Published in final edited form as:

Adv Exp Med Biol. 2018 ; 1105: 153–169. doi:10.1007/978-981-13-2200-6_10.

CHAPTER X: A HYBRID APPROACH FOR PROTEIN STRUCTURE DETERMINATION COMBINING SPARSE NMR WITH EVOLUTIONARY COUPLING SEQUENCE DATA

Yuanpeng Janet Huang¹, Kelly Brock³, Chris Sander², Debora S. Marks³, and Gaetano T. Montelione¹

¹Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ, 08854, U.S.A.

²Department of Cell Biology, Harvard Medical School and cBio Center, Dana-Farber Cancer Institute

³Department of Systems Biology, Harvard Medical School, Boston, MA, 02115, U.S.A.

Abstract

While 3D structure determination of small (<15 kDa) proteins by solution NMR is largely automated and routine, structural analysis of larger proteins is more challenging. An emerging hybrid strategy for modeling protein structures combines sparse NMR data that can be obtained for larger proteins with sequence co-variation data, called evolutionary couplings (ECs), obtained from multiple sequence alignments of protein families. This hybrid “EC-NMR” method can be used to accurately model larger (15 – 60 kDa) proteins, and more rapidly determine structures of smaller (5 – 15 kDa) proteins using only backbone NMR data. The resulting structures have accuracies relative to reference structures comparable to those obtained with full backbone and sidechain NMR resonance assignments. The requirement that evolutionary couplings (ECs) are consistent with NMR data recorded on a specific member of a protein family, under specific conditions, potentially also allows identification of ECs that reflect alternative allosteric or excited states of the protein structure.

Keywords

Hybrid Methods; Protein NMR Spectroscopy; Protein Families; Multiple Sequence Alignment; Maximum Entropy; Evolutionary Couplings; Automated NMR Data Analysis; AutoStructure / ASDP

X.1. Introduction

Solution-state NMR can generally provide accurate three-dimensional (3D) structures of small (MW < ~ 15 kDa) proteins (Mao et al. 2011; Mao et al. 2014). However, for larger proteins the efficient transverse spin relaxation of the ¹H-¹H network results in broad NMR linewidths, preventing collection of sufficient data to allow structural analysis. Perdeuteration and selective reprotonation (i.e. replacement of most ¹H atoms with ²H) decreases transverse relaxation rates of the remaining ¹H, ¹⁵N, and ¹³C nuclei, increasing the sensitivity and feasibility of NMR for larger proteins (Gardner et al. 1997). However,

perdeuteration also reduces the number of ^1H 's providing ^1H - ^1H NOEs, and generally excludes most sidechain protons, providing much fewer structural restraints. This incompleteness of NOE data can be compensated to some degree using conformational restraints based on chemical shift and orientation restraints from residual dipolar coupling (RDC) data. Although protein structure models based on such "sparse NMR data" can be improved using advanced knowledge-based molecular modeling methods (Raman et al. 2010; Lange et al. 2012; Sgourakis et al. 2014), the resulting structures are generally less accurate and precise than those obtained for smaller, fully-protonated proteins with complete sidechain resonance assignments.

It has long been a goal of bioinformatics research to use sequence co-variation to provide information about residue pair contacts, which could enable protein structure prediction and modeling (Gobel et al. 1994; Neher 1994; Taylor and Hatrick 1994; Shindyalov et al. 1994; Thomas et al. 1996). Historically, a key challenge was created by transitive correlations, or relay effects; i.e., to distinguish A-B covariation due to A->B interactions from A-C covariation due to relayed A->B->C interactions. Recently, methods have been developed using maximum entropy global statistical models and maximum likelihood parameter inference that distinguish direct evolutionary couplings from transitive correlations, allowing reliable analysis of evolutionary residue-residue couplings from multiple alignments of structurally related protein sequences (Lapedes et al. 2002; Morcos et al. 2011; Marks et al. 2011; Sulkowska et al. 2012; Kamisetty et al. 2013). Such evolutionary couplings (ECs), derived from evolutionary-correlated mutations, can provide accurate information about residue pair contacts in the 3D structures of proteins and protein complexes (Morcos et al. 2011; Marks et al. 2011; Sulkowska et al. 2012; Hopf et al. 2012; Marks et al. 2012; Kamisetty et al. 2013; Hopf et al. 2014; Michel et al. 2014; Ovchinnikov et al. 2014; Ovchinnikov et al. 2016; Anishchenko et al. 2017; Ovchinnikov et al. 2017; Simkovic et al. 2017). Most often, the highest scoring evolutionary couplings are between residues that indeed contact one another in the 3D structure. These contacts can then be used, together with molecular dynamics, knowledge-based, and/or energy minimization methods to model the native structure of the protein, with often correct identification of the protein fold (Marks et al. 2011; Sulkowska et al. 2012; Hopf et al. 2012; Sheridan et al. 2015; Ovchinnikov et al. 2015; Ovchinnikov et al. 2016; Ovchinnikov et al. 2017). Importantly, high-confidence ECs may also reflect protein-protein interactions (Hopf et al. 2014; Cheng et al. 2014; Ovchinnikov et al. 2014; dos Santos et al. 2015; Toth-Petroczy et al. 2016), alternative conformational or allosteric states (Morcos et al. 2013; Toth-Petroczy et al. 2016), and/or more subtle features of the protein structure and dynamics.

