Protein structure determination by combining sparse NMR data with evolutionary couplings

Yuefeng Tang[1,2,7], Yuanpeng Janet Huang[1,2,7], Thomas A Hopf[3,4], Chris Sander[5,8], Debora S Marks[3,8] & Gaetano T Montelione[1,2,6,8]

Accurate determination of protein structure by NMR spectroscopy is challenging for larger proteins, for which experimental data are often incomplete and ambiguous. Evolutionary sequence information together with advances in maximum entropy statistical methods provide a rich complementary source of structural constraints. We have developed a hybrid approach (evolutionary coupling–NMR spectroscopy; EC-NMR) combining sparse NMR data with evolutionary residue-residue couplings and demonstrate accurate structure determination for several proteins 6–41 kDa in size.

Solution-state NMR spectroscopy can generally provide accurate three-dimensional (3D) structures of small (molecular weight (MW) < ~15 kDa) proteins[1,2]. For larger proteins, however, broad line widths and resonance overlap make structure determination by NMR spectroscopy challenging. Perdeuteration[3,4], in which most ^{1}H nuclei are replaced with ^{2}H nuclei, using biosynthetic methods, generally increases the sensitivity and feasibility of NMR spectroscopy studies of larger proteins by decreasing the nuclear relaxation rates of the remaining ^{1}H, ^{15}N and ^{13}C nuclei[3]. Perdeuterated proteins provide good quality, but less complete, NMR data[3–5]. Structures generated with such 'sparse NMR data' are generally less accurate than those obtained for smaller proteins, for which all ^{1}H sites can be detected, complete backbone and side-chain resonance assignments can be determined, and extensive and accurate NMR restraints can be derived. Improved methods are therefore needed in order to enable structural biologists to routinely use sparse NMR data to generate accurate models of larger (i.e., 15 kDa to ~60 kDa) protein structures.

As a result of recent advances in sequencing technology and computational biology, complementary information about 3D structures can be obtained from evolutionary residue-residue couplings computed from multiple alignments of structurally related protein sequences. Such evolutionary couplings (ECs), derived from evolutionary correlated mutations using global statistical models and entropy maximization, provide accurate information about residue pair contacts[6–11], as the highest scoring ECs are between residues that are close in the 3D structure[6,7,12]. Contact restraints derived from ECs can be combined with molecular modeling methods to provide 3D structures of proteins[6,8,9,13]. However, the derived restraints, by definition, are an average over all 3D structures of the proteins in the multiple sequence alignment (MSA; i.e., the protein subfamily or family) and do not necessarily reflect the intricate details of residue interactions in any particular protein in the MSA data set. In addition, even when there is extensive sequence information, residue-residue contacts indicated by high-ranked ECs may contain false positives. Even partial experimental information about a particular protein can therefore be used to increase the atomic position accuracy of 3D structures computed from sequence information.

Here we describe a hybrid approach for protein structure determination, which can be used to mitigate both the sparseness of experimental NMR data and the accuracy limitations of structure modeling by evolutionary constraints, by providing more complete and accurate residue-pair contact information than either method alone. The method, outlined in **Figure 1**, involves simultaneous analysis of ECs, derived from MSAs, with NMR chemical shift, nuclear Overhauser effect (NOE) and residual dipolar coupling (RDC) data. The process rules out false positive ECs and assigns NOE spectroscopy (NOESY) cross-peaks while generating 3D structure models. We provide a detailed description of the EC-NMR method (**Supplementary Figs. 1** and **2**), together with specific protocols (Online Methods).

We tested the overall performance using experimental sparse NMR data for eight proteins ranging in size from 6 kDa to 41 kDa (**Table 1** and **Supplementary Tables 1–3**). These EC-NMR structures used backbone H^{N}, Cα, C′ and side-chain Cβ (and in some cases side-chain amide and methyl) resonance assignments, sparse NOESY-based restraint densities (0.09–2.0 long-range ($|i - j| > 5$) NOE restraints per residue), and backbone ^{15}N-^{1}H RDC data (**Supplementary Table 3**), together with EC restraints.

We compared the resulting EC-NMR structures with known 'reference structures', determined either by X-ray crystallography

[1]Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA. [2]Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA. [3]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [4]Department of Informatics, Technische Universität München, Garching, Germany. [5]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [6]Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA. [7]These authors contributed equally to this work. [8]These authors jointly supervised the work. Correspondence should be addressed to C.S. (ecnmr.authors@gmail.com), D.S.M. (debbie@hms.harvard.edu) or G.T.M. (gtm@rutgers.edu).

RECEIVED 6 JANUARY; ACCEPTED 26 MAY; PUBLISHED ONLINE 29 JUNE 2015; DOI:10.1038/NMETH.3455

Table 1 | Experimental data and comparisons of EC-NMR structures with benchmark reference structures

Protein name and UniProt identifier	*N*; MW (kDa)[a]	NOE data[b]	^{15}N-^{1}H RDC data[c]	Sequences in MSA[d]	r.m.s. deviation[o] (Å) relative to reference: N, Cα, C′, O backbone; all C, N, O, S atoms	PDB ID and method of structure determination
Smaller (<~15 kDa)						
Agrobacterium tumefaciens protein of unknown function A9CJD6_AGRTT5	64; 6.3	H^N-H^N only	None	10,962	1.5; 2.0[e] (1.5; 1.8[e])	2K2P NMR
Erwinia carotovora cold-shock-like protein Q6D6V0_ERWCT	66; 7.3	H^N-H^N only	2 alignment tensors	4,410	2.2; 3.0[f]	2K5N NMR
Arabidopsis thaliana ubiquitin-like domain Q9ZV63_ARATH	84; 9.7	H^N-H^N only	2 alignment tensors	4,964	1.9; 2.5[g]	2KAN NMR
Ralstonia metallidurans Rmet5065 Q1LD49_RALME	134; 15.0	H^N-H^N only	1 alignment tensor	2,620	2.0; 3.0[h] (2.0; 3.0[h])	2LCG NMR
Escherichia coli lipoprotein YiaD YIAD_ECOLI	141; 15.0	H^N-H^N only	2 alignment tensors	10,296	1.7; 2.3[i]	2K1S NMR
Larger (>~15 kDa)						
Homo sapiens H-ras oncogene protein p21 RASH_HUMAN	166; 18.9	H^N-H^N only	None	6,669	2.6; 3.6[j] (1.6; 2.6[j])	5P21 X-ray
Synechocystis Slr1183 P74712_SYNY3	194; 21.3	H^N-H^N, Me-Me, H^N-Me	2 alignment tensors	45,708	2.1; 3.0[k]	3MER X-ray
Escherichia coli maltose-binding protein	370; 40.7	H^N-H^N, Me-Me, H^N-Me	1 alignment tensor	12,416		
NTD (1–112; 259–329)					1.6; 2.4[l] (1.6; 2.5[l])	1DMB X-ray
CTD (113–258; 330–370)					1.9; 2.7[m] (1.9; 2.7[m])	1DMB X-ray
Full-length (1–370)					2.8; 3.4[n] (2.2; 2.8[n])	1DMB X-ray

