due to time limitations, the 20 lowest energy structures refined with this method all had better MR scores than the starting CNSw NMR ensemble and had better RMSD to the X-ray structure. In an additional refinement calculation with 5X more structures (100,000 calculated), several slightly lower energy structures were obtained (data not shown). These structures still had approximately the same RMSD to the X-ray structure as the previous smaller sampling refinement, indicating that the time needed to do more thorough conformational sampling with the current protocol is orders of magnitude longer than this longer two-week calculation.

Impact of NMR refinement methods on Molecular Replacement

NMR structures are calculated using various simulating annealing protocols and by minimizing an energy function that consists of both experimental information derived from interpretation of NMR data and force field terms that describe covalent and non-bonded interactions. The choice of protocol and force field affect the quality and accuracy of the resulting structure, even if all structures satisfy the experimental restraints. This is because the NMR restraints are sparse at certain locations such as loops and side chains on the protein surface, which can have exchange broadened peaks, and also because the upper and lower bounds for NOE distance restraints, hydrogen bonds, and dihedral angles allow for a range of conformations that satisfy the restraints.

Structures calculated using standard protocols exhibited improved MR performance using both MOLREP and Phaser in the order: DYANA, CYANA, Xplor-NIH, and CNSw (Figure 5). All structures were calculated with the same experimental restraints and none had NOE violations > 1 Å or dihedral angle violations > 1°. This trend was also observed for the backbone and side-chain RMSD (residues 6–138) to the X-ray structure for structures calculated by the different methods (Fig. 5). Although they follow a clear trend, the improvements in backbone and heavy atom RMSD observed when calculating the structures with DYANA, CYANA, Xplor, or CNSw were small and typically within the error bars for the measurements (Fig. 5A). This is in agreement with the previous observation that CNSw refinement of 26 NMR structures resulted in only a small and not significant improvement in RMSD to their corresponding X-ray structures after recalculation.[46] The MR performance metrics for both Phaser and MOLREP showed improvement in the shift of the average values that correlate with the small improvements in RMSD (Fig 5B-E). The general "rule-of-thumb" is that the Phaser translation function Z-score (TFZ) should be > 5 for a reliable MR solution.[63, 64] Weak solutions may start out with rotation function Z-scores (RFZ) < 5. For HPSC034, the mean RFZ scores for DYANA, CYANA, and Xplor structures are < 5, whereas the CNSw refined structures have a mean RFZ score > 5 (Fig. 5B). The Phaser log likelihood gain after packing and final refinement (LLG(R)), MOLREP RF/σ and TF/σ scores also improved after CNSw refinement. Improvements seen in the Phaser mean score for MR were larger than for MOLREP, consistent with the superior sensitivity of the likelihood based method implemented in Phaser compared to the Patterson based method used in MOLREP. Taken together, the MR results illustrate that even small improvements in agreement with the X-ray structures can enhance MR performance. It is standard protocol in the NESG Constorium to refine all structures with CNSw, and the deposited NMR structure of HSPC034 was calculated using Xplor followed by CNSw (PDB ID: 1XPW). Structures calculated by any method can be refined with CNSw, however for HSPC034, CYANA structures refined with CNSw had slightly higher, (but not significant) RMSD to X-ray compared to the Xplor structures refined with CNSw (1.33 ± 0.14 and 2.10 ± 0.13 for backbone and heavy atom RMSD). For this reason we used the CNSw-refined Xplor structures for the CNSw ensemble analysis.

Since it has been reported that using an average structure for the NMR ensemble can sometimes improve the MR performance,[5, 12, 65] we generated idealized average structures for each NMR ensemble using Rosetta parameters for idealization (residues 4–139). These structures had better RMSD to the X-ray structure than the average ensemble RMSD (Table 3), typically similar to or better than the best member of the ensemble. This correlates with the MR performance for the idealized average structures that were consistently improved compared to the average score for each ensemble (Figure 5).

Significantly improved MR performance was obtained by further refinement of CNSw NMR structures with a Rosetta protocol in the absence of NMR restraints (Figure 5). The protocol uses an iterative perturbation and minimization algorithm to search for alternative energy minima. A similar method was recently demonstrated for refinement of 8 NMR structures and gave better MR solutions in all cases.[1] Molecular replacement scores are reported for each ensemble of structures and idealized average structures using both Phaser and MOLREP in Supplementary Material S1–S4.

The Rosetta refined structures also have significantly improved quality scores and were equally consistent with the NMR data according to the RPF-DP scores. Both the individual, DP(each), and average distance, DP(ave), RPF-DP scores are reported in Table 3 as well as the backbone and heavy atom RMSD to the X-ray structure for residues 6–138. The RPF-DP scores are relatively insensitive to change in structure when the RMSD variation is small as seen here. However, the Rosetta structures and the X-ray structure correlate with the NMR NOESY peak lists data just as well as the NMR structures calculated by different methods. In addition, in Table 3 the structure quality Z-scores for PROCHECK (all dihedral angles) and MolProbity clashscores obtained from PSVS for each of the NMR calculated ensembles of 20 structures are listed. These two knowledge-based Z-scores have been found to be the most sensitive to X-ray structure resolution and to the accuracy of NMR structures.[60] For Rosetta refined structures, there is no apparent correlation with quality scores and RMSD to the X-ray structure or MR performance since the Rosetta refined structures had better quality Z-scores than the deposited X-ray structure, but did not have better MR performance. There is still a substantial gap between MR solutions for Rosetta refined NMR structures and for the X-ray structure (Figure 5). The gap can be attributed in part to real structural differences discussed earlier, to crystal-packing effects, and to inferior hydrogen bonding networks.

Aside from the RMSD to the X-ray structure, which is not known *a priori* to solving the X-ray structure by MR, we have found that the number of unsatisfied hydrogen bond donors and acceptors is a good predictor of MR performance (Figure 6). Values are also listed in Table 3 for the average of each ensemble. The X-ray structure has just three unsatisfied donors for three amide protons (S20, K24, and F126) and one acceptor (Q65 OE1). Further minor rotations of the side chain positions of D22 and Q65 leaves the structure with only two unsatisfied hydrogen bond donors and no unsatisfied acceptors. Since buried unsatisfied hydrogen bonds are energetically unfavorable, a large number of buried unsatisfied hydrogen bond donor or acceptors is a symptom of a poor structure. In fact, recent reviews of hydrogen bonding in proteins predict that all hydrogen bond donor or acceptors should be satisfied a significant fraction of the time by hydrogen bonds to protein atoms or to water.[66] The correlation of the number of unsatisfied buried hydrogen bonds donors and acceptors with the backbone RMSD to the X-ray structure (Figure 6A) has correlation coefficients of 0.96 and 0.75, respectively. The correlation coefficient between the number of unsatisfied donors and the Phaser LLG (TF) and (RF) metric (Figure 6B) is 0.91 for both. Due to the importance of hydrogen bonding and the correlation with improved agreement to the X-ray structure, we have done a more comprehensive hydrogen bond analysis of the structures calculated by different methods and compared them to those in the X-ray structure.

Hydrogen bond analysis

Hydrogen bonds have a significant influence on structure quality and also are very dependent on force field. Poor agreement between main chain hydrogen bonds was identified in a recent comparison of 78 NMR and corresponding X-ray structures.[67] For HSPC034, the deposited CNSw NMR structure has only 65% coincidence of backbone hydrogen bonds with the X-ray structure (% coincidence is 100 × number in both/total number found in X-ray and NMR structures), and 89% of the hydrogen bonds found in the X-ray structure (100 × number in both/number in X-ray). For the NMR ensembles, the number of hydrogen bonds is the average value for all 20 structures. In the NMR calculations, only 31 hydrogen bond restraints were included, using a conservative approach where only HNs in β-strands with NOEs to both the H^{N}_{i}, H^{α}_{i-1}, on either side of the O of the antiparallel β-strand were restrained. The hydrogen bond restraints were defined as upper and lower bound distances between the H^N--O, N--O, and C′--O atom pairs. These restraints allow for a generous range of hydrogen bond angles that includes all those observed in the X-ray structure for the corresponding hydrogen bonds. Due to the generosity of the restraint boundaries and the sparseness of hydrogen bond restraints, the force field used for refinement is influential in determining the hydrogen bond network as well as the geometries of the resulting hydrogen bonds.

Currently, improvements the hydrogen bond potential can be used within CNS and XPLOR-NIH to further improve the geometries and energetics of backbone-backbone hydrogen bonds in calculated NMR structures [68] The Xplor structures were further refined with an improved Xplor-NIH protocol, which includes the backbone hydrogen bond potential as well as other database potential terms, and are referred to as Xplor+. [69, 70] Refinement using the simulated annealing protocol and hydrogen bond potential of mean force (no other database terms) within Xplor-NIH was also tested,[68] and showed improvement in quality metrics and hydrogens bonding (data not shown). However, the structures refined with CNSw and Xplor+ were superior in all metrics reported in this paper.