While a breakthrough in the area of computational protein folding and protein structure prediction, the modeling of 3D structures from evolutionary couplings has a number of limitations. ECs provide information on residue-residue contacts present in many of the 3D structures of the proteins across the multiple sequence alignment (i.e., across the iso-structural protein subfamily or family), and may not accurately reflect the specific structural details of the particular protein under investigation. More specifically, there may be "structural drift" across the protein family, and sequence co-variation averaged over interaction constraints, including those in distantly related members of the family, may be inconsistent with the structure of the subject protein (Tang et al. 2015). In addition, even

when there is extensive sequence information, residue-residue contacts indicated by high-ranked ECs may not be consistent with the native structure under investigation, but rather reflect important but confounding effects, such as conformational alternatives, allosteric networks, excited-state conformations, homo-oligomerization, and/or indirect residue interactions via substrates or binding partners. They may also result from simple false positives in the parameter inference computation, especially when insufficiently diverse sequences are available. As a result, EC-derived models of proteins may differ in detail from the predominant native structure.

Residue contact information derived from sparse NMR data or from evolutionary couplings can provide highly complementary information. This creates the opportunity to combine the two for more reliable structure determination than can be achieved using either data type alone (Tang et al. 2015). Sparse NMR contact information is incomplete and often ambiguous in its assignment to specific ^1H - ^1H interactions. Nonetheless, all (or most) of the NOE, chemical shift, and RDC data should be consistent with the 3D structure model(s), averaged over an ensemble at finite temperature. EC-based contacts can complement this spectroscopic information to provide more complete contact information, and more accurate models, but potentially include interactions that are not consistent with the predominant structure of the subject protein under the conditions that the NMR data is acquired. The requirement that the overall structure be consistent with all of the experimental NMR data, however, provides “hard” constraints on the interpretation of ECs, allowing identification and removal of proposed residue pair contacts that are inconsistent with the dominant structure present under the solution conditions under investigation (Tang et al. 2015).

X.2. The EC-NMR Algorithm

The general EC-NMR method, as described by Tang et al. (2015) is outlined in Fig. 1. The overall process can be divided into three subprocesses. Step 1 provides a ranked list of direct evolutionary couplings (ECs) from multiple sequence alignments using either maximum entropy or pseudolikelihood models of the protein sequence, constrained by the statistics of the multiple sequence alignment, that have been developed to distinguish direct from transitive couplings (Morcos et al. 2011; Marks et al. 2011; Jones et al. 2012; Ekeberg et al. 2013; Kamisetty et al. 2013). In generating the multiple sequence alignment, it is important to carefully choose an appropriate range of evolutionary neighbors: not too many, so as to optimize specificity of structural constraints to the target of interest, and not too few, so as to retrieve as many sequences as possible at maximum sequence diversity and thus reduce sampling bias. In our published implementation of EC-NMR, the interaction parameters in the model, i.e., the evolutionary residue-residue couplings, were computed using pseudolikelihood maximization in the computer program *plmc*, part of the *Evcouplings* software suite (Ekeberg 2013; <https://github.com/debbiemarkslab/plmc>).

In Step 2, sparse NMR data is collected using uniformly ^{13}C , ^{15}N -enriched and/or ^2H , ^{13}C , ^{15}N -enriched protein samples prepared with ^1H - ^{13}C labeling of sidechain Leu, Val, and Ile($\delta 1$) methyl groups (Gardner et al. 1997; Rosen et al. 1996; Tugarinov et al. 2006), providing backbone ^1H , ^{13}C , and ^{15}N , as well as sidechain amide ^1H - ^{15}N and some methyl $^{13}\text{CH}_3$ resonance assignments. Backbone resonance assignments are determined, and

backbone dihedral angle restraints are defined from $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ chemical shift data using the program TALOS-N (Shen and Bax 2015). Unassigned NOESY peak lists are then generated from simultaneous 3D ^{15}N , ^{13}C -NOESY spectra, and, in some cases, ^{15}N - ^1H residual dipolar coupling (RDC) data are measured using one or more RDC alignment media. Sparse NMR data can generally be obtained for perdeuterated proteins with molecular weights as large as 40–70 kDa (Hiller et al. 2008; Raman et al. 2010; Lange et al. 2012), and have been used to determine chain folds for proteins as large as 82 kDa (Tugarinov et al. 2005; Grishaev et al. 2008).