[a]Number of residues (*N*) and MW of the protein construct studied by NMR spectroscopy, excluding affinity purification tags. [b]H^N-H^N NOESY cross-peak data include NOE data between backbone and side-chain amide H^N resonances. For P74712_SYNY3 and MBP, additional H^N-Me NOESY cross-peak data obtained for uniformly ^{15}N,^{13}C,^{2}H-enriched samples with $^{13}CH_3$ labeling of Ile(δ), Leu and Val methyls were also included. As only restraint lists are available for H-Ras oncogene protein p21, RASH_HUMAN, NOESY peak lists were back-calculated from the experimental NMR spectroscopy restraint list (2LCF) and chemical shift data (Biological Magnetic Resonance Data Bank (BMRB) ID 17610). [c]All experimental ^{15}N-^{1}H RDC data were measured in the laboratory of J. Prestegard. [d]Number of nonredundant sequences in MSA used to generate ECs. [e]Residue range for superimpositions and r.m.s. deviation calculations: 2–63. [f]Residue range for superimpositions and r.m.s. deviation calculations: 1–64. [g]Residue range for superimpositions and r.m.s. deviation calculations: 7–78. [h]Residue ranges for superimpositions and r.m.s. deviation calculations: 1–29, 36–58 and 62–135. [i]Residue ranges for superimpositions and r.m.s. deviation calculations: 15–39, 41–76, 79–120 and 127–141. [j]Residue ranges for superimpositions and r.m.s. deviation calculations: 1–29, 39–60 and 64–166. [k]Residue ranges for superimpositions and r.m.s. deviation calculations: 20–37, 41–134, 147–172 and 185–196. Residues 1–15 and 175–183 are not observed in the crystal structure. [l]Residue ranges for superimpositions and r.m.s. deviation calculations: 2–12, 14–112 and 259–329. [m]Residue ranges for superimpositions and r.m.s. deviation calculations: 115–117, 125–142, 144–172, 175–218, 221–227, 247–258 and 330–370. Interfacial residues 233–240 were exchange-broadened, precluding NMR spectra assignments. The sugar-binding site of MBP (1DMB) includes residues: K42, D65, E111, E153, Y155, E172, W230, W340 and R344. [n]Residue ranges for superimpositions and r.m.s. deviation calculations: 2–12, 14–112, 259–329, 115–117, 125–142, 144–172, 175–218, 221–227, 247–258 and 330–370. Interfacial residues 233–240 are exchange-broadened, precluding NMR spectra assignments. [o]Values in parentheses were calculated using additional simulated RDC data, as described in the text.

or by NMR spectroscopy using extensive backbone and side-chain ^{1}H, ^{13}C and ^{15}N resonance assignments (**Table 1**, **Fig. 1** and **Supplementary Fig. 3**). To assess the accuracy of these EC-NMR 3D structures we used three metrics: (i) accuracy of atomic positions, (ii) accuracy of the residue-pair contacts used to generate the structures, (iii) accuracy of side-chain χ_1 rotamer states for well-defined (i.e., converged), buried (i.e., not on the protein surface) side chains.

Relative to the known reference structures, the EC-NMR structures had accurate backbone and all-heavy-atom positions in six of eight proteins studied; i.e., <2 Å backbone atom positional r.m.s. deviations and <3 Å all-heavy-atom r.m.s. deviations relative to the reference structure (**Table 1**, **Fig. 1** and **Supplementary Figs. 4** and **5**). The remaining two proteins, human p21 H-Ras and maltose-binding protein (MBP) have no or limited RDC data, respectively, but their EC-NMR structures are nevertheless reasonably accurate; both proteins have backbone r.m.s. deviations <2.8 Å and all-heavy-atom r.m.s. deviations <3.6 Å relative to the reference structures. MPB consists of two structural domains. Considered separately, its two individual domains are even more accurate when compared to the reference X-ray crystal structure (N-terminal domain and C-terminal domain backbone r.m.s. deviation, 1.6 Å and 1.9 Å; and all-heavy-atom r.m.s. deviation 2.4 Å and 2.7 Å, respectively; **Table 1** and **Supplementary Fig. 6**) than is apparent from rigid-body superimposition for the entire protein. The difference in the accuracy of the individual domains relative to the whole MBP protein is likely due to its well-known interdomain flexibility[14].

For all eight proteins studied, the final residue pair contact list generated by the ASDP program had higher coverage of the short-distance contacts in the reference structure, lower false positive rate, higher precision and more long-range residue-residue contacts than either the initial EC list or the sparse NMR data alone (**Figs. 1**, **2a**, **Supplementary Note 1**, **Supplementary Fig. 7** and **Supplementary Table 4**). These results demonstrate that the EC-NMR method provides more accurate and complete structural contact information than is obtained using ECs or sparse NMR data alone.

We also compared the χ_1 side-chain dihedral angles for buried residues with well-defined atomic coordinates across the conformers of the NMR spectroscopy ensemble. Averaged over all eight EC-NMR structures, ~80% of these side chains had χ_1 rotamers matching corresponding reference structures (**Supplementary Table 5**). For the three largest proteins studied, 85%, 81% and 65% of these side chains have χ_1 rotamers that match the corresponding X-ray crystal structures (**Fig. 2b**,**c**, **Supplementary Fig. 8** and **Supplementary Table 5**).

We further compared the accuracy of EC-NMR structures relative to previously published NMR structures determined with more extensive side-chain resonance assignments (**Fig. 2d**). For p21 H-Ras (where no side-chain methyl NMR data were used in the EC-NMR calculations) the side-chain structure accuracy is similar to that of the published NMR structure PDB ID 2LCF[15], which had been determined using essentially complete side-chain resonance assignments obtained on a fully protonated sample. For MBP, the core side-chain accuracy of the EC-NMR structure was much better than PDB ID 1EZP, determined using similar sparse NMR data together with five