The coincidence of backbone hydrogen bonds found in both the X-ray structure and NMR structures computed by different methods decreased in the order: Rosetta > Xplor+ > CNSw > Xplor > CYANA > DYANA (Figure 7A). The best agreement was for Rosetta refined structures with an average 71% coincidence for the ensemble, considering the total number hydrogen bonds with DSSP energies greater than −0.5 Kcal/mol. The X-ray structure has 80 hydrogen bonds (for residues 6–138) calculated with this method. This corresponds to 60% of residues, which is a number typical for crystal structures. Consistent with other recent hydrogen bonds analyses,[67] it was observed that NMR structures calculated by Xplor, CYANA, and DYANA, have about the same total number of hydrogen bonds as the X-ray structure and that CNSw increased that number (Figure 7B). Although this resulted in an improvement in the coincidence of hydrogen bonds with the X-ray structure, the coincidence was attenuated by the increase in hydrogen bonds that were not found in the X-ray structure. In summary, the coincidence of hydrogen bonds between the NMR and X-ray structures was increased by either Xplor+ or CNSw refinement and was further improved by the Rosetta refinement.

NMR structures typically have fewer “strong” (low energy) hydrogen bonds and more “weak” (high energy) hydrogen bonds than X-ray structures, due in part to the broad range of N-H^N-O bond angles allowed when defining hydrogen bond restraints.[67, 71] We also observe that there are more “bivalent” hydrogen bonds and less long-range hydrogen bonds in HSPC034 NMR structures calculated by all methods. After filtering out the bivalent hydrogen bonds that are of similar strength and keeping the strongest one when there is one strong and one weak (see methods), there was no overall increase in the coincidence of the remaining hydrogen bonds with those in the X-ray structure, with the exception of the Rosetta calculations (Figure 7A). However, if just the long-range, non-bivalent hydrogen bonds are considered, there is improved coincidence with the X-ray structure (up to 86% for the Rosetta average) with the order: DYANA, CYANA, Xplor, CNSw, Xplor+, Rosetta.

The X-ray structure of HSPC034 has 63 long-range, eight i+2, and eight i+3 hydrogen bonds (Figure 7B). None of the NMR refined structures had as many long-range hydrogen bonds, although the Rosetta refinement came closest with 61.5. In all cases NMR refined structures have more i+2 hydrogen bonds than X-ray. These i+2 hydrogen bonds are found primarily in loop regions and often specify γ turns. Interestingly, Rosetta refinement of the X-ray structure (below) resulted in an increase in the number of i+2 hydrogen bonds to about 17.5 per model. This suggests that the Rosetta force fields favor short-range hydrogen bonds at the expense of long-range hydrogen bonds found in the X-ray structure, although to a lesser extent than the other force fields. Long-range hydrogen bonds that were found in the X-ray structure but not in Rosetta structures were primarily atypical hydrogen bonding patterns in the β-sheets. Rosetta correctly identified all of the i+3 hydrogen bonds in both 3^{10} helixes in all 20 lowest scoring structures in the ensemble.

Using hydrogen bond restraints obtained from Rosetta in NMR refinement

Since, *a priori* knowledge of the hydrogen bond network from a corresponding X-ray structure is typically not available, we ran Xplor followed by CNSw calculations including the 56 non bivalent hydrogen bonds identified in > 70% of Rosetta calculated structures and excluding the i+2 hydrogen bonds. Inclusion of these restraints did not result in any NOE violations > 0.1 Å. The backbone and heavy atom RMSD of these structures compared to the X-ray structure are better than the deposited CNSw structures, however the improvement is only significant for the heavy atom RMSD (+RosHBs, Figure 5). The i+3 hydrogen bonds in the 3^{10} helix of the calcium-binding loop were clearly identified. The MR performance was also improved with the score distributions shifted to higher scores. If it had been possible to identify all 61 non-bivalent hydrogen bonds from the X-ray structure (res. 6–138), then the RMSD to the X-ray structure and the MR performance could be improved a bit more (data not shown), and all these additional hydrogen bonds were also consistent with all NOE data. Improvements in side-chain hydrogen bonds to backbone and to other side chain atoms have not been examined in this study, but likely can account for some of the additional improvement in MR for Rosetta structures calculated without NMR restraints. Importantly, using the hydrogen bonds identified by Rosetta calculations as restraints is a way to use the Rosetta refinement method to improve our NMR structures and still refine them using the NMR restraint data.

CONCLUSIONS

Perfect agreement between the NMR and X-ray structures will always be impossible because of differences caused by crystal packing and different protein environments, as was seen here for HSPC034. However, it is clear from this study that changes in refinement methods can improve the agreement between NMR and X-ray structures and therefore improve the ability of NMR structures to be used for MR. The best MR performance was made possible by refinement of the NMR structures with the Rosetta force field using a new protocol in the absence of NMR restraints. Rosetta emphasizes short-range electrostatic interactions and rotamer sampling in order to optimize side chain packing. It also incorporates a knowledge-based hydrogen bond potential that is secondary structure dependent and superior to the simple distance-dependent Coulomb treatment of electrostatic interactions. [50] Rosetta refined structures had the best agreement with the backbone hydrogen bonds found in the X-ray structure, although there were still fewer long-range hydrogen bonds compared to the X-ray structure and there were still differences in long-range hydrogen bonding patterns, especially where there are atypical hydrogen bonds in a β-sheet.

It is clear that better identification of hydrogen bonds pairs is important to increasing the backbone similarity between the NMR and X-ray structure and will improve MR performance, and therefore should be a critical goal for NMR structure refinement. The 61 non-bivalent hydrogen bonds in the X-ray structure (res. 4–137) were consistent with NOEs and could be used to calculate a better NMR structure. However, we had no *a priori* knowledge of these hydrogen bonds. We could, however, identify 56 non-bivalent backbone hydrogen bonds from the Rosetta refined NMR structures. Structures calculated with these added restraints had no additional NOE violations, were more similar to the X-ray structure and had superior MR performance while retaining the benefit of having been refined against the NMR restraints. The Rosetta refined structures without NMR restraints had slightly better MR performance, which may be attributed at least in part to treatment of non-backbone hydrogen bonds and other electrostatics involving charged side chains.

Rosetta refinement of NMR structures without NMR restraints provides an independent exploration of the low energy landscape compared to conventional approaches. This gives Rosetta potential utility as an independent cross-validation technique for NMR models and restraints, which could aide in the identification of incorrect restraints. However, Rosetta refinement of NMR structures in the absence of experimental NMR restraints should not be considered an alternative method for generating NMR models. Ultimately, in order to take full advantage of Rosetta for calculating NMR models, it will be necessary to modify the Rosetta program to make use of experimental NMR restraints, a task that is underway in the Baker laboratory. While further improvements in NMR structures and hydrogen-bonding patterns can be made by collecting additional NMR data such as RDCs or measurement of small hydrogen bond coupling constants,[71, 72] we find that hydrogen bonds identified from Rosetta calculations can be used to improve calculated structures without additional data.

METHODS

Protein Purification

The human protein HSPC034 was cloned, expressed and purified using standard methods in order to produce SeMet or *U*-^{13}C, ^{15}N-labeled protein. The HSPC034 gene was cloned into pET14 vector and sequences was verified by DNA sequence analysis in both directions (D109 is not present). The protein, which contains 10 N-terminal residues (MGHHHHHHSH), was expressed in *E.coli* strain BL21-(Gold DE3) and purified by Ni-NTA affinity (Qiagen) followed by gel-filtration chromatography (HiLoad 26/60 Superdex 75 PG, Amersham Biosciences). The chromatography buffer was 20 mM Tris, 500 mM NaCl, 30 mM imidazole, pH 8.0, and the sample was eluted in the same buffer with 500 mM imidazole. Sample purity (> 97%) and molecular mass were confirmed by SDS-page and MALDI-TOF (17.5 kDa for [*U*-^{15}N; 5%-^{13}C]HSPC034). Analytical static light scattering measurements in-line with gel-filtration chromatography confirmed that the protein is monomeric in solution. For NMR, the labeled protein was concentrated and the buffer exchanged by ultracentrifugation and repeated dilution followed by concentration into the NMR buffer (below). For X-ray crystallography the protein was concentrated to 1.7 mg/ml and exchanged into 10 mM Tris-HCl, 5 mM DTT, and pH 7.5.

Crystallization and Crystal Structure Determination

Human protein HSPC034 containing SeMet was crystallized at room temperature by vapor diffusion in hanging drops. Drops were set up by mixing 2 μL of concentrated protein solution with 2 μl of reservoir solution (18% PEG and 200 mM $CaCl_2$). Crystals were cryoprotected in paratone-N for several seconds then flashed-cooled in liquid propane. Multiwavelength anomalous diffraction (MAD) data sets, at the edge, peak, and remote absorption of Se were collected at the National Synchrotron Light Source X4A beamline (Brookhaven National Laboratory, Brookhaven, New York). This beamline is equipped with a QUANTUM-4 charge-coupled device (CCD) detector. A total of 420 images for each of three wavelengths were recorded (210 images in the one direction and 210 images in the reverse direction). To reduce systematic errors in the scaling of Friedel pairs due to decay, the ϕ angle was changed by 180° after every 30 images. All synchrotron data were collected at 100 K and processed with the HKL software package.[73] The crystals belong to space group C2, with unit cell parameters a = 70.97 Å, b = 41.62 Å, c = 46.78 Å, and β = 102.2°. The asymmetric unit contains one protein molecule with a solvent level of 39%.

