Published in final edited form as:

Nat Methods. 2015 August ; 12(8): 751–754. doi:10.1038/nmeth.3455.

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], and **Gaetano T. Montelione**[1,2,6,8]

[1]Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, Piscataway, NJ, U.S.A.

[2]Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ, U.S.A.

[3]Department of Systems Biology, Harvard Medical School, Boston, MA, U.S.A.

[4]Department of Informatics, Technische Universität München, Garching, Germany

[5]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY, U.S.A.

[6]Department of Molecular Biology and Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, NJ, U.S.A.

Abstract

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

Solution-state NMR can generally provide accurate 3D structures of small (MW < ~ 15 kDa) proteins[1,2]. However, for larger proteins broad linewidths and resonance overlap make structure determination by NMR challenging. One important approach for addressing this problem is perdeuteration[3,4], in which most ^{1}H nuclei are replaced with ^{2}H, using biosynthetic methods. Perdeuteration generally increases the sensitivity and feasibility of NMR studies of larger proteins by decreasing the nuclear relaxation rates of the

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).

[7]These authors contributed equally to this work.

[8]These authors jointly supervised the work.

Author Contributions Y.T., Y.J.H., T.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.H., and D.S.M. performed calculations. Y.T., Y.J.H., T.H., C.S., D.S.M., and G.T.M. analyzed data. Y.T., Y.J.H., T.H., C.S., D.S.M. and G.T.M. wrote the manuscript.

The authors declare competing financial interests: G.T.M. is associated with Nexomics Biosciences, Inc., a scientific contract research organization.

remaining ^{1}H, ^{15}N, and ^{13}C nuclei[3]. Perdeuterated proteins, in which a subset of sites are selectively protonated, provide better 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 sidechain 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 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 (i.e., the protein subfamily or family) and do not necessarily reflect the intricate details of residue interactions within any particular protein of the multiple alignment. 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 novel hybrid approach for protein structure determination, which complements experimental sparse NMR data and mitigates specificity and accuracy limitations of structure modeling by evolutionary constraints. The new approach provides more complete and accurate residue pair contact information than either method alone. A general description of the EC-NMR method (Supplementary Figures 1 and 2), together with detailed protocols, is in On-Line Methods. The overall performance was tested using experimental sparse NMR data for 8 proteins ranging in size from 6 to 41 kDa (Table 1 and Supplementary Tables 1, 2, and 3). These EC-NMR structures utilized backbone H^{N}, Cα, C', and sidechain Cβ(and in some cases sidechain amide and methyl) resonance assignments, sparse NOESY-based restraint densities [0.09 to 2.0 long range ($|i - j| > 5$) NOE restraints per residue], backbone ^{15}N-^{1}H residual dipolar coupling (RDC) data (Supplementary Table 3), together with EC restraints.

The resulting EC-NMR structures were compared with known "reference structures", determined either by X-ray crystallography or by NMR using extensive backbone and sidechain ^{1}H, ^{13}C, and ^{15}N resonance assignments (Table 1, Fig. 1 and Supplementary Figure 3). Accuracy of these EC-NMR 3D structures was assessed using three metrics: (i) accuracy of atomic positions, (ii) accuracy of the residue pair contacts used to generate the structures, (iii) accuracy of sidechain 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 have accurate backbone and all-heavy-atom positions in 6 of 8 proteins studied; i.e. < 2 Å backbone atom positional root mean square deviations (RMSDs) and < 3 Å all-heavy atom RMSDs relative to the reference structure (Table 1; Fig. 1; Supplementary Figures 4&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 RMSDs < 2.8 Å and all-heavy-atom RMSDs < 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 / C-terminal domain backbone RMSD 1.6 Å / 1.9 Å, all-heavy-atom RMSD 2.4 Å / 2.7 Å; Table 1 and Supplementary Figure 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 its well-known interdomain flexibility[14].

For all 8 proteins studied, the final residue pair contact list generated by the *ASDP* program has higher coverage of the short-distance contacts in the reference structure, a lower false-positive rate, a higher precision, and a larger number of long-range residue-residue contacts than either the initial EC list or the sparse NMR data alone (Figs. 1 and 2a, Supplementary Figure 7 and Supplementary Table 4). The residue pair contact lists generated by the EC-NMR protocol provide more accurate and complete EC-NMR structures that those obtained using ECs or conventional sparse NMR methods alone.

A more detailed measure of accuracy resulted from comparison of the χ_1 side chain dihedral angles for buried residues with well-defined atomic coordinates across the conformers of the NMR ensemble. Averaged over all 8 EC-NMR structures, ~80% of these side chains have χ_1 rotamers that match the corresponding reference structures (Supplementary Table 5). For the 3 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 and 2c, Supplementary Figure 8, and Supplementary Table 5).