Step 3 identifies and iteratively refines distance constraints using both sources of information simultaneously, 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 ECs that are inconsistent with the NMR data, and to generate initial 3D models of the protein. This automated combined analysis of NMR and EC data is implemented in the NOESY assignment program *ASDP* (Huang et al. 2006). Intermediate 3D structures are generated from these combined NMR and evolutionary distance constraints using the program *CYANA* (Herrmann et al. 2002). 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 published implementation (Tang et al., 2015), the refinement step used a specific restrained energy minimization and knowledge-based modeling protocol with the program *Rosetta*, described by Mao et al. (2014), but alternative energy refinement protocols could also be used.

X.3. EC-NMR Results

Tang et al (2015) tested the overall performance of the EC-NMR method using experimental chemical shift, NOESY peak list, and RDC data for 8 proteins ranging in size from 6 to 41 kDa. These data were obtained from the archives of the Northeast Structural Genomics Consortium (www.nesg.org) (Everett et al. 2016). The resulting EC-NMR structures were compared with “reference structures”, which have been determined either by X-ray crystallography or by NMR using essentially complete backbone and sidechain resonance assignments. These EC-NMR structures were observed to have accurate backbone and all-heavy-atom positions; i.e. $< 2 \text{ \AA}$ backbone atom positional root mean square deviations (RMSDs) and $< 3 \text{ \AA}$ all-heavy atom RMSDs relative to the reference structure, in 6/8 proteins. The remaining two proteins studied, human p21 H-Ras and maltose binding protein had no or limited RDC data, respectively, but were nevertheless reasonably accurate; both protein structures had backbone RMSDs $< 2.8 \text{ \AA}$ and all-heavy-atom RMSDs $< 3.6 \text{ \AA}$ relative to the corresponding X-ray crystal structures (Tang et al. 2015).

For this monograph, we re-determined five of the EC-NMR structures reported by Tang et al. (2015) using the same archived NMR data, but an updated database of protein sequences, downloaded in April 2017. These five proteins and the NMR data used for this study are summarized in Table 1. For the four smaller protein targets, with molecular weights of 6 to 15 kDa, the NMR data include only H^N - H^N NOE data, along with restraints on backbone dihedral angles computed from $\text{C}^\alpha/\text{C}^\beta$ chemical shifts using Talos-N (Shen and Bax 2015).

For two of these four proteins, ^{15}N - ^1H RDCs were measured using two different molecular alignment conditions, for a third ^{15}N - ^1H RDCs were measured using only one alignment condition, and for the fourth no RDC data are available. These four EC-NMR structures were compared with NMR structures determined with complete sidechain proton assignments and much more extensive NOESY data. The results of these EC-NMR calculations are shown in Fig. 2.

These four EC-NMR 3D structures were assessed based on (i) accuracy of atomic positions (Table 2) and (ii) accuracy of sidechain χ_1 rotamer states for well-defined (i.e. converged), buried (i.e., not on the protein surface) side chains (Table 3). In each case, the representative structure from the NMR ensemble (either the EC-NMR ensemble or the reference NMR structure ensemble) was selected as the medoid conformer of the ensembles, as described elsewhere (Montelione et al. 2013; Tejero et al. 2013). The backbone RMSD's between EC-NMR structures ranges from 1.5 to 1.8 Å, while the RMSD's for all C, N, O and S atoms (both backbone and sidechain) range from 2.4 to 2.9 Å (Table 2). The χ_1 values of well-defined buried sidechain (17 – 38 sidechains in the 4 structures), compared for all conformers in the EC-NMR ensemble with all conformers in the reference ensemble, also agree in 73 – 85% of pair-wise comparisons (Table 3). Similar results were observed for the corresponding earlier EC-NMR structures of these same proteins reported by Tang *et al.* (2015). In both studies, the EC-NMR structures are significantly more accurate than models generated using either the EC or sparse NMR data alone. Remarkably, these EC-NMR structures determined using only H^{N} - H^{N} NOE data together with ECs have accuracies that compare with high quality NMR structures determined with complete backbone and sidechain resonance assignments, suggesting that when good quality ECs are available for small (< 15 kDa) proteins, it may only be necessary to complete the majority of backbone resonance assignments in order to determine a high-quality solution NMR structure.

As a fifth illustrative example, we also reanalyzed the EC-NMR structure of the 41 kDa *E. coli* maltose binding protein (MBP) bound to beta-cyclodextrin. The experimental NMR data for MBP include H^{N} - H^{N} NOE data, as well as Ile(81), Leu, and Val methyl proton assignments, providing also Me-Me and HN-Me NOEs, along with restraints on backbone dihedral angles computed from $\text{C}^{\alpha}/\text{C}^{\beta}$ chemical shifts using Talos-N (Shen and Bax 2015). These results (Figure 3) demonstrate high-quality EC-NMR structures are produced, with backbone RMSD's to the corresponding X-ray crystal structure of 2.5 Å for backbone atoms, and 3.2 Å for all C, N, O and S atoms (both backbone and sidechain). MBP is a two-domain protein, and the relative orientation of domains depends on which sugars are bound; the “open form” being preferred when bound to beta-cyclodextrin (Evenas et al. 2001). Considered separately, the two individual domains of MBP in the EC-NMR structure of the two-domain protein are even more accurate when compared to the reference X-ray crystal structure (N-terminal domain / C-terminal domain backbone RMSD 1.8 Å / 1.7 Å, all-heavy-atom RMSD 2.7 Å / 2.6 Å; Table 2) than is apparent from rigid body superimposition for the entire protein.