The computer program package SOLVE [74] was used to locate the heavy atom sites in the protein. Although there are three Met residues (out of 154 total residues), two are in the unstructured N-terminus and the anomalous signal from the one ordered Se was not enough to solve the structure directly from the MAD data. Estimation of the Se anomalous contribution by comparing of scaling of Friedel pairs as individual reflections and averaging as symmetry related reflections shows a difference of about 0.6%, which is low.

The structure was solved using additional data from a Sm derivative crystal that was generated by soaking a SeMet crystal for 24 hours in the mother liquor containing 4 mM of Sm acetate (SeMet + Sm). This data set was collected on the same X4A beamline using a Se peak absorption wavelength of 0.979Å. The contribution of Sm was estimated by analysis of the result of scaling three frames (3° oscillation) of derivative against the peak data of SeMet protein.[75] The value for χ^2 ~ 14.7 at 3.5 A resolution indicated that the Sm derivative data could be used to phase protein amplitudes. Phasing, heavy atom location, and occupancy refinement was carried out with the program SOLVE.[74] The combination of two data sets at peak wavelengths (SeMet and SeMet + Sm) was used to locate two sites for heavy atoms (Table 1). The ratios of heavy atom heights to background variations were 7.7 and 6.2, and the averaged merit factor for phases was 0.42 at 3.0 Å resolution. The program RESOLVE_BUILD (version 2.06) [76] was used to generate an initial partial model at 2.5 Å resolution. The best model was constructed from 121 amino acids, 67 of which had side chains. The R-free factor and standard crystallographic R-factor were 0.404 and 0.370, respectively. The missing residues in the partial model of the protein were built manually on a Silicon Graphics Octane workstation using interactive computer graphic programs CHAIN and O.[77, 78] Amino acid residues, which were initially assigned as Ala or Gly, were corrected. Refinement of the protein model was carried out by iterative refinement using CNS (version 1.1).[79] As intensities were much weaker and completeness dropped to ~60% in the highest shell, we reduced the resolution shell for refinement to 1.6 Å. Reflections included in the refinement were gradually extended from 2.5 Å to 1.6 Å with sigma cutoff $F \geq 2\sigma(F)$. To avoid model overfitting and overestimation of structure quality, 10% of reflections were randomly excluded from the refinement and later used to calculate R-free factor.[80] The target geometry parameters by Huber were used.[40] In the initial stages of refinement just torsion angles were refined and later the positional and individual temperature factors were refined. The final model was inspected and modified using the program CHAIN. Based on $2F_o$-F_c and F_o-F_c difference electron density maps, water molecules were added to the protein model. Two water molecules with lowest B-factor were interpreted as Sm^{+3} and Ca^{+2} ion sites. The final model consisting of protein residues 4–139, two cations, and 116 water molecules (R-free is 0.244 and standard crystallographic R-factor is 0.215) was deposited in the PDB with ID 1TVG. The X-ray data collection and refinement statistics are given in Table 1.

NMR data collection and NMR structure determination

All NMR data were collected at 298 K on 1.1 mM protein samples dissolved in 95% H_2O/5% D_2O solution containing 20 mM MES, 5 mM $CaCl_2$, 10 mM DTT, 0.02% NaN_3, at pH 6.5. Data were collected on Varian Inova 600 and 750-MHz spectrometers equipped with triple resonance gradient probes and a Varian Inova 600 with a cold probe. Spectra were process with NMRPipe [81] and analyzed with Sparky 3.110.[82] Backbone and sidechain chemical shifts were determined from 2D ^{1}H-^{15}N HSQC and ^{1}H-^{13}C HSQC, and 3D HNCO, HNCACB, CBCA(CO)NH, HNHA, (H)CC(CO)NH-TOCSY H(CC)(CO)NH-TOCSY, HCCH-COSY, H(C)CH-TOCSY, and (H)CCH-TOCSY spectra. NOESY peaks were picked in a ^{15}N-edited NOESY-HSQC (τ_m=100 ms) two ^{13}C-edited NOESY-HSQC (80 ms) optimized for either aliphatic or aromatic carbons. Additional NOEs were assigned from a 4D ^{13}C-^{13}C-HMQC-NOESY-HMQC (125 ms) recorded after lyophilization and exchange into 100% D_2O solution. All 2D and 3D pulse sequences were from the Varian BioPack library and the 4D NOESY was from Lewis Kay (University of Toronto). Stereospecific assignments of isopropyl methyl groups of Val and Leu residues were determined from the characteristic ^{1}H-^{13}C coupling in a high resolution ^{1}H-^{13}C HSQC of a [U-^{15}N, 5% ^{13}C]HSP034 sample.[83] Slowly exchanging amide protons were identified from a time-course analysis of 2D ^{1}H-^{15}N HSQC spectra recorded after exchange into D_2O. Dihedral restraints for φ and Ψ dihedral angles were derived from chemical shift data using the program TALOS (φ ± 40° and Ψ ± 50°).[84]

Resonance assignments, NOESY peak lists from four NOESY spectra, Talos derived dihedral restraints for 107 residues and a list of slowly exchanging amide protons (still observed after 1 hour) were used by the program AutoStructure version 2.1.1,[85] interfaced with Xplor-NIH [17, 86] to generate preliminary restraints and structures. AutoStructure generates restraints, including dihedral angle, NOE, and hydrogen bond distance restraints. These NOE distance restraints had uniform lower bounds of 1.8 Å and upper bounds of either 2.8, 3.2, 4.0 or 5.0, with all long-range NOEs and NOEs between side chains at 5.0 Å upper bounds. NOE assignments were examined and manually evaluated. Intermediate structures were used to identify consistently or egregiously violated NOEs, which were then subjected to manual assessment including nearby restraints to end up with the final NOE restraint list. Hydrogen bonds restraints were used for 31 slowly exchanging amide protons for which a CO backbone acceptor could be unambiguously identified from preliminary structures. Three restraints per hydrogen bonds were applied: H^N to O 1.7–2.3, N to O 2.7–3.2, H^N to C 2.8–3.4 Å. All final structure calculations used the same NOE, hydrogen bond, and dihedral restraints (with the exception of the Rosetta calculations). No pseudoatom corrections were used because sum averaging was used, with the exception of the DYANA calculations that treats pseudoatoms with center averaging (see below). All protocols took into account the *cis* Proline, P46. For the final NMR structure, 20 low energy structures calculated using the standard Xplor-3.84 routine sa.inp were used as input structures for a final refinement by restrained molecular dynamics in explicit water with CNS 1.1 using a standard protocol and deposited in the PDB with ID 1XPW.

Xplor

Xplor-NIH (version 2.15.0) [17, 86] software with the routines mkpsf.inp and generate_template.inp was used to generate starting structures followed by calculations with the simulated annealing protocol in the routine sa.inp. Starting from an extended structure, 130 structures were iteratively calculated and the first 20 structures with energies < 35 Kcal/mol were kept. Energies were typically < 35 Kcal/mol or > 4000 Kcal/mol. The topology and parameter files protein.top and protein.par designed to agree with bond lenths and angles from the CSDX force field.[40] The Xplor calculation has four stages (i) initial minimization, (ii) high-temperature torsional angle dynamics at 2,000K, (iii) cartesian dynamics cooling linearly 2,000 to to 100 K, (iv) final minimization. The number of steps in the high-temperature MD and cooling phases were 30,000 and 200,000. This was increased from the standard protocol values of 24,000 and 3,000, respectively, in order to improve convergence and accuracy as was also recently observed by Fossi.[87]

Xplor+

Each of the 20 Xplor structures calculated above with the standard protocol were further refined with an improved simulated annealing protocol that uses many of the updated features of Xplor-NIH. [69, 70] These include the IVM module for torsion angle and rigid body dynamics,[88] a radius of gyration term to represent the weak packing potential,[25] and database potentials of mean force to refine against Cα/Cβ chemical shifts,[23] multidimensional torsion angles [37, 38], and a backbone hydrogen bonding term.[68] The radius of gyration was applied to residues 2–138 with the target value of $2.2N^{0.38}$, where N is 137 residues.[25] The backbone hydrogen bonding term was used in free mode so that identification of backbone hydrogen bonding was fully automated without user input.[68]

CNSw

Each of the 20 Xplor structures calculated with the standard sa protocol were used as input for further refinement by restrained molecular dynamics in explicit water with CNS (version 1.1) [79] using a standard protocol.[46, 89] The protein structure is refined in a thin layer, 7 Å shell, of water molecules with a full MD force field that includes electrostatic and Lennard-Jones non-bonded potentials from the OPLS non-bonded parameters with slight modifications [39] in the PARALLHDG 5.3 force field. The protocol, starts with a energy minimization after addition of the explicit water, followed by slow heating from 100 to 500K, a dynamics refinement search at 500K, and slow cooling from 500 to 25K, with a final energy minimization at the end.[46, 89]

CNSw + Calcium

When Ca^{+2} was incorporated into the structure calculations, 6 additional restraints were added to coordinate the Ca^{+2} to six nearby O atoms identified from analysis of the X-ray structure (typically 2.4–2.7 Å). Coordination to the water molecule observed in the X-ray structure was not restrained. Structures were calculated using Xplor and further refined in explicit water in CNS, CNSw.