The value of the EC-NMR method in economizing the experimental NMR effort was assessed by comparing 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 was determined using essentially complete side-chain resonance assignments obtained on a fully protonated sample. For MBP, the core sidechain accuracy of the EC-NMR structure is significantly better than PDB ID 1EZP, determined using similar sparse NMR data together with 5 kinds of RDC data[4]. It is also similar to that of the solution NMR structure PDB ID 2D21, which was 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 On-Line Methods, further improved the accuracy of these EC-NMR structures (Table 1 and Fig. 2d).

In order to assess the robustness and sensitivity of the EC-NMR method to the amount of available sequence data, we computed evolutionary couplings (ECs) for randomly sampled subsets (50%, 25%, … 0.01%) of the full multiple sequence alignments (MSAs) 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}). ECs from these subsets were used for EC-NMR calculations (Supplementary Figure 9 and Supplementary Table 6). For this particular protein, the EC-NMR method breaks down at N_{eff} / L ~ 5, where L is the length of the protein; for larger sequence alignments (N_{eff} > ~ 1000) the backbone positional RMSDs between EC-NMR models and the X-ray crystal structure are consistently below ~ 3.5 Å. The more evolutionary sequence information is available, the better the resulting structures.

In developing the EC-NMR method, it is critical to have metrics which can be used to assess the reliability of the resulting structure models in the absence of a reference structure. For conventional NMR structures, methods are available which can 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 (e.g., backbone and side chain dihedral angle distributions, core atom packing, etc.) with those observed in high-resolution X-ray crystal structures. We assessed if these metrics can also discriminate "reliable" (backbone RMSD < 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 data. Each of the knowledge-based structure quality scores are 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 are able 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 Figure 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 Figure 10). Based on this analysis, we conclude that EC-NMR structures are "reliable" if they have NMR DP scores[18,19] greater than ~ 0.73, and knowledge-based Z scores computed with the PSVS server[17] more positive 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 (e.g. those arising from oligomer interfaces), 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 NMR structures obtained with complete side chain assignments and/or sophisticated side chain labeling methods.

For small proteins and domains up to ~ 140 residues (< ~15 kDa) with extensive sequence information, EC-NMR is a new, 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 makes extensive sidechain assignments difficult or impossible. **For larger proteins**, in the size range of 180–500 residues (20 to 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 utilize perdeuterated protein samples, and in protein structure determination by solid-state NMR methods. This advance expands the range of proteins for which accurate structures can be determined using either evolutionary coupling analysis or NMR spectroscopy data alone.

ONLINE METHODS

General description of the EC-NMR method

The computation of 3D structures from evolutionary couplings via distance constraints, while a breakthrough in the area of computational protein structure prediction, also has a number of limitations. Most importantly, the derived constraints, by definition, are an average over all 3D structures of the proteins in the multiple sequence alignment (i.e., the protein subfamily or family) and do not reflect the intricate details of residue interactions within any particular protein of the multiple alignment. In addition, even when there is extensive sequence information, residue-residue contacts indicated by high-ranked ECs may contain false positives as a result of insufficient data (undersampling) in the computational inference procedure of evolutionary couplings. The fraction of ECs, in some cases as much as ~30%, that are not consistent with a single folded native structure may also reflect real but confounding effects, such as conformational alternatives, homo-oligomerization, and/or indirect residue interactions via substrates.

The EC-NMR method involves three steps (Supplementary Fig. S1). Step A provides predicted residue pair contacts from sequence information. Evolutionary couplings are calculated for the protein from a multiple sequence alignment 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 multiple sequence alignment, 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 multiple sequence alignment, is used to remove the confounding effect of transitive correlations and thus reduce the number of false-positive 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. Sparse NMR data is collected using uniformly ^{13}C,^{15}N-enriched and/or ^{2}H,^{13}C,^{15}N-enriched protein samples prepared with ^{1}H-^{13}C labeling of sidechain Leu, Val, and Ile(δ) methyl groups[3,5,21]. Sequence-specific resonance assignments are determined for backbone $^{1}H^{N}$, ^{13}C, and ^{15}N resonances, as well as for sidechain $^{13}C\beta$ 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[22–24], and have been used in exceptional cases to determine chain folds for proteins as large as 82 kDa[25,26].

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 false positive (FP) ECs, and to generate initial 3D models of the protein. This automated combined analysis of NMR and EC data, ruling in ambiguous NOESY cross peak assignments and identifying ruling out false-positive EC contacts, is done using the NOESY assignment program *ASDP* [20]. 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 data, are then deconvoluted into atom-specific distance constraints, which are used to refine the protein structure using restrained energy minimization. In the current implementation, this refinement step uses the program *Rosetta* [2,28].