We also compared the accuracy of the EC-NMR structure of MBP relative to previously published NMR structures determined with more extensive sidechain assignments (Table 3). The core sidechains of the EC-NMR structure are significantly more accurate than PDB ID

1EZIP, determined using similar sparse NMR data together with 5 kinds of RDC data (Mueller et al. 2000). The core sidechain accuracy of the EC-NMR structure is similar to that of the solution NMR structure PDB ID 2D21, which was determined using extensive side chain resonance assignments provided by the sophisticated and expensive stereo-arrayed isotope labeling (SAIL) method (Kainosho et al. 2006). Based on RMSD relative to the X-ray crystal structure of beta-cyclodextrin-bound MBP, the overall structure of the EC-NMR models are more accurate than any previously published NMR structures. Similar results for MBP were also reported by Tang *et al.* (2015). Hence, we conclude that the EC-NMR method of Tang *et al.* (2015) can deliver structures with accurate backbone and core side chain atomic positions for larger (~ 40 kDa, or larger) proteins, with accuracy comparable or better than models obtained with sophisticated side chain labeling methods.

X.4. Sensitivity to Numbers of Sequence Homologs in Multiple Sequence Alignment

A prerequisite for the EC-NMR approach is extensive, diverse sequence data, required to obtain accurate co-evolutionary couplings between the residues (Marks et al. 2011; Hopf et al. 2012; Kamisetty et al. 2013). Recent experience suggests that more than 2^*L non-redundant sequences (N_{eff}) are generally required for confident predictions of overall protein fold from EC's alone, where L is the length of the target sequence (Marks et al. 2012; Michel et al. 2014; Ovchinnikov et al. 2014; Hopf et al. 2014; Kamisetty et al. 2013; Ovchinnikov et al. 2017). For a target protein that is 200 residues long, this typically requires on the order of 5000 sequences, before removal of redundancy, in an initial multiple sequence alignment of a family of structurally homologous proteins as inferred using standard sequence similarity methods with, if in doubt, a fairly conservative cutoff in sequence similarity, equivalent to typically not less than about 20–30% identical residues fairly evenly distributed over the entire length of the protein (Sander and Schneider 1991).

For EC-NMR, our goal is to obtain models with accuracies comparable to high-quality NMR structures; i.e. backbone positional root mean square deviations (RMSD's) relative to reference structures $< 2.5 \text{ \AA}$ and accurate core sidechain packing. Tang *et al.* (2015) analyzed a series of multiple sequence alignments, testing the number of sequences from $N_{\text{eff}}/L \sim 150$ down to $N_{\text{eff}}/L < 0.1$. In that analysis, using this implementation of the EC-NMR method and good quality NMR data for a perdeuterated, Ile(δ 1), Leu and Val $^{13}\text{CH}_3$ methyl labeled protein, the cutoff point for accurate modeling ($< 2.5 \text{ \AA}$ backbone RMSD) was estimated to be $N_{\text{eff}}/L \sim 5$, with little improvement in structural accuracy for higher values of N_{eff}/L (see Fig. 4 of Tang et al., 2015).

For the five EC-NMR structures described above, the number of non-redundant sequences N_{eff} ranged from ~7,100 to ~44,000 sequences (Table 2), with N_{eff}/L ranging from 113 to 241 sequences / residue. In order to assess the impact of the growth of the sequence databases over the last few years, we also compared these five EC-NMR structures, determined with protein sequence data available in April 2017, with the corresponding structures described by Tang *et al.* (2015), using protein sequence data downloaded in August 2013 (Table 2). Between these dates, the number of non-redundant sequences

available for each of these five proteins increased significantly; by about 10% (for *A. thaliana* Ubiquitin-like domain Q9ZV63_ARATH) to 12-fold (for *R. metallidurans* Rmet5065 Q1LD49_RALME). This observation is consistent with our estimate that the size of the relevant sequence databases is doubling every two to three years (Tang *et al.* 2015), and that many proteins which cannot yet be reliably studied using the EC-NMR method will become amenable as the sequence data base grows. However, as these targets already had high N_{eff}/L using the 2013 sequence databases (ranging from 20 to 174 non-redundant sequences per residue), despite this increase in sequence data, with the available protein NMR data there was little or no improvements in structural accuracy (Table 2). This is consistent with the conclusions of Tang *et al.* (2015), that good EC-NMR models can be produced with N_{eff}/L as low as 5 sequences / residue, with little improvement for higher values of N_{eff}/L .