CYANA

Structure calculations were performed using CYANA version 2.0 with a standard simulating annealing protocol.[90, 91] The torsion-angle MD dynamics has high temperature phase at 9,600 K of 2,000 steps and a cooling phase to 0 K in 8,000 steps. Structure calculations of 100 structures were sorted by the value of their target function and the lowest 20 were kept.

DYANA

Structure calculations using DYANA (version 1.5) [90] used the same simulated annealing protocol as above for CYANA. The NOE restraints for the DYANA calculations had pseudoatom corrections for stereochemically ambiguous protons added to the upper bounds: 1.0 A for methylene protons, 2.0 A for chemically equivalent aromatic protons, and 2.4 Å for pairs of methyl groups in leucine and valine residues,[92] because DYANA treats these protons with center averaging in the calculation.

Rosetta refinement

Rosetta refinement of the X-ray structure of HSPC034 was performed by using the Rosetta all atom refinement method consisting of random backbone perturbations and sampling of discrete side chain rotamers, followed by a Monte-Carlo minimization. Protons and residues 1–3 and 140–143 were added to the X-ray coordinates before using this structure to seed the calculation of 1000 refined structures.

Rosetta refinement of the NMR structure of HSPC034 was similar to the X-ray structure refinement. For each of the 20 structures in the starting NMR ensemble, 1000 structures were calculated generating 20,000 structures and the 20 lowest energy structures were retained for analysis. Calculations took 2.5 days on 128 processors on the Miami University EM64T cluster. An additional calculation of 100,00 structures (5000 for each of the 20 starting structures) took 13.5 days on the cluster.

Idealized average structures

Idealized average structures were calculated using an Xplor standard protocol followed by idealization with the Rosetta forcefield. The structures were first converted to Xplor format using PDBStat. The average structures for each ensemble of 20 structures (residues 4–139) was generated using the Xplor protocol rmsd_ens.inp, which takes into account rotation of symmetric sidechains, to give the best average structure (and minimize the RMSD). The resulting PDB file was truncated to leave only residues 4–139 and was converted into Rosetta format and idealized using the commands: (1) rosetta -score -fa_input -s ave_4to139.pdb –nstruct 1 –use_pdb_numbering, (2) rosetta -pose_idealize -s ave_4to139_0001.pdb -fa_input -nstruct 1

Molecular Replacement Calculations

MOLREP (version 9.2.10) [93] and Phaser (version 1.3.2) [63, 64, 94] were used for molecular replacement studies within CCP4 (version 6.0).[95] In each case, the ordered residues (4-139) of the HSPC034 protein was used as a template. Each of 20 NMR structures calculated using DYANA, CYANA, XPLOR-NIH, CNS with water refinement (CNSw), CNSw with calcium (CNSwCa), and CNSw subjected to high-resolution Rosetta refinement, as well as regularized average structures from each calculation, were used as independent search models for molecular replacement. Each NMR structure was superimposed to the X-ray structure (1TVG) prior to MR, making it possible to assess the accuracy of each MR solution. For MOLREP calculations the X-ray data was cut off at 3.0 Å resolution and the rotation and translation functions were calculated using the default parameters. For PHASER calculations, the MR solution was obtained using the automated searching algorithm with the X-ray data cut off at 3.0 Å resolution (45.6 Å –3.0 Å) and a RMS difference of 1.5 Å (a value typically used for new MR searches. The molecular weight used in Phaser calculations was 17417.

Hydrogen Bond Analysis

Backbone hydrogen bond analysis of HSPC034 X-ray structure and NMR structures (residues 6–138) calculated with different refinement methods were analyzed by DSSP, dictionary of protein secondary structure,[96] and by rules designed to filter out bivalent hydrogen bonds with similar DSSP energies. This was done in order to examine those hydrogen bonds that are strong and unshared in a given structure and to facilitate comparison of hydrogen bonds patterns between different structures. Only hydrogen bonds with DSSP energies less than −0.5 Kcal/mol are considered. If two hydrogen bonds share the same donor or acceptor, then the stronger one is kept if it is significantly stronger that the other (|delta E| > 1 Kcal/mol) and the second isn't too strong (E > −1.5 Kcal/mol). Otherwise, the two hydrogen bonds are considered to be bivalent because they have similar energy or are both stronger than the threshold and so both are filtered out for future analysis.

Structural Assessment Software

The Rutgers protein structure validation server (PSVS, http://www-nmr.cabm.rutgers.edu/PSVS) [60] runs PROCHECK v.3.5.4 [44, 97] and MolProbity,[98] ProsaII,[99] Verify3D,[100] RPF,[29] and PDB validation software [101] as well as other validation software. PDBStat was used for RMSD to X-ray structures (http://www-nmr.cabm.rutgers.edu/NMRsoftware/nmr_software.html). PyMOL was used to create protein figures (http://www.pymol.org). RPF scores were calculated within AutoStructure 2.1.1 by comparison of structural ensembles to manually optimized NOESY peak lists from the ^{15}N- and ^{13}C-edited NOESY-HSQC spectra, as output by the program Sparky, using the chemical shifts in BMRB format. Match tolerances of 0.03 ppm for direct H, 0.05 for indirect H, and 0.5 ppm for C/N were used. Individual RPF-DP scores were calculated for truncated structures (residues 4–139) by using distances of the individual structures, in order to make a fair comparison with the X-ray coordinates, DP(each). The average DP(each) is reported along with error bars in Table 3. Average RPF-DP scores were calculated using average distances based on all 20 structures with full-length (residues 1–143) coordinates, to obtain the optimal DP-score, DP(ave).

Inductively Coupled Plasma Mass Spectrometry

A 50 µl sample of the 1.1 mM [*U*-^{15}N; 5%-^{13}C]HSPC034 NMR sample was analyzed by ICP-MS (Agilent Technologies 4500 ICP-MS) along with 40 ul of control sample buffer filtrate that was obtained by ultracentrifugation of 200 ul of the remaining sample (Amicon Microcon 3). Calcium concentrations were 5.7 mM and 4.5 mM for the sample and the control respectively, which demonstrates 1.1 mM bound calcium, and stoichiometric (1:1) binding.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J. Liu and B. Rost for providing HSPC034 as target HR1958 of the Northeast Structural Genomics Consortium, T. W. Wietsma for ICP-MS analysis, and R. Tejero for improvements in PSVS, A. Srinivasan of the Miami University Research Computing Support group for development of hydrogen bond perl scripts, G. M. Clore for providing the Xplor+ protocol scripts, as well as Y. J. Huang, and J. R. Cort for useful discussions. Acquisition and processing of NMR spectra and structure calculations were performed in the Environmental Molecular Sciences Laboratory, 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. Rosetta calculations were run on the Miami University Redhawk cluster.

References

1. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature. 2007; 450:259–264. [PubMed: 17934447]
2. Rossman MG, Blow DM. The detection of sub-units within the crystallographic asymmetric unit. Acta Crystallogr A. 1962; 15:24–31.
3. Rossman, MG. The molecular replacement method. New York, NY: Gordon and Breach, Science Publishers, Inc; 1972.
4. Brunger AT, Campbell RL, Clore GM, Gronenborn AM, Karplus M, Petsko GA, Teeter MM. Solution of a protein crystal structure with a model obtained from NMR interproton distance restraints. Science. 1987; 235:1049–1053. [PubMed: 17782253]
5. Baldwin ET, Weber IT, St Charles R, Xuan JC, Appella E, Yamada M, Matsushima K, Edwards BF, Clore GM, Gronenborn AM. Crystal structure of interleukin 8: Symbiosis of NMR and crystallography. Proc Natl Acad Sci U S A. 1991; 88:502–506. [PubMed: 1988949]
6. Chen YW. Solution solution: Using NMR models for molecular replacement. Acta Crystallogr D. 2001; 57:1457–1461. [PubMed: 11567160]
7. Chen YW, Clore GM. A systematic case study on using NMR models for molecular replacement: P53 tetramerization domain revisited. Acta Crystallogr D. 2000; 56:1535–1540. [PubMed: 11092918]
8. Chen YW, Dodson EJ, Kleywegt GJ. Does NMR mean "not for molecular replacement"? using NMR-based search models to solve protein crystal structures. Structure. 2000; 8:213–220.
9. Shaanan B, Gronenborn AM, Cohen GH, Gilliland GL, Veerapandian B, Davies DR, Clore GM. Combining experimental information from crystal and solution studies: Joint X-ray and NMR refinement. Science. 1992; 257:961–964. [PubMed: 1502561]

10. Janes RW, Peapus DH, Wallace BA. The crystal structure of human endothelin. Nat Struct Biol. 1994; 1:311–319. [PubMed: 7664037]

11. Muller T, Oehlenschlager F, Buehner M. Human interleukin-4 and variant R88Q: Phasing X-ray diffraction data by molecular replacement using X-ray and nuclear magnetic resonance models. J Mol Biol. 1995; 247:360–372. [PubMed: 7707380]

12. Wilmanns M, Nilges M. Molecular replacement with NMR models using distance-derived pseudo B factors. Acta Crystallogr D. 1996; 52:973–982. [PubMed: 15299607]

13. Arseniev AS, Kondakov VI, Maiorov VN, Bystrov VF. NMR solution spatial structure of ‘short’ scorpion insectotoxin I5A. FEBS Lett. 1984; 165:57–62.