Alignments for generation of Evolutionary Couplings

Multiple sequence alignments (MSAs) were generated for each of the 8 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

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

Implementation of the EC-NMR method in the *ASDP* automated NOESY crosspeak assignment program

The EC-NMR method has been implemented within the automated NOESY cross peak assignment program *ASDP*[20]. 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 Fig. 2.

Step 1. Initial NOE-based distance constraints are generated from NOESY and chemical shift data using algorithms encoded in the *ASDP* program[20]. Secondary structures, including beta-strand alignments, are identified using previously described algorithms[20], 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[20], 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 elsewhere[20]. The cutoffs used in the EC-NMR analysis for identifying beta sheets are different from the cutoffs used for conventional NOESY analysis[20], 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 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 constraints 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 data also included 3D H^{N}-Me and Me-Me NOESY peak list data, which provide NOEs 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 isopropropyl groups were not included in the chemical shift lists.

The standard protocol of the *Talos+* computer 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° [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[20]. Structurally inconsistent RPCs (referred to here as SI-RPCs) are identified when ambiguous RPC distance constraints are violated by > 0.5 Å in more than 60% (e.g. 12 of 20) conformers. These SI-RPCs are excluded from the next cycle of *ASDP* calculatons.

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

Ambiguous RPC constraints 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 Residue Pair Contacts (RPCs), which are long-range residue pairs (i.e. $|i\text{-}j| \geq 5$) 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 of twenty conformers. These RPCs are added to the EC-NMR constraint 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 co-variation 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 Figure 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 re-generate 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 constraint 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 constraint 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 constraint 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 interresidue atom-atom constraints, 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 over-fitting.

Tutorial for EC-NMR Calculations

A web-base tutorial for running EC-NMR calculations is available on line at http://ec-nmr.nesg.org/tutorial.html. The on-line tutorial includes sample input and output data files. A step-by-step process is also provided in the following sub-sections.

EC pairs are generated from sequence data—EC pairs can be calculated using the *EVfold-plm* pipeline available at evfold.org (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 alternative 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 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, an 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*[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 calculation

Residual dipolar coupling data—Residual dipolar coupling data should be provided in the table format outlined in Sample Data. 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 the 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—Access to 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 Data 2.

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 list 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 Mao et al[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 constraints 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 constraint format. This can be done using a stand-alone 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 Hopf et al, 2014[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,

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

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 Figure 11). We refer to this as the Number of Reliable EC Pairs ($N_{reliable}$). Python code to identify high-confidence ECs is in Supplementary Data 1. 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 the Number of Reliable EC Pairs ($N_{reliable}$) 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 RMSD > 3.5 Å to the reference structure) for $N_{reliable} < \sim 25$ (Supplementary Figure 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 data; i.e. how well the model fits to derived restraint data. In our sensitivity analysis using reduced sequence data for EC-NMR studies of protein P74712, we observed that while some incorrect structures generated with highly-inaccurate EC data have significant numbers of restraint violations, some incorrect structures do not have significant violations of the interpreted restraints. This is not surprising, as the restraints themselves may be incorrectly interpreted from the NOE data when ECs with high false positive rates are used. Similar results have also been observed in analyses of the sensitivity of NMR restraint violations for validating NMR-derived structures[17]. Low restraint violations is a necessary, but not sufficient, condition for validating a distance-restraint-derived structure when the restraints themselves may be misinterpreted.

Other metrics used for NMR model validation include knowledge-based scores (e.g. 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[18,19].

In order 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 (results shown in Supplementary Fig. 12). This study demonstrates a good correlation between DP scores and protein model accuracy; nearly all models with DP score > 0.73 have backbone RMSD to 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* (<outputDir>/<proteinName>_DP.ovw) 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 on line 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 Figure 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 data collection, and analysis of reference NMR 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 elsewhere[50,51], except for RASH_HUMAN. These data sets, and the authors contributing to each of the corresponding Protein Data Bank entries 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 together on an on-line web site 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 ^{2}H,^{15}N,^{13}C-enriched samples with $^{13}CH_3$

labeling of Leu, Val, and Ile(δ) atoms[24]. For the six other protein NMR data sets, NOESY data were collected on uniformly $^{15}N,^{13}C$-enriched samples, essentially complete backbone and sidechain 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 sidechain H^N amide protons, as would be obtained on a $^{2}H,^{15}N,^{13}C$-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 χ1 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 (e.g., the EC-NMR structure or a NMR structure obtained from the PDB), side chains whose χ1 dihedral angle values had standard deviations 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 χ1 rotamer states for buried and converged sidechains 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 4 of the NMR data sets used in this study, experimental RDC data are available for two independent hydrodynamic alignment tensors (as summarized in Table 1). For 2 additional NMR protein data sets, RDC data is available for only 1 hydrodynamic alignment tensor, and for 2 proteins no RDC data is available. In order to assess if 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 2 proteins for which experimental RDC data were available for only 1 hydrodynamic alignment tensor (Q1LD49_RALME and MBP), and for 2 for which no experimental RDC data are available (A9CJD6_AGRTT5 and RASH_HUMAN p21 H-Ras), using the program ReDCat[59]. 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 the plots of Fig. 2d and in Supplementary Figures 3, 4, 5, and 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 RMSD 1.6 Å (previously 2.6 Å), all-heavy-atom RMSD 2.6 Å (previously 3.6

Å)]. Using RDCs for two hydrodynamic alignments also improves the buried χ1 rotamer match statistics to 87% (formally 85%) and 80% (formally 65%) for p21 H-Ras and MBP, respectively (Fig. 2d and Supplementary Table 5).