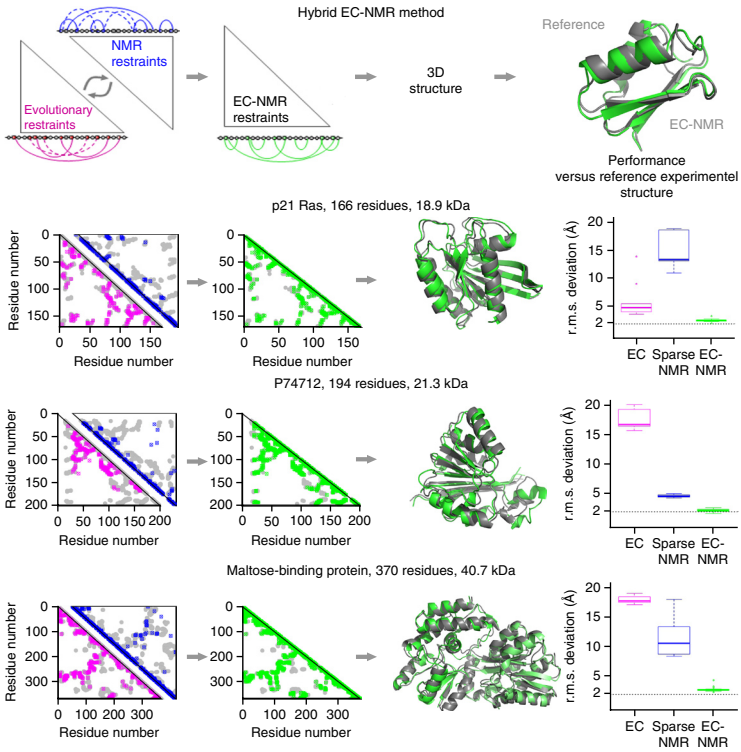

Figure 1 | The EC-NMR approach. EC information is interpreted together with ambiguous NOESY peak list data (top). Inconsistent ECs (dashed magenta contacts), NOESY noise peaks (dashed blue contacts) and ambiguous assignments of NOESY cross peaks (dotted blue contacts) are identified and/or resolved, and additional residue-pair contacts consistent with the NOE and EC data are discovered. Performance was assessed by comparing the resulting EC-NMR structure (green) with a reference X-ray crystal or NMR structure (gray). Below, the process of EC-NMR analysis using sparse NMR data for three example proteins with MW of 19–41 kDa is illustrated. Magenta contacts, initial EC residue-pair contacts. Blue contacts, contacts indicated by unambiguous NOESY peak assignments obtained by the ASDP program[18,19]. Green contacts, final residue pair contacts (RPCs) resulting from simultaneous analysis of EC and NMR data. Gray contacts, contacts in the reference X-ray crystal structure. Green ribbon structures, final EC-NMR structures. Gray ribbons, reference X-ray crystal structures. Box plots show the r.m.s. deviation to reference structures for backbone atoms of structures generated with EC data alone (magenta), sparse NMR data alone (blue) and the hybrid EC-NMR method (green). In box plots, the box in the middle indicates quartiles and median scores; the whiskers show the largest and smallest observation that falls within a distance of 1.5 times the nearest quartile; any additional points are shown as outliers. The EC-NMR protocol provides structures with backbone accuracy of ~2 Å (dashed gray line) relative to the corresponding X-ray crystal structures.

kinds of RDC data[4]. It is also similar to that of the solution NMR structure PDB ID 2D21, which had been determined using extensive stereospecific side-chain resonance assignments provided by the sophisticated and expensive stereo-arrayed isotope labeling method[16]. Additional backbone RDC data, calculated from the reference structure as described in Online Methods, further improved the accuracy of these EC-NMR structures (**Table 1**, in parentheses, and **Fig. 2d**).

To assess the robustness and sensitivity of the EC-NMR method to the amount of available sequence data, we computed ECs for randomly sampled subsets (50%, 25% and so on to 0.01%) of the full MSA for protein P74712 (194 residues; 21.2 kDa). The 19 subsets ranged in size from ~44,000 to 8 effective number of sequences (N_{eff}). We used ECs from these subsets for EC-NMR calculations (**Supplementary Fig. 9** and **Supplementary Table 6**). For this particular protein, the EC-NMR method breaks down at N_{eff}/L of ~5, where L is the length of the protein; for larger sequence alignments (N_{eff} >~1,000) the backbone positional r.m.s. deviations between EC-NMR models and the X-ray crystal structure were consistently below ~3.5 Å (**Supplementary Note 2**). The more evolutionary sequence information is available, the better the resulting structures.

It is critical to have metrics to assess the reliability of EC-NMR structure models in the absence of a reference structure. For conventional NMR structures, methods are available to discriminate correct from incorrect models[17]. These include the NMR discriminating power (DP) score[18,19], which tests for consistency of the structural models with the NOESY peak list data, and knowledge-based structure quality scores, which compare structural features (for example, backbone and side-chain dihedral angle distributions, core atom packing, etc.) with those observed in high-resolution X-ray crystal structures. We assessed whether these metrics can also discriminate 'reliable' (backbone r.m.s. deviation < 3.5 Å from the reference structure) from less accurate EC-NMR structures. Structure quality metrics were computed using various software packages integrated under the Protein Structure Validation Server (PSVS)[17]. DP scores range from 0 to 1, with higher values indicating better agreement between the model and the NMR spectroscopy data. Each of the knowledge-based structure quality scores were reported by PSVS as statistical Z scores relative to a collection of high-resolution X-ray crystal structures[17]; better structure quality scores have more positive Z scores. These metrics can be used to distinguish between EC-NMR models of protein P74712 generated with varying amounts of sequence data, with better scores for structures generated using more sequence information (**Supplementary Fig. 9**). These metrics also score the reference X-ray crystal and NMR structures used in this study as 'reliable' structures, and identify the models generated using ECs or sparse NMR data alone as 'less accurate' structures (**Supplementary Fig. 10**). From this analysis, we concluded that EC-NMR structures are 'reliable' if they had NMR DP scores[18,19] greater than ~0.73, and knowledge-based Z scores computed with the PSVS server[17] more positive than (i.e., greater than) $Z = -2$.

Our study demonstrates the complementary value of evolutionary sequence information and sparse NMR data for protein structure determination. The experimentally reliable, but ambiguous, contact information in sparse NOESY data can rule out ECs that are not relevant to the structure of the specific target protein (for example, those arising from oligomer interfaces) (**Supplementary Note 3**), and the ECs provide information about residue-residue contacts not contained in or incompletely covered by the NOESY and RDC data. The largely automated EC-NMR method delivers structures of perdeuterated, selectively protonated proteins with atomic positions comparable in accuracy to those in NMR structures obtained with complete side-chain assignments and/or sophisticated side-chain labeling methods.

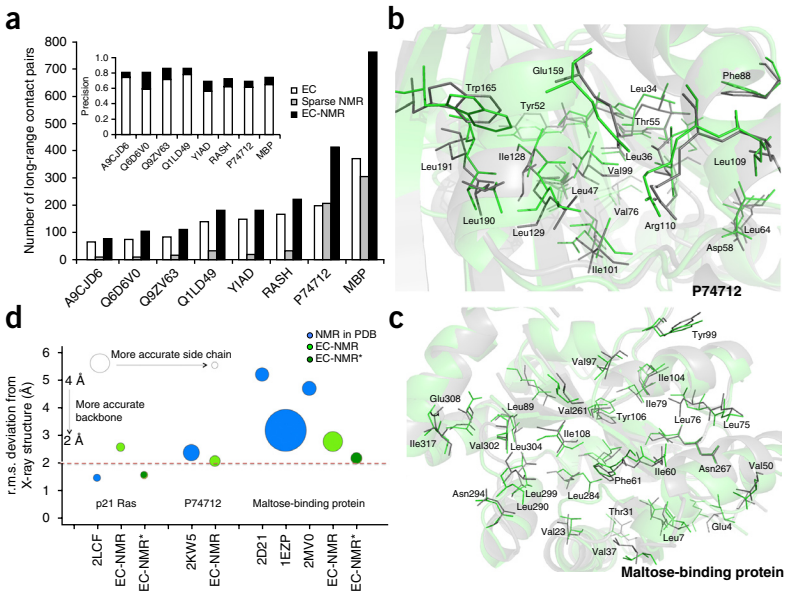

Figure 2 | Performance of the EC-NMR method. (**a**) Number of long-range residue pair contacts (i.e., between residue pairs (*i*, *j*) where $|i - j| \geq 5$) for the initial EC list (white histograms), the initial unambiguous sparse NOESY data (gray), and the final EC-NMR residue contact list (black). For smaller (<150 residues) proteins, the NMR spectroscopy data include only H^N-H^N NOE data, whereas for larger proteins (>150 residues) the NMR spectroscopy data also include NOE data to Val, Leu, and Ile(δ) methyl protons. Inset, precision of contacts, relative to the corresponding reference structures, is higher for final residue-pair contact list (solid) than for the initial EC list (open), as false positives are identified and removed by the EC-NMR algorithm. (**b**,**c**) Comparison of buried side-chain conformations in EC-NMR structures (green) and the corresponding X-ray crystal structure (gray). (**d**) Comparison of backbone r.m.s. deviation and buried side-chain χ_1 rotamers, relative to crystal structures. EC-NMR structures were determined using exclusively the experimental NMR spectroscopy data (no RDC data for p21 H-Ras, two RDC alignment tensors for P74712, and one RDC alignment tensor for MBP, light green). Results obtained after adding additional RDC data calculated from the reference structure are also shown for comparison (EC-NMR*, two hydrodynamic alignments of p21 H-Ras, or a second hydrodynamic alignment for MBP, dark green). The size of the circles corresponds to the percentage of core side-chains with χ_1 rotamers different from that observed in the crystal structure; smaller circles indicate a better match of side-chain conformations to the crystal structure.