14. Williamson MP, Havel TF, Wuthrich K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. J Mol Biol. 1985; 182:295–315. [PubMed: 3839023]

15. Clore GM, Gronenborn AM. New methods of structure refinement for macromolecular structure determination by NMR. Proc Natl Acad Sci U S A. 1998; 95:5891–5898. [PubMed: 9600889]

16. Clore GM, Schwieters CD. Theoretical and computational advances in biomolecular NMR spectroscopy. Curr Opin Struc Biol. 2002; 12:146–153.

17. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The xplor-NIH NMR molecular structure determination package. J Magn Reson. 2003; 160:65–73. [PubMed: 12565051]

18. Clore GM, Gronenborn AM. Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy. Crit Rev Biochem Mol Biol. 1989; 24:479–564. [PubMed: 2676353]

19. Garrett DS, Kuszewski J, Hancock TJ, Lodi PJ, Vuister GW, Gronenborn AM, Clore GM. The impact of direct refinement against three-bond HN-$C^{\alpha}H$ coupling constants on protein structure determination by NMR. J Magn Reson Ser B. 1994; 104:99–103. [PubMed: 8025816]

20. Osapay KA, Case DA. A new analysis of proton chemical shifts in proteins. J Am Chem Soc. 1991; 113:9436–9444.

21. Williamson MP, Asakura T. Empirical comparisons of models for chemical-shift calculation in proteins. J Magn Reson Ser B. 1993; 101:63–71.

22. Kuszewski J, Gronenborn AM, Clore GM. A potential involving multiple proton chemical-shift restraints for nonstereospecifically assigned methyl and methylene protons. J Magn Reson B. 1996; 112:79–81. [PubMed: 8661311]

23. Kuszewski J, Qin J, Gronenborn AM, Clore GM. The impact of direct refinement against $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts on protein structure determination by NMR. J Magn Reson Ser B. 1995; 106:92–96. [PubMed: 7850178]

24. Tjandra N, Garrett DS, Gronenborn AM, Bax A, Clore GM. Defining long range order in NMR structure determination from the dependence of heteronuclear relaxation times on rotational diffusion anisotropy. Nat Struct Biol. 1997; 4:443–449. [PubMed: 9187651]

25. Kuszewski J, Gronenborn AM, Clore GM. Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. J Am Chem Soc. 1999; 121:2337–2338.

26. Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH. Nuclear magnetic dipole interactions in field-oriented proteins: Information for structure determination in solution. Proc Natl Acad Sci U S A. 1995; 92:9279–9283. [PubMed: 7568117]

27. Tjandra N, Omichinski JG, Gronenborn AM, Clore GM, Bax A. Use of dipolar ^{1}H-^{15}N and ^{1}H-^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution. Nat Struct Biol. 1997; 4:732–738. [PubMed: 9303001]

28. Bewley CA, Gustafson KR, Boyd MR, Covell DG, Bax A, Clore GM, Gronenborn AM. Solution structure of cyanovirin-N, a potent HIV-inactivating protein. Nat Struct Biol. 1998; 5:571–578. [PubMed: 9665171]

29. Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): Structure quality assessment measures based on information retrieval statistics. J Am Chem Soc. 2005; 127:1665–1674. [PubMed: 15701001]

30. Linge JP, Nilges M. Influence of non-bonded parameters on the quality of NMR structures: A new force field for NMR structure calculation. J Biomol NMR. 1999; 13:51–59. [PubMed: 10905826]

31. Billeter M, Qian YQ, Otting G, Muller M, Gehring W, Wuthrich K. Determination of the nuclear magnetic resonance solution structure of an antennapedia homeodomain-DNA complex. J Mol Biol. 1993; 234:1084–1093. [PubMed: 7903398]

32. Prompers JJ, Folmer RH, Nilges M, Folkers PJ, Konings RN, Hilbers CW. Refined solution structure of the Tyr41-->His mutant of the M13 gene V protein. A comparison with the crystal structure. Eur J Biochem. 1995; 232:506–514. [PubMed: 7556200]

33. Kordel J, Pearlman DA, Chazin WJ. Protein solution structure calculations in solution: Solvated molecular dynamics refinement of calbindin D_*9k*. J Biomol NMR. 1997; 10:231–243. [PubMed: 9390401]

34. Xia B, Tsui V, Case DA, Dyson HJ, Wright PE. Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized born model, and with explicit water. J Biomol NMR. 2002; 22:317–331. [PubMed: 12018480]

35. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M. Refinement of protein structures in explicit solvent. Proteins. 2003; 50:496–506. [PubMed: 12557191]

36. Kuszewski J, Gronenborn AM, Clore GM. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci. 1996; 5:1067–1080. [PubMed: 8762138]

37. Kuszewski J, Gronenborn AM, Clore GM. Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. J Magn Reson. 1997; 125:171–177. [PubMed: 9245376]

38. Kuszewski J, Clore GM. Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force. J Magn Reson. 2000; 146:249–254. [PubMed: 11001840]

39. Jorgensen WJ, Tirado-Rives J. The OPLS potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc. 1988; 110:1657–1666.

40. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr A. 1991; 47:392–400.

41. Allen FH, Bellard S, Brice MD, Cartwright BA, Doubleday A, Higgs H, Hummelink T, Hummelink-Peters BG, Kennard O, Motherwell WDS, Rodgers JR, Watson DG. The cambridge crystallographic data centre: Computer-based search, retrieval, analysis and display of information. Acta Crystallog B. 1979; 35:2331–2339.

42. Hendrickson WA. Stereochemically restrained refinement of macromolecular structures. Methods Enzymol. 1985; 115:252–270. [PubMed: 3841182]

43. Vriend G. WHAT IF: A molecular modeling and drug design program. J Mol Graphics. 1990; 8:52–56.

44. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. J Biomol NMR. 1996; 8:477–486. [PubMed: 9008363]

45. Nabuurs SB, Nederveen AJ, Vranken W, Doreleijers JF, Bonvin AM, Vuister GW, Vriend G, Spronk CA. DRESS: A database of REfined solution NMR structures. Proteins. 2004; 55:483–486. [PubMed: 15103611]

46. Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CA, Nabuurs SB, Guntert P, Livny M, Markley JL, Nilges M, Ulrich EL, Kaptein R, Bonvin AM. RECOORD: A recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins. 2005; 59:662–672. [PubMed: 15822098]

47. Chen J, Won HS, Im W, Dyson HJ, Brooks CL 3rd. Generation of native-like protein structures from limited NMR data, modern force fields and advanced conformational sampling. J Biomol NMR. 2005; 31:59–64. [PubMed: 15692739]

48. Lee SY, Zhang Y, Skolnick J. TASSER-based refinement of NMR structures. Proteins. 2006; 63:451–456. [PubMed: 16456861]

49. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with rosetta in CASP6. Proteins. 2005; 61 (Suppl 7):128–134. [PubMed: 16187354]

50. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science. 2003; 302:1364–1368. [PubMed: 14631033]

51. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci U S A. 2006; 103:5361–5366. [PubMed: 16567638]

52. Rohl CA. Protein structure estimation from minimal restraints using rosetta. Methods Enzymol. 2005; 394:244–260. [PubMed: 15808223]

53. Rohl CA, Baker D. De novo determination of protein backbone structure from residual dipolar couplings using rosetta. J Am Chem Soc. 2002; 124:2723–2729. [PubMed: 11890823]

54. Misura KM, Baker D. Progress and challenges in high-resolution refinement of protein structure models. Proteins. 2005; 59:15–29. [PubMed: 15690346]

55. Monleon D, Colson K, Moseley HN, Anklin C, Oswald R, Szyperski T, Montelione GT. Rapid analysis of protein backbone resonance assignments using cryogenic probes, a distributed linux-based computing architecture, and an integrated set of spectral analysis tools. J Struct Funct Genomics. 2002; 2:93–101. [PubMed: 12836666]

56. Moseley HN, Sahota G, Montelione GT. Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. J Biomol NMR. 2004; 28:341–355. [PubMed: 14872126]

57. Szyperski T, Yeh DC, Sukumaran DK, Moseley HN, Montelione GT. Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. Proc Natl Acad Sci U S A. 2002; 99:8009–8014. [PubMed: 12060747]