Box Plots

Box plots were used to present RMSD 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 (P), Recall (R), and Performance (F), and RMS 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:

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

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

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

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[60]. 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) Root Mean Square Deviations (RMSDs) were computed using the *fit* command, for specified residue ranges, as implemented in the *PyMol* software[61].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank all of the members of the Northeast Structural Genomics Consortium who generated and archived NMR data used in this work, particularly scientists in the laboratories of C. Arrowsmith, M. Kennedy, G.T. Montelione, T. Szyperski, and J. Prestegard. We also 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 National Institutes of Health: grant 1R01-GM106303 to C.S. & D.M. and Protein Structure Initiative grant U54-GM094597 to G.T.M.

References

1. Mao B, Guan R, Montelione GT. Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure. 2011; 19:757–766. [PubMed: 21645849]
2. Mao B, Tejero R, Baker D, Montelione GT. Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. Journal of the American Chemical Society. 2014; 136:1893–1906. [PubMed: 24392845]
3. Gardner KH, Rosen MK, Kay LE. Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. Biochemistry. 1997; 36:1389–1401. [PubMed: 9063887]
4. Mueller GA, et al. Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. Journal of molecular biology. 2000; 300:197–212. [PubMed: 10864509]
5. Rosen MK, et al. Selective methyl group protonation of perdeuterated proteins. Journal of molecular biology. 1996; 263:627–636. [PubMed: 8947563]
6. Marks DS, et al. Protein 3D structure computed from evolutionary sequence variation. PloS one. 2011; 6:e28766. [PubMed: 22163331]
7. Morcos F, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:E1293–E1301. [PubMed: 22106262]
8. Hopf TA, et al. Three-dimensional structures of membrane proteins from genomic sequencing. Cell. 2012; 149:1607–1621. [PubMed: 22579045]
9. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nature biotechnology. 2012; 30:1072–1080.
10. Hopf TA, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife. 2014; 3:e03430.
11. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife. 2014; 3:e02030. [PubMed: 24842992]
12. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:10340–10345. [PubMed: 22691493]
13. Nugent T, Jones DT. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:E1540–E1547. [PubMed: 22645369]
14. Evenas J, et al. Ligand-induced structural changes to maltodextrin-binding protein as studied by solution NMR spectroscopy. Journal of molecular biology. 2001; 309:961–974. [PubMed: 11399072]
15. Araki M, et al. Solution structure of the state 1 conformer of GTP-bound H-Ras protein and distinct dynamic properties between the state 1 and state 2 conformers. The Journal of biological chemistry. 2011; 286:39644–39653. [PubMed: 21930707]
16. Kainosho M, et al. Optimal isotope labelling for NMR protein structure determinations. Nature. 2006; 440:52–57. [PubMed: 16511487]
17. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. Proteins. 2007; 66:778–795. [PubMed: 17186527]

18. 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. Journal of the American Chemical Society. 2005; 127:1665–1674. [PubMed: 15701001]
19. Huang YJ, Rosato A, Singh G, Montelione GT. RPF: a quality assessment tool for protein NMR structures. Nucleic Acids Res. 2012; 40:W542–W546. [PubMed: 22570414]
20. 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]