For small proteins and domains up to ~140 residues (under ~15 kDa) with extensive sequence information, EC-NMR is a powerful and efficient approach for protein structure determination. It can be particularly valuable for determining structures of proteins for which backbone assignments can be determined, but for which poor signal-to-noise ratio makes extensive side-chain assignments difficult or impossible. For larger proteins, in the size range of 180–500 residues (20–60 kDa), ECs can be combined with sparse NMR data obtained on perdeuterated, selectively protonated protein samples to provide structures that are more accurate and complete than those obtained using such sparse NMR data alone. The EC-NMR method should also be valuable for determining NMR structures of membrane proteins, which typically rely on perdeuterated protein samples, and in protein structure determination by solid-state NMR spectroscopy methods. This advance expands the range of proteins for which accurate structures can be determined using either evolutionary coupling analysis or sparse NMR data alone.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Protein Data Bank: EC-NMR restraint lists and atomic coordinates have been deposited under accession codes 2N4C, 2N4D, 2N49, 2N4F, 2N4A, 2N4B, 2N48, 2N42, 2N46, 2N47, 2N44 and 2N45.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all of the members of the Northeast Structural Genomics Consortium who generated and archived NMR spectroscopy data used in this work, particularly scientists in the laboratories of C. Arrowsmith, M. Kennedy, G.T.M., T. Szyperski and J. Prestegard. We thank J. Aramini, G. Liu, G.V.T. Swapna, H. Valafar, M. Nilges and F. Xu for helpful discussions. This work was supported by grants from the US National Institutes of Health grant 1R01-GM106303 to C.S. and D.S.M. and Protein Structure Initiative grant U54-GM094597 to G.T.M.

AUTHOR CONTRIBUTIONS

Y.T., Y.J.H., T.A.H., C.S., D.S.M. and G.T.M. designed the research. Y.J.H. wrote ASDP program code. Y.T., Y.J.H., T.A.H. and D.S.M. performed calculations. Y.T., Y.J.H., T.A.H., C.S., D.S.M. and G.T.M. analyzed data. Y.T., Y.J.H., T.A.H., C.S., D.S.M. and G.T.M. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Mao, B., Guan, R. & Montelione, G.T. *Structure* **19**, 757–766 (2011).
2. Mao, B., Tejero, R., Baker, D. & Montelione, G.T. *J. Am. Chem. Soc.* **136**, 1893–1906 (2014).
3. Gardner, K.H., Rosen, M.K. & Kay, L.E. *Biochemistry* **36**, 1389–1401 (1997).
4. Mueller, G.A. *et al. J. Mol. Biol.* **300**, 197–212 (2000).
5. Rosen, M.K. *et al. J. Mol. Biol.* **263**, 627–636 (1996).
6. Marks, D.S. *et al. PLoS ONE* **6**, e28766 (2011).
7. Morcos, F. *et al. Proc. Natl. Acad. Sci. USA* **108**, E1293–E1301 (2011).
8. Hopf, T.A. *et al. Cell* **149**, 1607–1621 (2012).
9. Marks, D.S., Hopf, T.A. & Sander, C. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
10. Hopf, T.A. *et al. eLife* **3**, e03430 (2014).
11. Ovchinnikov, S., Kamisetty, H. & Baker, D. *eLife* **3**, e02030 (2014).
12. Sulkowska, J.I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J.N. *Proc. Natl. Acad. Sci. USA* **109**, 10340–10345 (2012).
13. Nugent, T. & Jones, D.T. *Proc. Natl. Acad. Sci. USA* **109**, E1540–E1547 (2012).
14. Evenas, J. *et al. J. Mol. Biol.* **309**, 961–974 (2001).
15. Araki, M. *et al. J. Biol. Chem.* **286**, 39644–39653 (2011).
16. Kainosho, M. *et al. Nature* **440**, 52–57 (2006).
17. Bhattacharya, A., Tejero, R. & Montelione, G.T. *Proteins* **66**, 778–795 (2007).
18. Huang, Y.J., Powers, R. & Montelione, G.T. *J. Am. Chem. Soc.* **127**, 1665–1674 (2005).
19. Huang, Y.J., Rosato, A., Singh, G. & Montelione, G.T. *Nucleic Acids Res.* **40**, W542–W546 (2012).

ONLINE METHODS

General description of the EC-NMR method. The EC-NMR method involves three steps (**Supplementary Fig. 1**). Step A provides predicted residue-pair contacts from sequence information. Evolutionary couplings are calculated for the protein from a MSA of the protein family. Ideally, the protein sequence alignment required for the calculation is centered around the protein of interest and has a carefully chosen range of evolutionary neighbors: not too many, so as to optimize specificity of structural constraints, and not too few, so as to retrieve as many sequences as possible and thus reduce sampling bias. The specificity-sensitivity trade-off is managed in part by limiting the number of gaps allowed in the columns of the MSA, which tend to increase with evolutionary distance. A maximum entropy model of the protein sequences, constrained by the amino acid residue pair frequencies observed in the MSA, is used to remove the confounding effect of transitive correlations and thus reduce the number of false positive (FP) predicted inter-residue contacts, which would result from the application of local mutual information methods. In the current implementation, the interaction parameters in the model, i.e., the evolutionary residue-reside couplings, are computed using pseudo-likelihood maximization in the EVcouplings[9] computational procedure.

Step B acquires sparse NMR data from protein samples in solution, using ^{13}C,^{15}N–enriched and/or ^{2}H,^{13}C,^{15}N–enriched samples prepared with ^{1}H-^{13}C labeling of side chain Leu, Val, and Ile(δ) methyl groups[3,5,20]. Sequence-specific resonance assignments are determined for backbone $^{1}H^{N}$, ^{13}C and ^{15}N resonances, as well as for side-chain $^{13}C\beta$ and amide $^{1}H^{N}$-^{15}N resonances. For larger proteins, some methyl $^{13}CH_3$ resonance assignments are also required. NOESY peak lists are then generated from simultaneous 3D ^{15}N,^{13}C-NOESY spectra and ^{15}N-^{1}H residual dipolar coupling (RDC) data are measured using one or more hydrodynamic alignment media (referred to here as RDC hydrodynamic alignment tensors). Such 'sparse NMR data' can generally be obtained for perdeuterated proteins with molecular weights as large as 40–60 kDa[21–23], and have been used in exceptional cases to determine chain folds for proteins as large as 82 kDa[24,25].