X.5. Conclusions and Future Prospects

Evolutionary information and sparse NMR data, used together with knowledge-based modeling, are highly complementary for protein structure determination. The EC-NMR approach improves the accuracy of models generated by EC data alone, by requiring that EC-based contacts are consistent with experimental NMR data collected for one member of the protein family under specific conditions. This requirement eliminates important, but confounding, EC-derived contact restraints that may arise from structural drift across the protein family, and allosteric networks and/or excited states which may also be detected as evolutionary co-variation. More specifically, the experimentally reliable, but ambiguous, contact information of sparse NOESY peak list data, together with orientation restraints from RDC data and backbone dihedral restraints from chemical shift data, can rule out ECs that are not relevant to the structure of the specific target protein. Simultaneously, ECs complement the sparse NOESY and RDC data that can be obtained on largely perdeuterated protein samples, a requirement for studies of larger proteins and membrane proteins reconstituted in micelles or nanodisks. In this way, complementarity EC and NMR data provide much more complete and accurate residue contact information than can be obtained from either method alone.

The EC-NMR method outlined in this monograph is **largely automated**, and provides high-quality 3D structures with accurate backbone and core sidechain conformations (Tang *et al.* 2015). **For small proteins** and domains up to 150 residues ($< \sim 15$ kDa) with extensive sequence information, EC-NMR is a new, powerful, and efficient approach for protein structure determination using only backbone NMR data. **For larger proteins**, up to 400–500 residues (40–60 kDa, or larger), for which extensive side chain resonance assignment is challenging if not prohibitive, ECs can be combined with sparse NMR data obtained on perdeuterated protein samples to provide structures that are more accurate and complete than those obtained using such NMR data alone. In the method outlined here, ECs are combined with NMR data to determine both small and larger soluble protein structures, but the same approach should be applicable to membrane proteins (Hopf *et al.* 2012; Ovchinnikov *et al.* 2017) and for RNA structure determination (Weinreb *et al.* 2016). This advance significantly expands the range of biomolecules for which accurate structures can be determined using either evolutionary coupling analysis or NMR spectroscopy data alone.

The EC-NMR method requires large multiple sequence alignments, which are only currently available for a fraction of known proteins. However, as the sequence databases continue to grow, more proteins will be amenable to this approach. Fortunately, combining ECs together with sparse NMR data reduces the requirements for the amount and diversity of sequence information.

In this work, we used a simple restrained energy minimization protocol of *Rosetta* in the final refinement step (Mao et al. 2014). This protocol improves both backbone and sidechain structure accuracy. It is advantageous because it is relatively fast, and can be implemented with limited computer resources. However, the resolution adapted recombination protocol (Resrac) developed by Lange and Baker (Raman et al. 2010; Lange and Baker 2012) has significant advantages for generating accurate structures of proteins from sparse NMR data (Raman et al. 2010; Lange et al. 2012). The Rasrec protocol has also been used successfully for modeling protein structures for EC data (Braun et al. 2015). While it is much more computationally demanding, currently limiting its broad application, the Rasrec protocol has the potential to provide more accurate EC-NMR structures with less complete and/or more noisy EC and sparse NMR data.

The EC-NMR method also allows identification of ECs which are not consistent with the NMR data collected for the target protein under specific conditions. While these are “false positives” relative to the modeling of this particular state of the protein, ECs with strong signals and high reliability that are not consistent with this particular state of the protein structure can provide information on alternative conformations accessible to the protein, excited states, and potentially provide information on allosteric networks. Further investigations of the combined use of ECs and NMR data to characterize the multiple conformational states of proteins and their energy landscapes is an exciting emerging area which can be explored using these powerful hybrid methods.

Acknowledgements

This work was supported by National Institutes of Health grants 1R01-GM120574 (to G.T.M.) and 1R01-GM106303 (C.S. & D.M.). 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.

Bibliography

- Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D (2017) Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci U S A* 114 (34):9122–9127. doi:10.1073/pnas.1702664114 [PubMed: 28784799]
- Braun T, Koehler Leman J, Lange OF (2015) Combining Evolutionary Information and an Iterative Sampling Strategy for Accurate Protein Structure Prediction. *PLoS Comput Biol* 11 (12):e1004661. doi:10.1371/journal.pcbi.1004661
- Cheng RR, Morcos F, Levine H, Onuchic JN (2014) Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc Natl Acad Sci U S A* 111 (5):E563–571. doi:10.1073/pnas.1323734111 [PubMed: 24449878]
- dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci Rep* 5:13652. doi:10.1038/srep13652 [PubMed: 26338201]

- Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys* 87 (1): 012707
- Evenas J, Tugarinov V, Skrynnikov NR, Goto NK, Muhandiram R, Kay LE (2001) Ligand-induced structural changes to maltodextrin-binding protein as studied by solution NMR spectroscopy. *Journal of molecular biology* 309 (4):961–974. doi:10.1006/jmbi.2001.4695 [PubMed: 11399072]
- Everett JK, Tejero R, Murthy SB, Acton TB, Aramini JM, Baran MC, Benach J, Cort JR, Eletsky A, Forouhar F, Guan R, Kuzin AP, Lee HW, Liu G, Mani R, Mao B, Mills JL, Montelione AF, Pederson K, Powers R, Ramelot T, Rossi P, Seetharaman J, Snyder D, Swapna GV, Vorobiev SM, Wu Y, Xiao R, Yang Y, Arrowsmith CH, Hunt JF, Kennedy MA, Prestegard JH, Szyperski T, Tong L, Montelione GT (2016) A community resource of experimental data for NMR / X-ray crystal structure pairs. *Protein Sci* 25 (1):30–45. doi:10.1002/pro.2774 [PubMed: 26293815]
- Gardner KH, Rosen MK, Kay LE (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry* 36 (6):1389–1401 [PubMed: 9063887]
- Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18 (4):309–317. doi:10.1002/prot.340180402 [PubMed: 8208723]
- Grishaev A, Tugarinov V, Kay LE, Trewella J, Bax A (2008) Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *J Biomol NMR* 40 (2):95–106. doi:10.1007/s10858-007-9211-5 [PubMed: 18008171]
- Herrmann T, Guntert P, Wuthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of molecular biology* 319 (1):209–227 [PubMed: 12051947]
- Hiller S, Garces RG, Malia TJ, Orekhov VY, Colombini M, Wagner G (2008) Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. *Science* 321 (5893):1206–1210. doi:10.1126/science.1161302 [PubMed: 18755977]
- Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149 (7):1607–1621. doi:10.1016/j.cell.2012.04.012 [PubMed: 22579045]
- Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Sander C, Bonvin AM, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* In press
- Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* 62 (3):587–603. doi: 10.1002/prot.20820 [PubMed: 16374783]
- Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28 (2):184–190. doi:10.1093/bioinformatics/btr638 [PubMed: 22101153]
- Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei Ono A, Guntert P (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature* 440 (7080):52–57. doi:10.1038/nature04525 [PubMed: 16511487]
- Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* 110 (39):15674–15679. doi:10.1073/pnas.1314045110 [PubMed: 24009338]
- Lange OF, Baker D (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* 80 (3):884–895 [PubMed: 22423358]
- Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci U S A* 109 (27):10873–10878. doi:10.1073/pnas.1203013109 [PubMed: 22733734]
- Lapedes A, Giraud B, Jarzynski C (2002) Using sequence alignments to predict protein structure and stability with high accuracy. National Laboratory Report LA-UR-02-4481. URL [http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-02-4481and arXiv:1207.2484\[q-bio.QM\]\(2012 copy\)](http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-02-4481and arXiv:1207.2484[q-bio.QM](2012 copy)).

- Mao B, Guan R, Montelione GT (2011) Improved technologies now routinely provide protein NMR structures useful for molecular replacement. *Structure* 19 (6):757–766. doi:10.1016/j.str.2011.04.005 [PubMed: 21645849]
- Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *J Am Chem Soc* 136 (5):1893–1906. doi:10.1021/ja409845w [PubMed: 24392845]
- Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6 (12):e28766. doi:10.1371/journal.pone.0028766
- Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30 (11):1072–1080. doi:10.1038/nbt.2419 [PubMed: 23138306]
- Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A (2014) PconsFold: improved contact predictions improve protein models. *Bioinformatics* 30 (17):i482–i488. doi:10.1093/bioinformatics/btu458 [PubMed: 25161237]
- Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL (2013) Recommendations of the wwPDB NMR Validation Task Force. *Structure* 21 (9):1563–1570. doi:10.1016/j.str.2013.07.021 [PubMed: 24010715]
- Morcos F, Jana B, Hwa T, Onuchic JN (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci U S A* 110 (51):20533–20538. doi:10.1073/pnas.1315625110 [PubMed: 24297889]
- Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 108 (49):E1293–1301. doi:10.1073/pnas.1111471108 [PubMed: 22106262]
- Mueller GA, Choy WY, Yang D, Forman-Kay JD, Venters RA, Kay LE (2000) 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* 300 (1):197–212. doi:10.1006/jmbi.2000.3842 [PubMed: 10864509]
- Neher E (1994) How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A* 91 (1):98–102 [PubMed: 8278414]
- Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* 3:e02030. doi:10.7554/eLife.02030
- Ovchinnikov S, Kim DE, Wang RY, Liu Y, DiMaio F, Baker D (2016) Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins* 84 Suppl 1:67–75. doi:10.1002/prot.24974 [PubMed: 26677056]
- Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* 4:e09248. doi:10.7554/eLife.09248
- Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D (2017) Protein structure determination using metagenome sequence data. *Science* 355 (6322):294–298. doi:10.1126/science.aah4043 [PubMed: 28104891]
- Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsy A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) NMR structure determination for larger proteins using backbone-only data. *Science* 327 (5968):1014–1018. doi:10.1126/science.1183649 [PubMed: 20133520]
- Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, Kay LE (1996) Selective methyl group protonation of perdeuterated proteins. *Journal of molecular biology* 263 (5):627–636 [PubMed: 8947563]
- Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9 (1):56–68. doi:10.1002/prot.340090107 [PubMed: 2017436]