58. Chandonia JM, Brenner SE. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. Proteins. 2005; 58:166–179. [PubMed: 15521074]

59. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B. Automatic target selection for structural genomics on eukaryotes. Proteins. 2004; 56:188–200. [PubMed: 15211504]

60. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. Proteins. 2007; 66:778–795. [PubMed: 17186527]

61. Gaskell A, Crennell S, Taylor G. The three domains of a bacterial sialidase: A β-propeller, an immunoglobulin module and a galactose-binding jelly-roll. Structure. 1995; 3:1197–1205. [PubMed: 8591030]

62. Wendt KS, Vodermaier HC, Jacob U, Gieffers C, Gmachl M, Peters JM, Huber R, Sondermann P. Crystal structure of the APC10/DOC1 subunit of the human anaphase-promoting complex. Nat Struct Biol. 2001; 8:784–788. [PubMed: 11524682]

63. Read RJ. Pushing the boundaries of molecular replacement with maximum likelihood. Acta Crystallogr D. 2001; 57:1373–1382. [PubMed: 11567148]

64. Storoni LC, McCoy AJ, Read RJ. Likelihood-enhanced fast rotation functions. Acta Crystallogr D. 2004; 60:432–438. [PubMed: 14993666]

65. Anderson DH, Weiss MS, Eisenberg D. A challenging case for protein crystal structure determination: The mating pheromone er-1 from *euplotes raikovi*. Acta Crystallogr D. 1996; 52:469–480. [PubMed: 15299668]

66. Fleming PJ, Rose GD. Do all backbone polar groups in proteins form hydrogen bonds? Protein Sci. 2005; 14:1911–1917. [PubMed: 15937286]

67. Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV. Comparison of X-ray and NMR structures: Is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? Proteins. 2005; 60:139–147. [PubMed: 15856480]

68. Grishaev A, Bax A. An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. J Am Chem Soc. 2004; 126:7281–7292. [PubMed: 15186165]

69. Legler PM, Cai M, Peterkofsky A, Clore GM. Three-dimensional solution structure of the cytoplasmic B domain of the mannitol transporter IImannitol of the escherichia coli phosphotransferase system. J Biol Chem. 2004; 279:39115–39121. [PubMed: 15258141]

70. Cai M, Huang Y, Suh JY, Louis JM, Ghirlando R, Craigie R, Clore GM. Solution NMR structure of the barrier-to-autointegration factor-emerin complex. J Biol Chem. 2007; 282:14525–14535. [PubMed: 17355960]

71. Lipsitz RS, Sharma Y, Brooks BR, Tjandra N. Hydrogen bonding in high-resolution protein structures: A new method to assess NMR protein geometry. J Am Chem Soc. 2002; 124:10621–10626. [PubMed: 12197765]

72. Gsponer J, Hopearuoho H, Cavalli A, Dobson CM, Vendruscolo M. Geometry, energetics, and dynamics of hydrogen bonds in proteins: Structural information derived from NMR scalar couplings. J Am Chem Soc. 2006; 128:15127–15135. [PubMed: 17117864]

73. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. 1997; 276:307–326.

74. Terwilliger TC, Berendzen J. Automated MAD and MIR structure solution. Acta Crystallogr D. 1999; 55:849–861. [PubMed: 10089316]

75. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. HKL-3000: The integration of data reduction and structure solution--from diffraction images to an initial model in minutes. Acta Crystallogr D. 2006; 62:859–866. [PubMed: 16855301]

76. Terwilliger TC. Maximum-likelihood density modification using pattern recognition of structural motifs. Acta Crystallogr D. 2001; 57:1755–1762. [PubMed: 11717487]

77. Sack J. CHAIN - a crystallographic modeling program. J Mol Graphics. 1997; 15:132–134.

78. Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A. 1991; 47:110–119. [PubMed: 2025413]

79. 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. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D. 1998; 54:905–921. [PubMed: 9757107]

80. Brunger AT. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. Nature. 1992; 355:472–475. [PubMed: 18481394]

81. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. J Biomol NMR. 1995; 6:277–293. [PubMed: 8520220]

82. Goddard, TD.; Kneller, DG. Sparky 3. <http://www.cgl.ucsf.edu/home/sparky>

83. Neri D, Szyperski T, Otting G, Senn H, Wuthrich K. 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 ^{13}C labeling. Biochemistry. 1989; 28:7510–7516. [PubMed: 2692701]

84. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR. 1999; 13:289–302. [PubMed: 10212987]

85. Huang YJ, Tejero R, Powers R, Montelione GT. A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins. 2006; 62:587–603. [PubMed: 16374783]

86. Schwieters CD, Kuszeski JJ, Clore GM. Using xplor-NIH for NMR molecular structure determination. Prog Nucl Mag Res Sp. 2006; 48:47–62.

87. Fossi M, Oschkinat H, Nilges M, Ball LJ. Quantitative study of the effects of chemical shift tolerances and rates of SA cooling on structure calculation from automatically assigned NOE data. J Magn Reson. 2005; 175:92–102. [PubMed: 15949752]

88. Schwieters CD, Clore GM. Internal coordinates for molecular dynamics and minimization in structure determination and refinement. Journal of Magnetic Resonance. 2001; 152:288–302. [PubMed: 11567582]

89. Linge JP, Habeck M, Rieping W, Nilges M. ARIA: Automated NOE assignment and NMR structure calculation. Bioinformatics. 2003; 19:315–316. [PubMed: 12538267]

90. Guntert P, Mumenthaler C, Wuthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol. 1997; 273:283–298. [PubMed: 9367762]

91. Guntert P. Automated NMR structure calculation with CYANA. Methods Mol Biol. 2004; 278:353–378. [PubMed: 15318003]

92. Wuthrich, K. NMR of proteins and nucleic acids. New York: Wiley; 1986.

93. Vagin A, Teplyakov A. MOLREP: An atomated program for molecular replacement. J Appl Crystallog. 1997; 30:1022–1025.

94. McCoy AJ, Grosse-Kunstleve LC, Read RJ. Likelihood-enhnaced fast translation functions. Acta Crystallogr D. 2005; 61:458–464. [PubMed: 15805601]

95. Collaborative Computational Project, Number 4. The CCP4 suite: Programs for protein crystallography. Acta Crystallogr D. 1994; 50:760–763. [PubMed: 15299374]

96. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]

97. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: A program to check the stereochemical quality of protein structures. J Appl Crystallog. 1993; 47:283–291.

98. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by C^{α} geometry: Φ,Ψ and C^{β} deviation. Proteins. 2003; 50:437–450. [PubMed: 12557186]

99. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins. 1993; 17:355–362. [PubMed: 8108378]

100. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: Assessment of protein models with three-dimensional profiles. Methods Enzymol. 1997; 277:396–404. [PubMed: 9379925]

101. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]

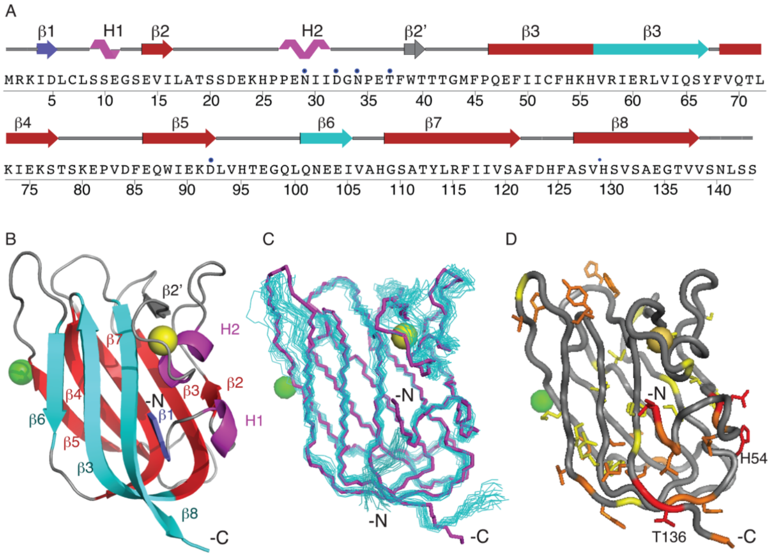

Figure 1.

Structure of HSPC034, (A) Secondary structure superimposed on sequence (adapted from PDBsum, http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/). Residues that coordinate metal ions are marked with a blue dot. (B) Ribbon representation of X-ray structure of HSPC034 for deposited coordinates, residues 4–139. The Calcium ion is shown in yellow and a Samarium (III) ion in green. (C) Backbone atoms for 20 NMR structures optimally superimposed with respect the N, Cα, and C′ coordinates of the X-ray structure residues 6–138. NMR residues 2–141 are shown. (D) NOE violations indicated on X-ray structure. Red violations are > 2 Å, orange are 1–2 Å, and yellow are 0.5–1 Å. Violations are not show for residues 1–3 and 140–141. Figures B, C, and D were generated using PyMOL (DeLano Scientific).