Step C identifies and iteratively refines residue-pair contact distance restraints using both sources of information, and determines a small set of accurate 3D structures. Chemical shift, NOESY peak list, EC and RDC data are interpreted together to determine NOESY cross-peak assignments, rule out FP ECs and to generate initial 3D models of the protein. This automated combined analysis of NMR spectroscopy and EC data, ruling in ambiguous NOESY cross-peak assignments and ruling out FP EC contacts, is done using the NOESY assignment program ASDP[26]. Intermediate 3D structures are generated from these combined NMR and evolutionary distance constraints using the program CYANA[27]. The resulting residue-pair contacts, derived by the combined analysis of EC and NMR spectroscopy data, are then deconvoluted into atom-specific distance restraints, which are used to refine the protein structure by restrained energy minimization. In the current implementation, this refinement step uses the program Rosetta[2,28].

Alignments for generation of evolutionary couplings. MSAs were generated for each of the eight target proteins using the jackhmmer algorithm[29] for different sequence alignment depths, following a search of the UniProt database of protein sequences for potential homologs. The depth of the specific MSA used for each protein was chosen based on a minimum coverage of the protein for the maximum number of sequences. In the current implementation, minimum coverage is defined as no more than 10% of columns in the alignment with more than 50% gaps across the set of all sequences. Sequence fragments of less than 70% of the full length of the search protein were removed, and sequences with more than 70% identity were down-weighted, as previously described[6,8].

Calculation of evolutionary couplings. ECs were calculated using the EVfold-plm pipeline available at http://evfold.org, as described elsewhere[8]. For structure modeling using ECs alone, secondary structure prediction clashes with EC pairs were removed from the restraint list[8]. EC score files for each protein used in this study are available at http://ec-nmr.nesg.org/.

Implementation of the EC-NMR method in the ASDP automated NOESY cross-peak assignment program. The EC-NMR method has been implemented within the automated NOESY cross-peak assignment program ASDP[27]. This version of ASDP (version 2.0), along with specific instructions for EC-NMR analysis including the specific parameters used in this study, are available from http://ec-nmr.nesg.org/.

The five major steps of the iterative EC-NMR analysis process are outlined in **Supplementary Figure 2**.

Step 1. Initial NOE-based distance restraints are generated from NOESY and chemical shift data using algorithms encoded in the ASDP program[26]. Secondary structures, including beta-strand alignments, are identified using previously described algorithms[26], based on the chemical shift index method[30], together with characteristic secondary-structure NOE patterns[31]. Additional NOE assignments are ruled in and ruled out using the ASDP software[26], based on uniqueness relative to the chemical shift list, NOESY cross-peak symmetry patterns, and the network anchoring algorithms of the ASDP program, as described[26]. The cutoffs used in the EC-NMR analysis for identifying beta sheets are different from the cutoffs used for conventional NOESY analysis[26], because backbone H^{α}-H^{α} and H^{N}-H^{α} NOEs are missing from sparse NMR data sets. When using the subset of H^{N}-H^{N} NMR spectroscopy data available for fully protonated proteins, no other parameters were changed for ASDP analysis. For perdeuterated proteins, a deuterium correction to the ^{13}C chemical shifts[32] is applied in ASDP automatically, and longer distance cutoffs (up to 6 Å) are used for the NOEs since such interactions are often observable for longer distances in such perdeuterated proteins.

EC-based RPC ambiguous restraints were generated as follows: ambiguous distance restraints (≤5 Å) are generated between every two carbon atoms (C_i, C_j) for each residue pair (i, j) in the EC list. For each protein, the number of EC pairs used as input to the ASDP calculations was L, the number of residues in the target protein sequence (excluding any purification tags). ECs are ranked based on EC reliability scores[8]. The weights (w) are initially set to $w = 1.0$ for the first $L/2$ ECs on the EC list, and $w = 0.5$ for the second $L/2$ ECs in the list.

Step 2. One hundred decoy models were generated using the noeassign module of the program CYANA[27], with 3D H^{N}-H^{N} NOESY peak list data, ^{1}H-^{15}N RDC values (if available), dihedral angle restraints generated from backbone chemical shift data

using Talos+[33], together with unique NOE-based distance constraints identified by ASDP and L EC-based inter-residue ambiguous distance constraints from Step 1. In this process, CYANA provides analysis of ambiguous restraints for unassigned NOESY cross-peaks. For larger (>20 kDa) perdeuterated proteins, NMR spectroscopy data also included 3D H^N-Me and Me-Me NOESY peak list data, which provide NOE data involving Val $^{13}C\gamma H_3$, Leu $^{13}C\delta H_3$ and Ile $^{13}C\delta H_3$ methyl groups. Stereospecific assignments of Val and Leu isopropyl groups were not included in the chemical shift lists.

The standard protocol of the Talos+ program was used to generate backbone dihedral angle restraints based on $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, and residues with 'good' Talos+ scores (i.e., Talos+ reliability score of 10) were restrained to ϕ and ψ ranges of ±20° (ref. 33). A deuterium correction to ^{13}C chemical shifts was also applied in Talos+ calculations for perdeuterated proteins[33].

Step 3. The top 20 decoy models from CYANA are identified using a combined score comprised of the NMR RPF Recall[18] score and CYANA target function. The NMR RPF Recall score measures the fraction of NOESY cross-peaks that can be explained by the decoy model structure. These 3D decoy structures are then used to rule in and rule out potential NOE and EC assignments using the ASDP program.

Structurally inconsistent NOE assignments from step 1 are excluded as described in the published description of the ASDP algorithm[26]. Structurally inconsistent RPCs (referred to here as SI-RPCs) are identified when ambiguous RPC distance restraints are violated by >0.5 Å in more than 60% (for example, 12 of 20) conformers. These SI-RPCs are excluded from the next cycle of ASDP calculations.

Using these decoy models, standard rules of ASDP are then used to make new NOESY cross-peak assignments, which are added as unique restraints to the distance restraint list as input for the iterative run of step 2.

Ambiguous RPC restraints that are satisfied by all 20 decoy conformers (i.e., no violation > 0.5 Å) are reassigned weight w = 1.0. No changes are made for the remaining RPCs, which have small violations among the 20 conformers. All RPCs are then again defined as ambiguous distance restraints, between all C atoms of residue i and all C atoms of residue j.

In addition, the ASDP software also identifies new RPCs, which are long-range residue pairs (i.e., $|i-j| \geq 5$) that have at least one inter-atom (i.e., any H, N or C atom with a resonance assignment) distance ≤5 Å apart in all 20 conformers. These RPCs are added to the EC-NMR restraint list as ambiguous distance restraints between all C atoms of residue i and all C atoms of residue j, with weight w = 1.0. These RPCs based on intermediate structures often, but not always, correspond to EC pairs with low ranking scores in the covariance analysis.

Steps 2–3 are then repeated two more cycles, resulting in an ensemble of 20 protein structure models (incrementing the cycle count: cycle = 3 in **Supplementary Fig. 2**).