Additional References for On Line Methods

21. Tugarinov V, Kanelis V, Kay LE. Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. Nature Protocols. 2006; 1:749–754. [PubMed: 17406304]
22. Hiller S, et al. Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. Science. 2008; 321:1206–1210. [PubMed: 18755977]
23. Raman S, et al. NMR structure determination for larger proteins using backbone-only data. Science. 2010; 327:1014–1018. [PubMed: 20133520]
24. Lange OF, et al. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:10873–10878. [PubMed: 22733734]
25. Tugarinov V, Choy WY, Orekhov VY, Kay LE. Solution NMR-derived global fold of a monomeric 82-kDa enzyme. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102:622–627. [PubMed: 15637152]
26. Grishaev A, Tugarinov V, Kay LE, Trewhella J, Bax A. Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. Journal of biomolecular NMR. 2008; 40:95–106. [PubMed: 18008171]
27. Herrmann T, Guntert P, Wuthrich K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. Journal of molecular biology. 2002; 319:209–227. [PubMed: 12051947]
28. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol. 2004; 383:66–93. [PubMed: 15063647]
29. Eddy SR. Accelerated Profile HMM Searches. PLoS computational biology. 2011; 7:e1002195. [PubMed: 22039361]
30. Wishart DS, Sykes BD. The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. Journal of biomolecular NMR. 1994; 4:171–180. [PubMed: 8019132]
31. Wuthrich, K. NMR of proteins and nucleic acids. Wiley: 1986.
32. Maltsev AS, Ying J, Bax A. Deuterium isotope shifts for backbone (1)H, (1)(5)N and (1)(3)C nuclei in intrinsically disordered protein alpha-synuclein. Journal of biomolecular NMR. 2012; 54:181–191. [PubMed: 22960996]
33. Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. Journal of biomolecular NMR. 2009; 44:213–223. [PubMed: 19548092]
34. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28:184–190. [PubMed: 22101153]
35. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:15674–15679. [PubMed: 24009338]
36. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Physical review. E, Statistical, nonlinear, and soft matter physics. 2013; 87:012707.

37. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet. 2013; 14:249–261. [PubMed: 23458856]
38. Ulrich EL, et al. BioMagResBank. Nucleic Acids Res. 2008; 36:D402–D408. [PubMed: 17984079]
39. Bartels C, Xia TH, Billeter M, Guntert P, Wuthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. Journal of biomolecular NMR. 1995; 6:1–10. [PubMed: 22911575]
40. Diercks T, Coles M, Kessler H. An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments. Journal of biomolecular NMR. 1999; 15:177–180. [PubMed: 20872110]
41. Shen Y, Bax A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. Journal of biomolecular NMR. 2013; 56:227–241. [PubMed: 23728592]
42. Zweckstetter M. Prediction of Sterically Induced Alignment in a Dilute Liquid Crystalline Phase: Aid to Protein Structure Determination by NMR. Journal of the American Chemical Society. 2000; 122:3791–3792.
43. Valafar H, Prestegard JH. REDCAT: a residual dipolar coupling analysis tool. Journal of magnetic resonance. 2004; 167:228–241. [PubMed: 15040978]
44. Lovell SC, et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins. 2003; 50:437–450. [PubMed: 12557186]
45. Laskowski RA, Moss DS, Thornton JM. Main-chain bond lengths and bond angles in protein structures. Journal of molecular biology. 1993; 231:1049–1067. [PubMed: 8515464]
46. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins. 1993; 17:355–362. [PubMed: 8108378]
47. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature. 1992; 356:83–85. [PubMed: 1538787]
48. Shen Y, et al. Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A. 2008; 105:4685–4690. [PubMed: 18326625]
49. Acton TB, et al. Preparation of protein samples for NMR structure, function, and small-molecule screening studies. Methods Enzymol. 2011; 493:21–60. [PubMed: 21371586]
50. Baran MC, Huang YJ, Moseley HN, Montelione GT. Automated analysis of protein NMR assignments and structures. Chemical reviews. 2004; 104:3541–3556. [PubMed: 15303826]
51. Huang YJ, et al. An integrated platform for automated analysis of protein NMR structures. Methods Enzymol. 2005; 394:111–141. [PubMed: 15808219]
52. Araki M, et al. Solution structure of the state 1 conformer of GTP-bound H-Ras protein and distinct dynamic properties between the state 1 and state 2 conformers. J Biol Chem. 2011; 286:39644–39653. [PubMed: 21930707]
53. Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. Journal of molecular graphics. 1996; 14:51–55. 29–32. [PubMed: 8744573]
54. Montelione GT, et al. Recommendations of the wwPDB NMR Validation Task Force. Structure. 2013; 21:1563–1570. [PubMed: 24010715]
55. Tejero R, Snyder D, Mao B, Aramini JM, Montelione GT. PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. Journal of Biomolecular Nmr. 2013
56. Prestegard JH, Bougault CM, Kishore AI. Residual dipolar couplings in structure determination of biomolecules. Chemical reviews. 2004; 104:3519–3540. [PubMed: 15303825]
57. Bax A. Weak alignment offers new NMR opportunities to study protein structure and dynamics. Protein Science. 2003; 12:1–16. [PubMed: 12493823]
58. Al-Hashimi HM, et al. Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. J Magn Reson. 2000; 143:402–406. [PubMed: 10729267]
59. Valafar H, Prestegard JH. REDCAT: a residual dipolar coupling analysis tool. J Magn Reson. 2004; 167:228–241. [PubMed: 15040978]

60. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. Journal of molecular biology. 1999; 285:1735–1747. [PubMed: 9917408]