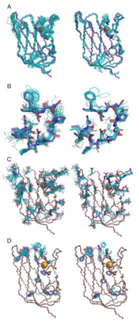

Figure 2.
Comparison of CNSw refined NMR structures (left) to Rosetta refined NMR structures (right). The ensemble of 20 structures is shown as lines for (A) backbone atoms, (B) calcium-binding loop, residues 28–38, (C) side chains for residues DNEQ, and (E) side chains for residues WYF. In all cases, the structures are superimposed on the X-ray structure shown with thick lines.

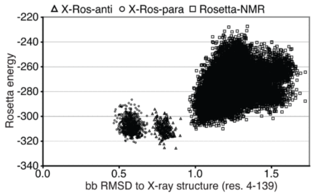

Figure 3.

A) Rosetta all atom energy versus backbone RMSD (residues 4–139) for NMR Rosetta refined structures (squares), and X-ray Rosetta refined structures with parallel (circles) and antiparallel (triangles) β1-strand structures.

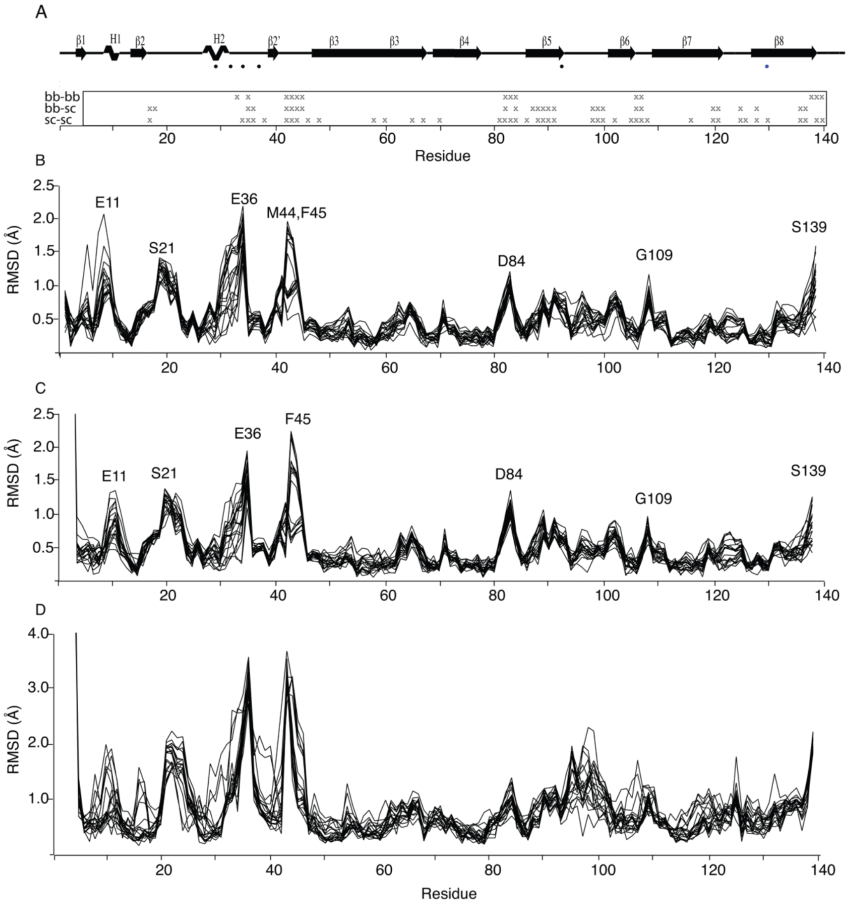

Figure 4.

(A) Residues that have nearby symmetry related molecules in the X-ray structure are marked with an x for backbone-backbone (bb-bb), backbone-sidechain (bb-sc) and sidechain-sidechain (sc-sc) close contact less than 5 Å. The secondary structure schematic with the metal coordinating residues marked with a filled circle shown above. (B) Backbone RMSD per residue for idealized X-ray structure vs. the X-ray structure after backbone superposition for residues 6–138 for Xros-para, (C) for Xros-anti (D) for Rosetta refined CNSw NMR structures.

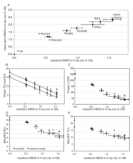

Figure 5.

(A) Backbone RMSD (x-axis) versus heavy atom RMSD (y-axis) for residues 6–138. (B–F) Molecular replacement solution metrics (y-axis) versus backbone RMSD for residues 6–138 (x-axis) for Phaser (B) TFZ, triangles, and RFZ, circles, (C) LLG (R) squares, MOLREP (D) RF/s and (E) TF/s. From left to right, the structures are: X-ray, XRos-anti, XRos-para, Rosetta, CNSwRosHB (+RosHBs), CNSw, Xplor, Cyana, Dyana represented with circles. All error bars are one standard deviation for the ensemble. Idealized-average structures for the ensemble are shown with open symbols or Xs. For the X-ray structure, the idealized structure is indicated. Error bars are one standard deviation for the ensembles and are not shown in (B) for the x-axis for clarity.

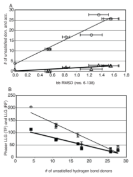

Figure 6.

A) Average number of unsatisfied hydrogen bond donors (open circles) and acceptors (open triangles) versus the backbone RMSD to the X-ray structure (residues 6-138) for each ensemble of 20 structures as in Figure 2. (B) Phaser LLG (RF) (filled squares) and LLG (TF) (open circles) vs. the average number of unsatisfied hydrogen bond donors in each ensemble of 20 structures.

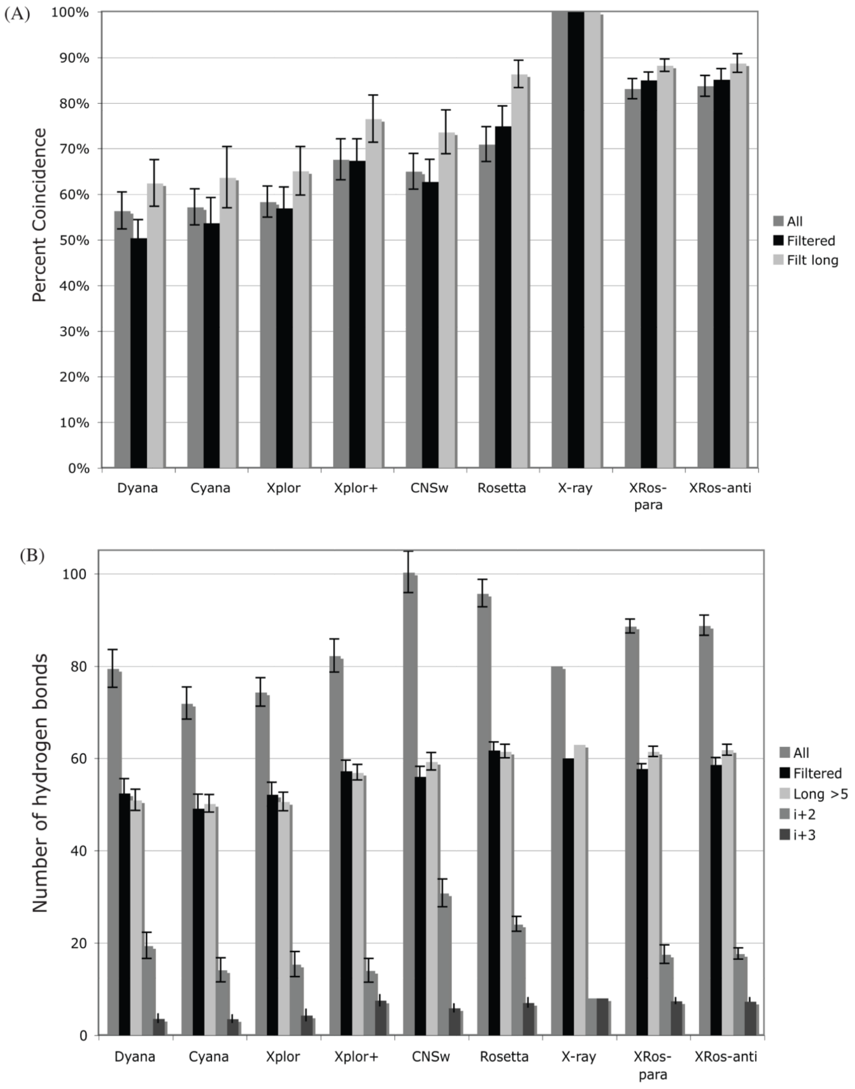

Figure 7.

(A) Percent coincidence for all hydrogen bonds (All), filtered to remove bivalent hydrogen bonds (Filtered), and long-range (| j | > 5) filtered hydrogen bonds (Filt long). Coincidence is (number in both)/(number in X-ray only + number in NMR only + both). (B) Counts of hydrogen bonds in each ensemble of 20 structures (1 structure for X-ray) for all hydrogen bonds, filtered to remove bivalent hydrogen bonds, all long range (HN_i to CO_{i-j}, | j | > 5) (, i +2 (j = 2), and i+3 (j = 3) hydrogen bonds.