Step 4. The protein structure models from cycle 3 are then used to identify NOE peaks and RPCs that are inconsistent with these intermediate structures. These 'noise' data are then removed from the input data, and steps 1–3 are then repeated again. The parameter run is incremented.

Using intermediate structures to clean up the *de novo* initial distance restraints helps to regenerate better conformers for subsequent restrained-energy optimization. 'Noise NOESY cross-peaks' are defined as all NOESY cross-peaks with initial NOE assignments from step 1 for which the corresponding restraint is violated by >10 Å in all 20 conformers from cycle 3. 'Noise ECs' are initial ECs from step 1 for which the corresponding ambiguous restraint is violated with distance >10 Å in all 20 conformers from cycle 3. These 'noise' NOESY cross-peaks and ECs are removed from the EC-NMR restraint list.

Step 5. The resulting 20 NMR structure models are further energy refined using a standard restrained Rosetta refinement protocol[2]. Specific atom-atom Rosetta refinement restraints were generated for each atom pair in residue pairs in the EC list, which have minimal (over all atoms in the side chains) residue-residue interatomic distance ≤5 Å in all 20 models. Upper-bound restraints of 7 Å are used for all of these specific inter-residue atom-atom restraints, in order to allow the Rosetta force field to attain low-energy structures and to avoid generating overly constrained structures.

The variables cycle and run are used here to control the repeated analyses of steps 1–3. These parameters are defined in **Supplementary Figure 2**. When the process begins, cycle is set to 0 and run is set to 1. After steps 2–3 are repeated for 3 cycles, step 4 is executed. If any 'noise NOEs' and/or 'noise ECs' are identified, steps 1–3 are repeated again. The iterative process ends with run = 2. No further runs are then executed to avoid potential overfitting.

Tutorial for EC-NMR calculations. A web-based tutorial for running EC-NMR calculations is available at http://ec-nmr.nesg.org/tutorial.html. The tutorial includes sample input and output data files. A step-by-step process is also provided below.

EC pairs are generated from sequence data. EC pairs can be calculated using the EVfold-plm pipeline available at http://evfold.org/. ECs can also be identified using alternative software implemented subsequent to the original EVfold process, including PSICOV[34], GREMLIN[11,35] or other methods[12,36,37], although these methods have not been tested here. EC pairs are sorted based on the coupling scores and the top L EC pairs with highest coupling scores are used.

Resonance assignment table. The NMR resonance assignment table is prepared in either BMRB 2.x or 3.x format[38]. The ASDP software does interpret the ambiguity code column, which should be correctly prepared, as these data are needed for denoting stereospecific assignments of Leu and Val isopropyl methyl groups and individual assignments of side amide hydrogens.

NOESY peak lists. Peak lists are generated from 2D, 3D, 4D and/or pseudo4D NOESY data using standard automated peak picking programs, and generally should be manually edited to eliminate obvious noise peaks. These peak lists are prepared in X-Easy format[39]. For pseudo 4D NOESY data[40], the pseudo chemical shifts for the indirect proton dimension should be labeled as 999 in the peak list.

Backbone dihedral angle restraints. Dihedral angle restraints may be generated automatically from backbone chemical shift using TALOS-N (ref. 41) (or TALOS+[33]), or defined by alternative automated and/or manual methods. When using the ASDP program, dihedral angle restraints should be prepared in Cyana format. For perdeuterated samples, the talosn command shall use [–iso] to provide appropriate deuterium correction to chemical shifts. The Talos2dyana.com script from the TalosN package can be used to generate restraints in Cyana format for EC-NMR calculations.

doi:10.1038/nmeth.3455

Residual dipolar coupling data. Residual dipolar coupling data should be provided in the table format outlined in sample data available on the EC-NMR website (http://ec-nmr.nesg.org/tutorial.html). The RDC list supports multiple interatomic vectors in multiple media, including N-H, N-CA (intra) and N-C′ (sequential) vectors with error and weight factors. The RDC file shall also provide the D_a (magnitude) and R (Rhombicity) notation typical of programs such as PALES[42] and ReDCat[43].

Parameter table for ASDP. When using the ASDP program, the par.tbl parameter table from the sample data should be used as the default parameter table.

Control file. For each project, ASDP requires a control file which specifies the protein name, sequences, input files and instructions to the program on how to run structure calculations. An example control file is provided with sample data. The flag EC = <EC pairs> should be included in the control file. The tolerance for the pseudo proton should be set as 999 in the control-file.

Generation of EC NMR structures with ASDP. The ASDP software, together with a short tutorial, is available at: http://www-nmr.cabm.rutgers.edu/NMRsoftware/asdp/Quick_Starts.html. Additional instructions for using ASDP are at: http://www.nmr2.buffalo.edu/nesg.wiki/AutoStructure_Structure_Determination_Program. The ASDP commands used to run EC-NMR calculations are in **Supplementary Note 4**.

Refinement of EC NMR structures with Rosetta. ASDP can use various programs to generate 3D structures from the NOESY-based distance restraints that the program derives from the NOESY peak and chemical shift lists. For EC-NMR calculations, the program has been most thoroughly tested using CYANA for structure generation. Each of the resulting NMR structure models are then further energy refined using the restrained Rosetta refinement protocol outlined in ref. 2. Detailed protocols for Restrained Rosetta refinement are available at http://www.nmr2.buffalo.edu/nesg.wiki/Rosetta_High_Resolution_Protein_Structure_Refinement_Protocol.

The script getCC.pl in the ASDP-2.0 package is used to generate specific atom-atom Rosetta refinement restraints for each atom pair in residue pairs of EC list, which have minimal interatomic distance ≤5 Å in all 20 models. Upper-bound restraints of 7 Å are used for all of these specific atom-atom constraints. The input files for the getCC.pl script are the PDB file of the final models (<proteinName>.pdb in the final ASDP cycle) and the final EC pairs (<proteinName >.ec in the final ASDP cycle). The resulting output file final.upl is then used for restrained Rosetta refinement, as described elsewhere[2]. The distance upper bounds are loosened by 30% before converting to the Rosetta restraint format. This can be done using a standalone version of Rosetta or, alternatively, using the Restrained Rosetta Refinement server[2] available at: http://psvs-1_4-dev.nesg.org/consRosetta.html.

Identification of high-confidence EC pairs. To assess the confidence of EC pairs computationally, we follow, in the current implementation, the approach introduced in more detail in ref. 10 that measures how much each EC score is an outlier from the distribution of non-informative background couplings between the majority of positions. Based on the approximately symmetrical distribution of background coupling scores around 0, we estimate the level of background noise from the absolute value of the most negative EC score. The reliability score $Q(i, j)$ of an EC score $EC(i, j)$ is then calculated by measuring how far it exceeds the level of background noise

$$Q(i,j)=\frac{EC(i,j)}{\left|\min_{i,j}\left(EC(i,j)\right)\right|}$$

This measure depends solely on the shape of the EC scores distribution and has been shown to be a useful predictor for the accuracy of ECs[10]. For the purpose of this work, we define high-confidence ECs as all pairs with $Q(i, j)$ >2, i.e., couplings that exceed the background noise by a factor of at least 2 (**Supplementary Fig. 11**). We refer to this as the number of reliable EC pairs ($N_{reliable}$). Python code to identify high-confidence ECs is in **Supplementary Note 5**. For each of the 19 randomly generated MSAs for the protein P74712 (194 residues; 21.2 kDa), as described in the main text, we predicted $N_{reliable}$ based on a score threshold that is determined solely on the statistics of the distribution of the EC coupling scores, using no information about the structure. The EC-NMR method failed (backbone r.m.s. deviation > 3.5 Å to the reference structure) for $N_{reliable}$ <~25 (**Supplementary Fig. 9**) and this can be used as guidance for minimal requirements of sequence information for successful application of the EC-NMR.