61. The PyMOL Molecular Graphics System. Schrodinger, LLC.

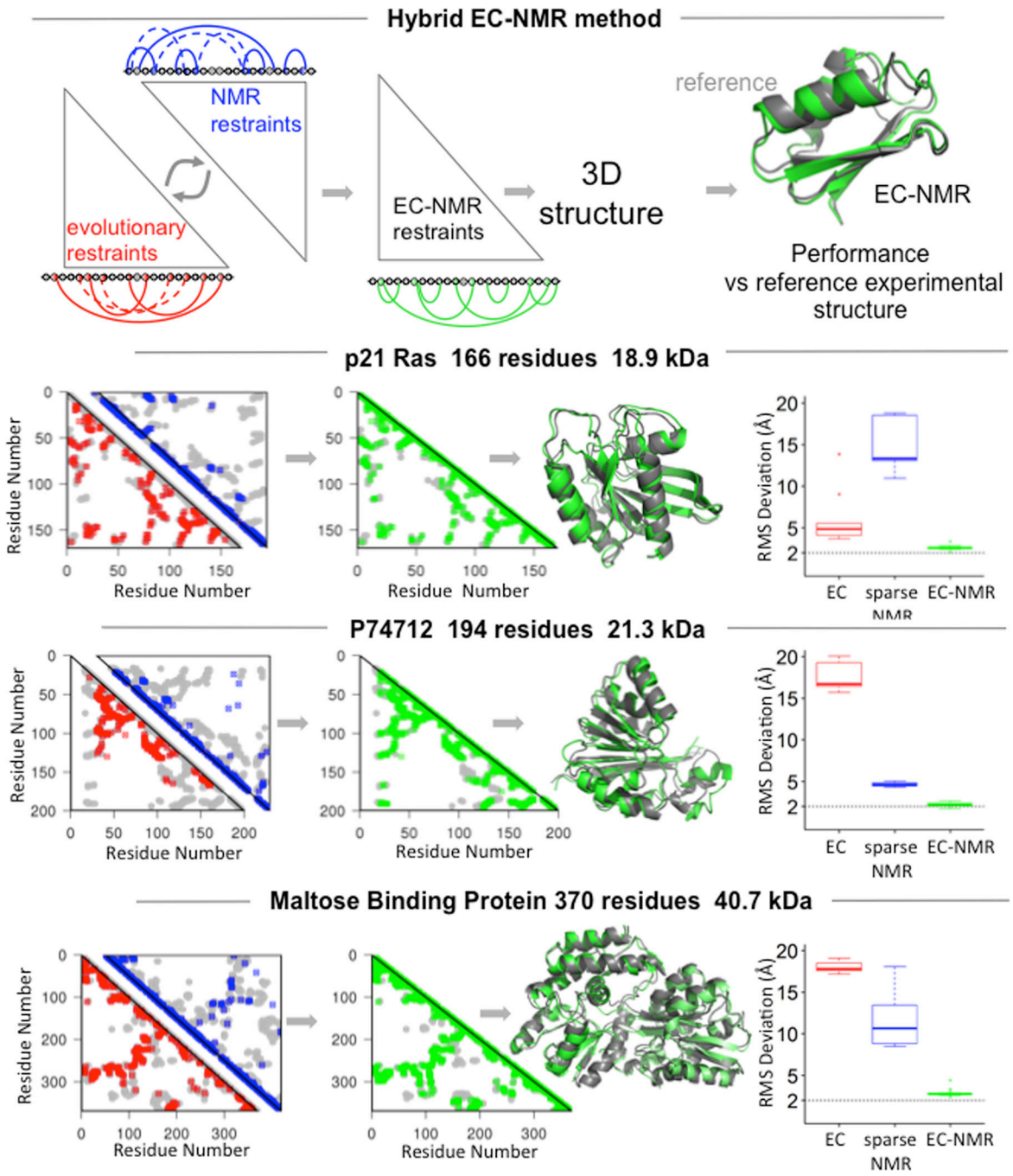

Fig. 1. The EC-NMR process

Top panel. EC information is interpreted together with ambiguous NOESY peak list data. Inconsistent ECs (dashed red 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 is assessed by comparing the resulting EC-NMR structure (green) with a reference X-ray crystal or NMR structure (grey). Each of the lower three horizontal panels illustrates the process of EC-NMR analysis using sparse NMR data for proteins with

MW of 19 – 41 kDa. Red contacts – initial EC residue-pair contacts. Blue contacts – contacts indicated by unambiguous NOESY peak assignments obtained by the *ASDP* program[29]. Green contacts – final Residue Pair Contacts (RPCs) resulting from simultaneous analysis of EC and NMR data. Grey contacts – contacts in the reference X-ray crystal structure. Green ribbon structures – final EC-NMR structures. Grey ribbons – reference X-ray crystal structures. Box plots show the RMS deviation to reference structures for backbone atoms of structures generated with EC data alone (red), 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/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 grey line) relative to the corresponding X-ray crystal structures.