Table 1

Human HSPC034 X-ray data collection and refinement statistics.

Data Processing and Refinement Statistics				
X-ray source	X4A			
Temperature (K)	100			
Data	SeMet+Sm	SeMet		
Wavelength (Å)	0.97896	0.97949	0.97917	0.97239
	Se peak	Se edge	Se peak	Se remote
Space group	C2	C2	C2	C2
Cell dimensions (Å) and angles (°) a	70.974	73.245	73.318	73.832
b	41.617	42.335	42.351	42.646
c	46.779	47.253	47.267	47.598
β	102.19	102.714	102.711	102.712
Number of molecules in the asymmetric unit	1	1	1	1
SeMAD data statistics				
Resolution (Å)	30.0–1.4	30.0–1.4	30.0–1.4	30.0–1.4
Number observed reflections	175535	160779	184924	161697
Number unique reflections [a]	48903	47579	47766	47917
Completeness (%)	94.5 (60.0)	86.8 (44.5)	89.4 (52.5)	85.7 (37.6)
R_{merge} (%)	7.1 (40.7)	5.9 (35.3)	6.3 (37.7)	5.8 (37.8)
<I/σ(I)>	22.8 (5.4)	18.1 (1.9)	19.2 (2.0)	18.2 (1.6)
Summary of Structure Quality Statistics				
Resolution limits (Å)	30.0–1.6			
Number of unique reflections F≥1σ(F)	32,294			
Completeness (%)	93.5 (84.5)			
R_{cryst} (%)[a,b]	21.5 (22.7)			
R_{free} (%)[a,c]	24.4 (28.0)			
Number of protein atoms	1084			
Number of protein residues	136			
Number of water molecules	116			
Number of ions	2			
RMSD from ideal geometry				
bond length (Å)	0.005			
bond angles (°)	1.30			
Averaged B value (Å^2)	14.70			
Ramachandran plot summary from PROCHECK[97]				
most favored	85.4%			
allowed	13.8%			
generously allowed	0.8%			
disallowed	0.0%			
Structure quality factors generated using PSVS-1.3 [60]			Mean score	Z-score

Procheck G-factor (φ/Ψ) [c]	−0.5	−1.7
Procheck G-factor (all dihedral angles) [c]	−0.3	−1.5
Verify3D	0.5	0.2
ProsaII (−ve)	−1.5	−1.5
MolProbity clashscore	14.6	−1.0

[a] The value in parentheses are for the highest resolution shell 1.60 Å – 1.70 Å for the refinement, and 1.40 Å – 1.45 Å for data processing

[b] $R_{cryst} = \Sigma hkl \, ||F_o| - |F_c|| / \Sigma hkl \, |F_o|$, whrere F_o and F_c are the observed and calculated structure factors, respectively.

[c] R_{free} is computed for 10% reflections randomly selected and omitted from the refinement.

Table 2

Statistics for human HSPC034 NMR structure determination.

Structure Calculation Statistics		
Completeness of resonance assignments for residues 1–143		
backbone	97.6%	
side chains [a]	92.5%	
Conformationally-restricting NOE restraints		
intra-residue [i = j]	2	
sequential [$\lvert i - j\rvert = 1$]	160	
medium range [$1 < \lvert i - j\rvert < 5$]	121	
long range [$\lvert i - j\rvert \leq 5$]	640	
total	923	
NOE restraints per residue [b]	6.6	
Dihedral angle restraints		
total	210	
φ	103	
Ψ	107	
Hydrogen bond restraints		
total (3 per hydrogen bond)	93	
long range [$\lvert i – j\rvert \leq 5$]	93	
Total number of conformationally-restricting restraints	1206	
Number of restraints per residue [b]	8.5	
Number of long-range restraints per residue [b]	5.0	
Number of structures calculated	20	
Number of structures used	20	
Structure Validation Statistics		
Distance violations/structure		
> 0.1 Å	0	
RMSD of distance violation/restraint	0.002 Å	
maximum distance violation	0.04 Å	
Dihedral angle violations/structure		
> 0.001°	0	
RMSD of dihedral angle violation/restraint	0.05°	
maximum dihedral angle violation	0.80°	
Average RMSD to the average structure	residues 4–139	ordered residues [c]
backbone atoms (N, Cα, C′)	0.7 + 0.1 Å	0.7 + 0.1 Å
heavy atoms	1.2 + 0.1 Å	1.2 + 0.1 Å
RMSD from ideal geometry		
bond length (Å)	0.004	
bond angles (°)	0.6	
Ramachandran plot summary from PROCHECK[44]		

most favored	84.8 %	86.9%
additionally allowed	13.1 %	12.4%
generously allowed	1.3 %	0.7%
disallowed	0.8 %	0.0%
Structure quality factors generated using PSVS-1.3 [60].	Mean score	Z-score
Procheck G-factor (φ/Ψ) [c]	−0.6	−2.2
Procheck G-factor (all dihedral angles) [c]	−0.5	−2.8
Verify3D	0.4	−1.3
ProsaII (−ve)	0.3	−1.7
MolProbity clashscore	22.7	−2.4
RPF R/P/DP scores [d]	0.90/0.88/0.78	

[a] Lys NH_3^+, Arg NH_2, Cys SH, Ser/Thr OH, Pro N, N-terminal NH_3^+, C-terminal carbonyl, sidechain carbonyl and aromatic quaternary carbons were not considered to be routinely assignable resonances.

[b] For 140 residues with conformationally-restricting NOE restraints, residues 2–141.

[c] Ordered residues ranges: 4–16, 21–31, 33–67, 70–139, with the sum of φ and Ψ order parameters > 1.8.

[d] RPF scores defined in reference,[29] calculated for ensemble residues 1–143. RPF-DP score is DP(ave.)

Table 3

Structure quality validation Z-scores for HSPC034 using PSVS [60] for residues 4–139. RMSD to the X-ray structure is given for backbone (bb) N, C^{α}, C′ and all heavy atoms (hv) for residues 6–138. Standard deviations are given in parenthesis.

	RMSD (6-138)		Procheck		MolProbity	Unsatisfied H-bond		DP(each)[b]	DP(ave)[c]
	bb	**hv**	**φ-Ψ**	**all**	**Clashscore**	**donors**	**acceptors** [a]	**(4–139)**	**(1–143)**
X-ray	0	0	−1.7	−1.5	−1.0	3	1	0.64	–
X-Ros-Anti	0.3 (0.04)	1.2 (0.04)	−1.4	−0.3	0.7	10.9	1.9	0.67 (0.01)	0.74
X-Ros-Para	0.6 (0.03)	1.2 (0.03)	−1.5	−0.5	0.7	10.8	1.6	0.67 (0.01)	0.74
Rosetta	0.8 (0.1)	1.6 (0.01)	−1.5	−0.4	0.8	16.8	2.0	0.65 (0.01)	0.75
CNSw61HB	0.9 (0.1)	1.7 (0.1)	−2.4	−2.8	−2.6	13.9	0.8	0.64 (0.02)	0.78
CNSwRosHB	1.0 (0.1)	1.8 (0.1)	−2.4	−3.0	−2.3	17.0	1.6	0.64 (0.02)	0.78
CNSwCa	1.0 (0.1)	1.8 (0.1)	−2.4	−3.0	−2.6	17.8	1.1	0.64 (0.01)	0.78
CNSw	1.2 (0.2)	2.0 (0.2)	−2.5	−3.0	−2.4	17.3	1.1	0.63 (0.02)	0.78
Xplor+	1.2 (0.1)	2.0 (0.1)	−2.0	−2.1	−2.6	21.9	2.4	0.62 (0.02)	0.77
Xplor	1.4 (0.2)	2.2 (0.1)	−3.3	−5.9	−4.2	27.7	3.0	0.62 (0.01)	0.75
CYANA	1.6 (0.1)	2.4 (0.1)	−3.2	−5.4	−1.0	25.7	2.7	0.61 (0.02)	0.76
DYANA	1.6 (0.1)	2.4 (0.1)	−3.4	−5.9	−3.5	25.7	2.7	0.62 (0.02)	0.77
Idealized average structures									
X-ray (ideal)	0.3	0.6	−2.0	−2.1	−0.9				
X-Ros-Anti	0.5	1.1	−1.4	−0.9	0.0				
X-Ros-Para	0.5	1.2	−1.5	−1.0	0.1				
Rosetta	0.7	1.5	−1.3	−1.1	−0.2				
CNSwCa	0.8	1.6	−2.1	−3.9	−5.7				
CNSw	0.9	1.7	−1.9	−3.7	−0.8				
Xplor	1.1	1.9	−2.1	−4.8	−1.0				
CYANA	1.2	2.1	−1.7	−4.5	−0.5				
DYANA	1.2	2.0	−1.9	−4.9	−1.2				

[a] Calculated with WHAT IF [43]

[b] Individual RPF-DP, DP(each), scores calculated for truncated structures (residues 4–139) with average and standard deviation reported.

[c] Average RPF-DP, DP(ave), scores were calculated using all 20 full-length structures.