Assessment of structure reliability. One of the metrics used in protein NMR structure validation is an analysis of restraint violations interpreted from the NMR spectroscopy data; i.e., how well the model fits to derived restraint data. Low restraint violations is a necessary, but not sufficient, condition for validating a distance restraint–derived structure when the restraints themselves may be misinterpreted[17]. Other metrics used for NMR model validation include knowledge-based scores (for example, Molprobity[44], ProCheck[45], ProsaII[46] and Verify3D[47]), which assess how well the structure fits with the known conformational features of proteins, such as the dihedral angle and structure packing distributions observed in high-resolution X-ray crystal structures. Using statistics normalized to a set of high-resolution crystal structures, computed with the Protein Structure Validation Server (PSVS), it has been demonstrated that accurate conventional NMR structures have Z scores more positive than $Z = -2$ to -3 for these structure quality assessment metrics[17]. Other useful validation metrics are RPF-DP scores, which compare models against the unassigned NOESY data and resonance assignments[18,19]. RPF-DP scores are correlated with structure accuracy for fully protonated proteins, with reliable models having DP scores greater than ~0.70–0.75 (refs. 18,19).

To verify this NMR DP threshold for deuterated proteins protonated only on amide and I(δ)LV methyl sites, we carried out a comprehensive study of the correlation between these scores and model accuracy. This analysis was done using CS-Rosetta[48] decoys generated with backbone chemical shift data obtained on three perdeuterated, I(δ)LV -methyl protonated test proteins (**Supplementary Fig. 12**). This study demonstrated a good correlation between DP scores and protein model accuracy; nearly all models with DP score >0.73 have backbone r.m.s. deviation to the corresponding reference structure < ~4 Å. Hence, we conclude that 'reliable models' will have DP scores > ~0.73, whether they are from fully protonated or deuterated protein samples.

The NMR DP scores[18] reported by ASDP provide a global measure of how well the structures fit with the NMR NOE data. Reliable models will generally have DP scores >0.73. NMR DP scores can also be computed independently of the ASDP program using the RPF-DP server available at http://nmr.cabm.rutgers.edu/rpf/. The RPF-DP program can also be downloaded to run on local machines. Reliable EC-NMR structures also have structure quality *Z* scores[17] > −2 for Procheck(backbone), Procheck(all dihedral), Verify3D, MolProbity and Prosa II knowledge-based structure quality assessment metrics (**Supplementary Fig. 9**). Structure quality *Z* scores can be computed using the on-line Protein Structure Validation Software Suite Server (PSVS) accessible at http://psvs-1_5-dev.nesg.org/. Detailed instructions on using the PSVS server are available at http://www.nmr2.buffalo.edu/nesg.wiki/PSVS.

Sample preparation, NMR spectroscopy data collection, and analysis of reference NMR spectroscopy protein structures. Isotope-enriched samples were prepared using standard methods[49], and NMR data collection and analysis was carried out by the Northeast Structural Genomics Consortium, as described[50,51], except for RASH_HUMAN. These data sets, and the authors contributing to each of the corresponding PDB IDs and DOIs, together with a summary of the distance restraint and RDC data used for generating each of these reference NMR structures, are outlined in **Supplementary Table 2**. In this study, data for RASH_HUMAN was obtained from PDB ID 2LCF[52], as experimental NOESY peaks lists were not available. Instead, H^N-H^N NOESY peaks were back-calculated from the distance restraint and resonance assignment lists using an interproton cutoff of 5 Å; no NOEs to methyl protons were assumed. The NMR data sets used in this study, together with the EC lists and resulting EC NMR structures, are all collected at http://ec-nmr.nesg.org/.

Data sets for maltose-binding protein (MALE_ECOLI) bound to β-cyclodextrin and protein P74712 (P74712_SYNY3) were recorded on ^2H,^{15}N,^{13}C-enriched samples with $^{13}CH_3$ labeling of Leu, Val and Ile(δ) atoms[23]. For the six other protein NMR data sets, NOESY data were collected on uniformly ^{15}N,^{13}C-enriched samples, essentially complete backbone and side-chain resonance assignments were determined using standard methods[50,51]. For EC-NMR studies, the resonance assignment lists for these six proteins were modified to exclude all entries except the backbone and side-chain H^N amide protons, as would be obtained on a ^2H,^{15}N,^{13}C-enriched sample. These 'sparse NMR data sets' were analyzed to provide interproton distance restraints by the EC-NMR protocol using ASDP. Statistics on the sparseness of the resulting NOESY-based distance restraints are summarized in **Supplementary Table 3**.

Rotamer comparisons between EC-NMR and reference X-ray crystal structures. The χ_1 rotamers for all residues in each reference X-ray crystal structure were assigned to the nearest g^+, t or g^- conformational state. Side chains with solvent accessible surface area (SASA) less than 40 $Å^2$ in the reference X-ray crystal structure (calculated using the program Molmol[53]) were considered as buried side chains. In considering NMR structure ensembles (for example, the EC-NMR structure or a NMR structure obtained from the PDB), side chains whose χ_1 dihedral angle values had standard deviation of <30 degrees were considered as 'converged side chains'. Rotamer states for residues with both buried and converged side chains were compared between the reference X-ray crystal (or the 'representative' NMR conformer) and each member of the ensemble of NMR structures. The percentages of χ_1 rotamer states for buried and converged side chains that are consistent between the representative (medoid) conformer[54,55] selected from the ensemble of NMR structures and the reference X-ray crystal are summarized in **Supplementary Table 5**.

Impact of using RDC data for two independent hydrodynamic alignment tensors. Significantly improved restraining power can be obtained by combining RDCs measured using more than one hydrodynamic alignment tensor[56–58]. For four of the NMR data sets used in this study, experimental RDC data are available for two independent hydrodynamic alignment tensors (**Table 1**). For two additional NMR protein data sets, RDC data are available for only one hydrodynamic alignment tensor, and for two proteins no RDC data are available. To assess whether EC-NMR structures can potentially be improved using RDC data obtained with multiple hydrodynamic alignments, as a proof of principle we simulated additional RDC data for the two proteins for which experimental RDC data were available for only one hydrodynamic alignment tensor (Q1LD49_RALME and MBP), and for two for which no experimental RDC data are available (A9CJD6_AGRTT5 and RASH_HUMAN p21 H-Ras), using the program ReDCat[43]. These results are shown in parentheses in **Table 1**. The impact of having two sets of RDC data, each measured with a distinct hydrodynamic alignment, is also illustrated in **Figure 2d** and **Supplementary Figures 3–6**.