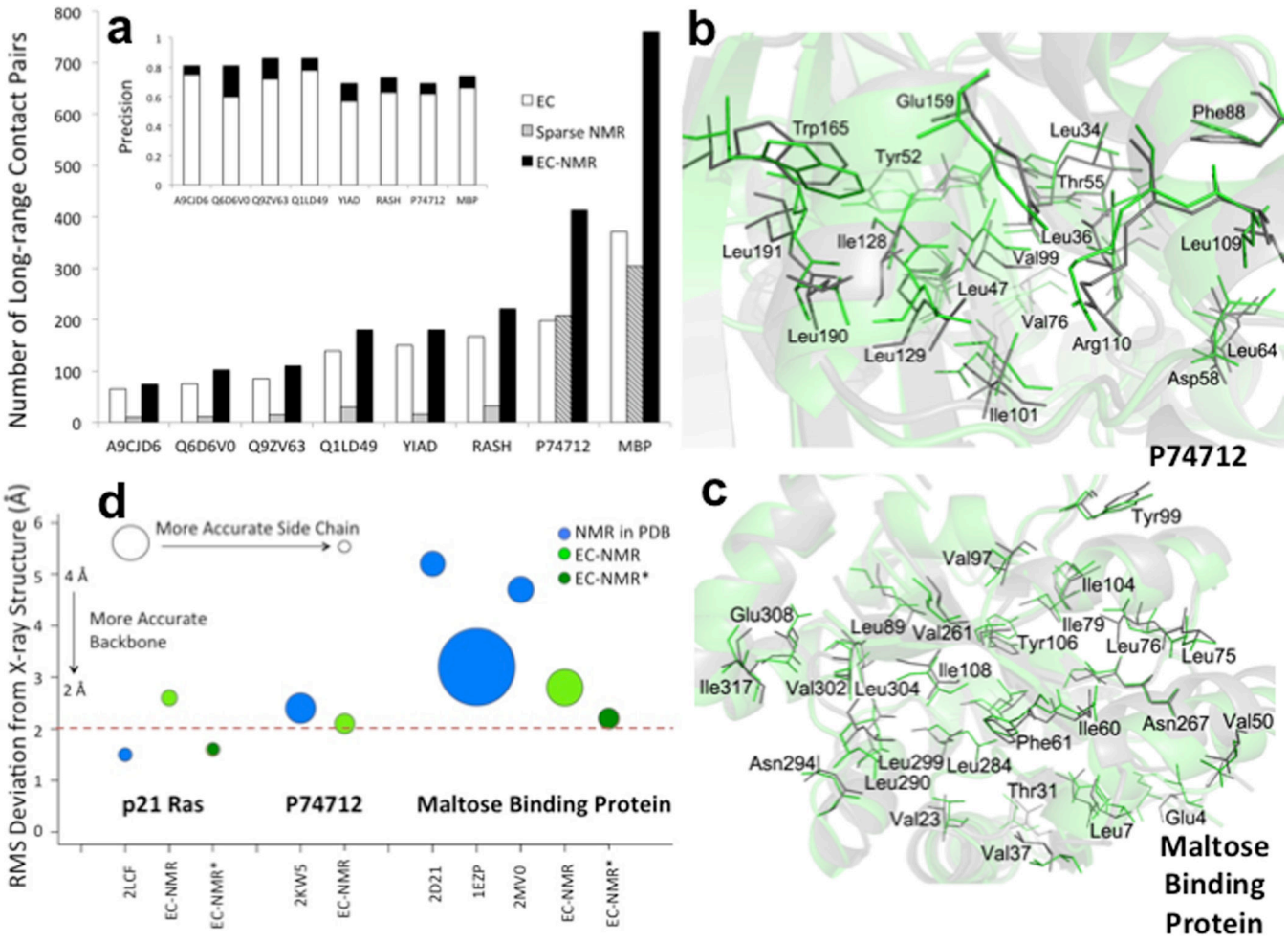

Fig. 2. Performance of the EC-NMR method

(**a**) Number of long-range residue pair contacts (i.e., between residue pairs (i, j) where |i – j| ≥ 5) for the initial EC list (white histograms), the initial unambiguous sparse NOESY data (grey), and the final EC-NMR residue contact list (black). For smaller (< 150 residues, grey-open) proteins, the NMR data include only H^N-H^N NOEs, while for larger proteins (> 150 residues, gray-hashed) the NMR data also include NOEs to Val, Leu, and Ile(δ) methyl protons. Inset – The Precision of contacts, relative to the corresponding reference structures, is higher for final Residue Pair Contact list (solid histograms) than for the initial EC list (open histograms), as false-positives are identified and removed by the EC-NMR algorithm. (**b,c**) Comparison of buried sidechain conformations in EC-NMR structures and the corresponding X-ray crystal structure. (**d**) Comparison of backbone RMSD and buried sidechain χ1 rotamers, relative to crystal structures. EC-NMR structures were determined using exclusively the experimental NMR 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 sidechains with χ1 rotamers different from that observed in the

crystal structure; smaller circles indicate a better match of sidechain conformations to the crystal structure.