- Sgourakis NG, Natarajan K, Ying J, Vogeli B, Boyd LF, Margulies DH, Bax A (2014) The structure of mouse cytomegalovirus m04 protein obtained from sparse NMR data reveals a conserved fold of the m02-m06 viral immune modulator family. *Structure* 22 (9):1263–1273. doi:10.1016/j.str.2014.05.018 [PubMed: 25126960]
- Shen Y, Bax A (2015) Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods in molecular biology* 1260:17–32. doi: 10.1007/978-1-4939-2239-0_2 [PubMed: 25502373]
- Sheridan R, Fieldhouse RJ, Hayat S, Sun Y, Antipin Y, Yang L, Hopf T, Marks DS, Sander C (2015) EVfold.org: Evolutionary couplings and protein 3D structure prediction. *bioRxiv* bioRxiv 021022. doi:10.1101/021022
- Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* 7 (3):349–358 [PubMed: 8177884]
- Simkovic F, Ovchinnikov S, Baker D, Rigden DJ (2017) Applications of contact predictions to structural biology. *IUCrJ* 4 (Pt 3):291–300. doi:10.1107/S2052252517005115
- Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci U S A* 109 (26):10340–10345. doi:10.1073/pnas.1207864109 [PubMed: 22691493]
- Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods* 12 (8): 751–754. doi:10.1038/nmeth.3455 [PubMed: 26121406]
- Taylor WR, Hatrick K (1994) Compensating changes in protein multiple sequence alignments. *Protein Eng* 7 (3):341–348 [PubMed: 8177883]
- Tejero R, Snyder D, Mao B, Aramini JM, Montelione GT (2013) PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. *J Biomol NMR* 56 (4):337–351. doi:10.1007/s10858-013-9753-7 [PubMed: 23897031]
- Thomas DJ, Casari G, Sander C (1996) The prediction of protein contacts from multiple sequence alignments. *Protein Eng* 9 (11):941–948 [PubMed: 8961347]
- Toth-Petroczy A, Palmedo P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS (2016) Structured states of disordered proteins from genomic sequences. *Cell* 167 (1):158–170 e112. doi:10.1016/j.cell.2016.09.010 [PubMed: 27662088]
- Tugarinov V, Choy WY, Orekhov VY, Kay LE (2005) Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc Natl Acad Sci U S A* 102 (3):622–627. doi:10.1073/pnas.0407792102 [PubMed: 15637152]
- Tugarinov V, Kanelis V, Kay LE (2006) Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nature Protocols* 1 (2):749–754. doi:nprot.2006.101 [pii] 10.1038/nprot.2006.101 [PubMed: 17406304]
- Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell* 165 (4):963–975. doi:10.1016/j.cell.2016.03.030 [PubMed: 27087444]

Fig. 1. 3D structure determination by the hybrid EC-NMR method.

The hybrid EC-NMR strategy combines Evolutionary Coupling (EC) information from protein sequences with sparse experimental nuclear magnetic resonance (NMR) data.

Fig. 2. EC-NMR structures determined using only H^N - H^N NOESY data superimposed on reference conventional NMR structures.

The representative structure from the ensemble of conformers generated by the EC-NMR method (green) is superimposed on a representative structure from reference NMR structure ensemble. For each protein, the left image is a superimposition of backbone atoms, and the right image a superimposition of backbone and well-defined core sidechain atoms.

Fig. 3. EC-NMR structure of *E. coli* maltose binding protein superimposed on reference X-ray crystal structure.

The top horizontal panels illustrate EC-NMR analysis process using sparse NMR data. Red contacts – initial EC residue-pair contacts. Blue contacts – contacts indicated by unambiguous NOESY peak assignments obtained by the *ASDP* program (Huang et al. 2006). Green contacts – final residue pair contacts resulting from simultaneous analysis of EC and NMR data. Grey contacts – contacts in the reference X-ray crystal structure. Box plots – RMSD 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).

Superimposed backbone and core sidechain structures are for full length MBP, and for the individual N-terminal domain (NTD) and C-terminal domain (CTD) in the full-length EC-NMR structure. Green ribbon structures – final EC-NMR structure of MBP. Grey ribbons – reference X-ray crystal structure.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.

Experimental data and 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	PDB ID of Reference Structure and Method of Structure Determination
<u>Proteins < ~15 kDa</u>					
<i>A. tumefaciens</i> Protein of Unknown Function A9CJD6_AGRIT5	64 / 6.3	H ^N -H ^N only	None	28,265	2K2P NMR
<i>E. carotovora</i> Cold-shock-like protein Q6D6V0_ERWCT	66 / 7.3	H ^N -H ^N only	2 alignment tensors	7,108	2K5N NMR
<i>A. thaliana</i> Ubiquitin-like domain Q9ZV63_ARATH	84 / 9.7	H ^N -H ^N only	2 alignment tensors	5,396	2KAN NMR
<i>R. metallidurans</i> Rmet5065 Q1LD49_RALME	134 / 15.0	H ^N -H ^N only	1 alignment tensor	31,674	2LCG NMR
<u>Protein > ~15 kDa</u>					
<i>E. coli</i> Maltose Binding Protein MALE_ECOLI NTD (1–112; 259–329) CTD (113–258; 330–370) Full-length (1–370)	370 / 40.7	H ^N - H ^N , Me-Me, H ^N -Me	1 alignment tensor	43,759	1DMB Xray 1DMB Xray 1DMB Xray

^aNumber of residues (N) and molecular weight (MW) of the protein construct studied by NMR, excluding affinity purification tags.