Adding additional RDC data for two independent hydrodynamic alignments had little impact on the accuracy of the small proteins studied. It did, however, improve the accuracy of the larger proteins. For human p21 H-Ras, adding RDC data computed for two distinct hydrodynamic alignment tensors significantly improved the EC-NMR model accuracy; backbone r.m.s. deviation 1.6 Å (previously 2.6 Å), all-heavy-atom r.m.s. deviation 2.6 Å (previously 3.6 Å). Using RDCs for two hydrodynamic alignments also improves the buried χ_1 rotamer match statistics to 87% (previously 85%) and 80% (previously 65%) for p21 H-Ras and MBP, respectively (**Fig. 2d** and **Supplementary Table 5**).

Box plots. Box plots were used to present r.m.s. deviation comparisons. In these plots, box in the middle indicates quartiles and median scores; the 'whiskers' show the largest/smallest observation that falls within a distance of 1.5 times the nearest quartile. Any additional points are shown as outliers.

Calculation of precision, recall, performance, and r.m.s. deviations. The precision (P), recall (R), and performance (F) statistics were computed for sets of EC contacts or expanded lists of Residue Pair Contacts (RPCs) resulting from the EC-NMR protocol as:

$$P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \quad (1)$$

$$R = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \quad (2)$$

$$F = (2\times \mathrm{R}\times\mathrm{P})/(\mathrm{R}+\mathrm{P}) \quad (3)$$

doi:10.1038/nmeth.3455

In this analysis a TP contact is defined for residue pair (i, j) if any atom of residue i is ≤5 Å apart from any atom of residue j in the reference structure. An EC (or RPC) for which a contact is not indicated in the reference structure is a FP. A contact in the reference structure which is not included in the EC (or RPC) list is a FN.

When X-ray crystal structures were used as the reference structure, hydrogens were added using the Reduce program of the MolProbity software package[59]. When NMR ensembles were used as the reference structure, a TP was defined if this criterion was satisfied for at least 60% of the conformers in the NMR ensemble. The Precision statistic is the fraction of TPs in all the predicted contacts. Recall (R) is the fraction of TPs identified compared to all the contacts observed in the reference structure. These P, R and F statistics assume that the experimental X-ray crystal or NMR structure is the 'ground truth', and the EC or RPC contacts are the 'prediction'. They differ from those used in assessing NMR models against NMR NOESY peak list data (NMR RPF[18]), in which the model is taken as the 'prediction' and the NOESY data is the 'ground truth'.

Backbone (defined as N, Cα, C′, and O atoms) and all-heavy-atom (N, C, O, S) r.m.s. deviations were computed using the fit command, for specified residue ranges, as implemented in PyMOL software (The PyMOL Molecular Graphics System; Schrodinger, LLC).

20. Tugarinov, V., Kanelis, V. & Kay, L.E. *Nat. Protoc.* **1**, 749–754 (2006).
21. Hiller, S. *et al. Science* **321**, 1206–1210 (2008).
22. Raman, S. *et al. Science* **327**, 1014–1018 (2010).
23. Lange, O.F. *et al. Proc. Natl. Acad. Sci. USA* **109**, 10873–10878 (2012).
24. Tugarinov, V., Choy, W.Y., Orekhov, V.Y. & Kay, L.E. *Proc. Natl. Acad. Sci. USA* **102**, 622–627 (2005).
25. Grishaev, A., Tugarinov, V., Kay, L.E., Trewhella, J. & Bax, A. *J. Biomol. NMR* **40**, 95–106 (2008).
26. Huang, Y.J., Tejero, R., Powers, R. & Montelione, G.T. *Proteins* **62**, 587–603 (2006).
27. Herrmann, T., Güntert, P. & Wüthrich, K. *J. Mol. Biol.* **319**, 209–227 (2002).
28. Rohl, C.A., Strauss, C.E., Misura, K.M. & Baker, D. *Methods Enzymol.* **383**, 66–93 (2004).
29. Eddy, S.R. *PLoS Comput. Biol.* **7**, e1002195 (2011).
30. Wishart, D.S. & Sykes, B.D. *J. Biomol. NMR* **4**, 171–180 (1994).
31. Wüthrich, K. *NMR of Proteins and Nucleic Acids* (Wiley, 1986).
32. Maltsev, A.S., Ying, J. & Bax, A. *J. Biomol. NMR* **54**, 181–191 (2012).
33. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. *J. Biomol. NMR* **44**, 213–223 (2009).
34. Jones, D.T., Buchan, D.W., Cozzetto, D. & Pontil, M. *Bioinformatics* **28**, 184–190 (2012).
35. Kamisetty, H., Ovchinnikov, S. & Baker, D. *Proc. Natl. Acad. Sci. USA* **110**, 15674–15679 (2013).
36. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. *Phys. Rev. E* **87**, 012707 (2013).
37. de Juan, D., Pazos, F. & Valencia, A. *Nat. Rev. Genet.* **14**, 249–261 (2013).
38. Ulrich, E.L. *et al. Nucleic Acids Res.* **36**, D402–D408 (2008).
39. Bartels, C., Xia, T.H., Billeter, M., Güntert, P. & Wüthrich, K. *J. Biomol. NMR* **6**, 1–10 (1995).
40. Diercks, T., Coles, M. & Kessler, H. *J. Biomol. NMR* **15**, 177–180 (1999).
41. Shen, Y. & Bax, A. *J. Biomol. NMR* **56**, 227–241 (2013).
42. Zweckstetter, M. & Bax, A. *J. Am. Chem. Soc.* **122**, 3791–3792 (2000).
43. Valafar, H. & Prestegard, J.H. *J. Magn. Reson.* **167**, 228–241 (2004).
44. Lovell, S.C. *et al. Proteins* **50**, 437–450 (2003).
45. Laskowski, R.A., Moss, D.S. & Thornton, J.M. *J. Mol. Biol.* **231**, 1049–1067 (1993).
46. Sippl, M.J. *Proteins* **17**, 355–362 (1993).
47. Luthy, R., Bowie, J.U. & Eisenberg, D. *Nature* **356**, 83–85 (1992).
48. Shen, Y. *et al. Proc. Natl. Acad. Sci. USA* **105**, 4685–4690 (2008).
49. Acton, T.B. *et al. Methods Enzymol.* **493**, 21–60 (2011).
50. Baran, M.C., Huang, Y.J., Moseley, H.N. & Montelione, G.T. *Chem. Rev.* **104**, 3541–3556 (2004).
51. Huang, Y.J. *et al. Methods Enzymol.* **394**, 111–141 (2005).
52. Araki, M. *et al. J. Biol. Chem.* **286**, 39644–39653 (2011).
53. Koradi, R., Billeter, M. & Wüthrich, K. *J. Mol. Graphics* **14**, 51–55 (1996).
54. Montelione, G.T. *et al. Structure* **21**, 1563–1570 (2013).
55. Tejero, R., Snyder, D., Mao, B., Aramini, J.M. & Montelione, G.T. *J. Biomol. NMR* **56**, 337–351 (2013).
56. Prestegard, J.H., Bougault, C.M. & Kishore, A.I. *Chem. Rev.* **104**, 3519–3540 (2004).
57. Bax, A. *Protein Sci.* **12**, 1–16 (2003).
58. Al-Hashimi, H.M. *et al. J. Magn. Reson.* **143**, 402–406 (2000).
59. Word, J.M., Lovell, S.C., Richardson, J.S. & Richardson, D.C. *J. Mol. Biol.* **285**, 1735–1747 (1999).