Table 1

Experimental data and comparisons of EC-NMR structures with benchmark reference structures.

Protein Name and Uniprot ID	N[a] / MW[a] (kDa)	NOE Data[b]	^{15}N-^{1}H RDC Data[c]	No. Sequences in MSA[d]	RMSD (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)						
A. tumefaciens Protein of Unknown Function A9CJD6_AGRTT5	64 / 6.3	H^N- H^N only	None	10,962	1.5 [e] / 2.0 [e] (1.5 [e] / 1.8 [e])	2K2P NMR
E. carotovora Cold-shock-like protein Q6D6V0_ERWCT	66 / 7.3	H^N- H^N only	2 alignment tensor	4,410	2.2 [f] / 3.0 [f]	2K5N NMR
A. thaliana Ubiquitin-like domain Q9ZV63_ARATH	84 / 9.7	H^N- H^N only	2 alignment tensors	4,964	1.9 [g] / 2.5 [g]	2KAN NMR
R. metallidurans Rmet5065 Q1LD49_RALME	134 / 15.0	H^N- H^N only	1 alignment tensor	2,620	2.0 [h] / 3.0 [h] (2.0 [h] / 3.0 [h])	2LCG NMR
E. coli lipoprotein YiaD YIAD_ECOLI	141 / 15.0	H^N- H^N only	2 alignment tensors	10,296	1.7 [i] / 2.3 [i]	2K1S NMR
Larger (> ~15 kDa)						
H. sapiens H-ras oncogene protein p21 RASH_HUMAN	166 / 18.9	H^N- H^N only°	None	6,669	2.6 [j] / 3.6 [j] (1.6 [j] / 2.6 [j])	5P21 Xray
Slr1183 P74712_SYNY3	194/ 21.3	H^N- H^N, Me-Me, H^N-Me only	2 alignment tensors	45,708	2.1 [k] / 3.0 [k]	3MER Xray
E. coli Maltose Binding Protein MALE_ECOLI	370 / 40.7	H^N- H^N Me-Me, H^N Me only	1 alignment tensor	12,416		
NTD (1–112; 259–329)					1.6 [l] / 2.4 [l] (1.6 [l] / 2.5 [l])	1DMB Xray
CTD (113–258; 330–370)					1.9 [m] / 2.7 [m] (1.9[m] / 2.7 [m])	1DMB Xray
Full-length (1–370)					2.8 [n] / 3.4 [n] (2.2 [n] / 2.8 [n])	1DMB Xray

[a] Number of residues (N) and molecular weight (MW) of the protein construct studied by NMR, excluding affinity purification tags.

[b] H^N-H^N NOESY cross peak data include NOEs between backbone and sidechain amide H^N resonances. For P74712_SYNY3 and MALE_ECOLI, 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 constraint list (2LCF) and chemical shift data (BMRB ID 17610).

[c] All experimental ^{15}N-^{1}H RDC data were measured in the laboratory of James Prestegard.

[d] Number of non-redundant sequences in multiple sequence alignment used to generate ECs

[e] Residues ranges for superimpositions and rmsd calculations: 2–63

[f] Residues ranges for superimpositions and rmsd calculations: 1–64

[g] Residues ranges for superimpositions and rmsd calculations: 7–78

[h] Residues ranges for superimpositions and rmsd calculations: 1–29, 36–58, 62–135

[i] Residues ranges for superimpositions and rmsd calculations: 15–39,41–76,79–120,127–141

[j] Residues ranges for superimpositions and rmsd calculations: 1–29, 39–60, 64–166

[k] Residues ranges for superimpositions and rmsd calculations: 20–37, 41–134, 147–172, 185–196. Residues 1–15 and 175–183 are not observed in the crystal structure.

[l] Residues ranges for superimpositions and rmsd calculations: 2–12,14–112,259–329

[m] Residues ranges for superimpositions and rmsd calculations: 115–117,125–142,144–172, 175–218, 221–227, 247–258, 330–370. Interfacial residues 233–240 are exchange-broadened, precluding NMR assignments. The sugar binding site of MBP (1DMB) includes residues: K42, D65, E111, E153, Y155, E172, W230, W340, and R344

[n] Residues ranges for superimpositions and rmsd calculations: 2–12,14–112,259–329, 115–117,125–142,144–172, 175–218, 221–227, 247–258, 330–370. Interfacial residues 233–240 are exchange-broadened, precluding NMR assignments.