^bH^N-H^N NOESY cross peak data include NOEs between backbone and sidechain amide H^N resonances. For MALE_ECOLI, additional H^N-Me NOESY cross peak data obtained for uniformly ¹⁵N,¹³C,²H-enriched samples with ¹³CH₃ labeling of Ile(61), Leu, and Val methyls were also included.

^cAll experimental ¹⁵N-¹H RDC data were measured in the laboratory of James Prestegard.

^dNumber of non-redundant sequences in multiple sequence alignment used to generate ECs (N_{eff})

^eResidues ranges for superimpositions and RMSD calculations: 2–63

^fResidues ranges for superimpositions and RMSD calculations: 1–64

^gResidues ranges for superimpositions and RMSD calculations: 7–78

^hResidues ranges for superimpositions and RMSD calculations: 1–29, 36–58, 62–135

ⁱResidues ranges for superimpositions and RMSD calculations: 2–12,14–112,259–329

^jResidues 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

^kResidues 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.

Table 2.

Accuracy of EC-NMR Structures

Protein Name and Uniprot ID	Sequence Database Download (Month/Year)	No. Sequences in MSA N_{eff} (N_{eff}/L)	RMSD (Å) Relative to Reference: N, C α , C', O backbone / all C, N, O, S atoms
<i>A. tumefaciens</i> Protein of Unknown Function A9CJD6_AGRIT5 L=63	Aug 2013	10,964 (174)	1.5±0.2 / 2.2±0.2
	Apr 2017	28,265 (449)	1.8±0.2 / 2.4±0.1
<i>E. carotovora</i> Cold-shock-likeprotein Q6D6V0_ERWCT L=63	Aug 2013	4,410 (70)	1.9±0.3 / 2.9±0.3
	Apr 2017	7,107 (113)	1.7±0.6 / 2.6±0.4
<i>A. thaliana</i> Ubiquitin-like domain Q9ZV63_ARATH L=73	Aug 2013	4,964 (68)	1.4±0.1 / 2.0±0.1
	Apr 2017	5,396 (74)	1.5±0.1 / 2.4±0.3
<i>R. metallidurans</i> Rmet5065 Q1LD49_RALME L=131	Aug 2013	2,620 (20)	1.9±0.3 / 3.0±0.2
	Apr 2017	31,674 (241)	1.7±0.3 / 2.9±0.2
<i>E. coli</i> Maltose Binding Protein MALE_ECOLI Full-length ¹ (396 residues) L=388	Aug 2013	12,416 (32)	2.9±0.4 / 3.5±0.4
	Apr 2017	43,759 (112)	2.5±0.3 / 3.2±0.3
<i>E. coli</i> Maltose Binding Protein MALE_ECOLI N-terminal domain ²	Aug 2013		1.6±0.1 / 2.5±0.2
	Apr 2017		1.8±0.2 / 2.7±0.2
<i>E. coli</i> Maltose Binding Protein MALE_ECOLI C-terminal domain ³	Aug 2013		1.9±0.3 / 2.7±0.2
	Apr 2017		1.7±0.2 / 2.6±0.2

¹Residues superimposed: 1–370.²Residues superimposed: 1–112; 259–329.³Residues superimposed: 113–258; 330–370.

Table 3.Assessment of the accuracy of well-defined, buried side chain χ_1 dihedral angles.

Protein NMR Data Set	Reference NMR Structure	Number of buried, well-defined sidechains ¹	χ^I rotamer agreement (%)
A9CJD6_AGRT5	2K2P	20	84
Q6D6V0_ERWCT	2K5N	17	75
Q9ZV63_ARATH	2KAN	21	73
Q1LD49_RALME	2LCG	38	85

Maltose Binding Protein					
NMR Structure	Number of buried, well-defined, side chains ¹	χ^I rotamer agreement (%)	Number of common buried, well-defined sidechains ¹	χ^I rotamer agreement (%)	RMSD to X-ray crystal structure ² Full-length / NTD / CTD (Å)
2D21	105	76	15	57	5.4 / 1.6 / 1.5
1EZP	33	26	15	23	3.3 / 2.8 / 2.6
2MV0	80	75	15	60	4.7 / 2.0 / 3.7
EC-NMR	102	73	15	57	2.5 / 1.8 / 1.7

¹ Side chains that are buried (average SASA < 40 Å² in the NMR structures) and well-defined (χ_1 angle S.D. < 30 degrees in the NMR ensemble).¹ Side chains that are buried (SASA < 40 Å² in the X-ray structure) and well-defined (χ_1 angle S.D. < 30 degrees in the NMR ensemble).² The reference X-ray crystal structure is PDB ID 1DMB.