Published in final edited form as:

Methods Enzymol. 2019 ; 614: 363–392. doi:10.1016/bs.mie.2018.11.004.

Combining Evolutionary Co-variance and NMR Data for Protein Structure Determination

Yuanpeng Janet Huang^{1,2}, Kelly P. Brock³, Yojiro Ishida^{1,4}, G.V.T. Swapna^{1,2}, Masayori Inouye^{1,4}, Debora S. Marks³, Chris Sander⁵, and Gaetano T. Montelione^{1,2,4,*}

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

²Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ, 08854, U.S.A

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

⁴Department of Biochemistry and Molecular Biology, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, Piscataway, NJ, 08854, U.S.A

⁵Department of Cell Biology, Harvard Medical School and cBio Center, Dana-Farber Cancer Institute

Abstract

Accurate protein structure determination by solution-state NMR is challenging for proteins greater than about 20 kDa, for which extensive perdeuteration is generally required, providing experimental data that is incomplete (sparse) and ambiguous. However, the massive increase in evolutionary sequence information coupled with advances in methods for sequence covariance analysis can provide reliable residue-residue contact information for a protein from sequence data alone. These “evolutionary couplings (ECs)” can be combined with sparse NMR data to determine accurate 3D protein structures. This hybrid “EC-NMR” method has been developed using NMR data for several soluble proteins, and validated by comparison with corresponding reference structures determined by X-ray crystallography and/or conventional NMR methods. For small proteins, only backbone resonance assignments are utilized, while for larger proteins both backbone and some sidechain methyl resonance assignments are generally required. ECs can be combined with sparse NMR data obtained on deuterated, selectively-protonated protein samples to provide structures that are more accurate and complete than those obtained using such sparse NMR data alone. EC-NMR also has significant potential for analysis of protein structures from solid-state NMR data, and for studies of integral membrane proteins. The requirement that ECs are consistent with NMR data recorded on a specific member of a protein family, under specific conditions, also allows identification of ECs that reflect alternative allosteric or excited states of the protein structure.

*To whom correspondence may be addressed.

Keywords

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

1. Introduction

One key goal of structural bioinformatics over the past few decades has been the reliable prediction of tertiary structure contacts from protein sequence data (Gobel, Sander, Schneider, & Valencia, 1994; Neher, 1994; Shindyalov, Kolchanov, & Sander, 1994; Taylor & Hatrick, 1994; Thomas, Casari, & Sander, 1996). Residues that contact one another may co-vary in the course of evolution; e.g., when a large, buried hydrophobic residue which is packed against small hydrophobic residues in the protein core is mutated to a small hydrophobic residue, the space created may be compensated by mutation of the contacting residue(s) to larger hydrophobic residues. Detection of such sequence correlations from multiple sequence alignment data can therefore potentially provide information about residue contacts, which in turn can be used to guide protein structure prediction and molecular modeling. Contact prediction has been pursued by many groups, and has been a central activity of the Critical Assessment of Protein Structure Prediction experiment (Schaarschmidt, Monastyrskyy, Kryshchuk, & Bonvin, 2018).

In interpreting contacts from sequence co-variance, a challenge is created by *transitive correlations*, or relay effects; i.e., the challenge of distinguishing A-B covariation due to A->B interactions from A-C covariation due to relayed A->B and B->C interactions. Because multiple relayed pathways are possible, co-variance due to such non-direct interactions may be more pronounced than co-variance due to direct residue-residue interactions. Unless properly weighted, these relay effects can result in hot spots of co-variance due to local networks in the protein structure (Morcos et al., 2011). These networks of correlations confound the extraction of directly-coupled residue pairs useful for contact prediction and for providing distance restraints for molecular modeling.

Motivated by the potential of using sequence-based residue covariance restraints for molecular modeling and protein structure prediction, several groups have developed maximum entropy global statistical models and maximum likelihood parameter inference methods that can reliably distinguish direct evolutionary couplings from transitive correlations, allowing analysis of direct residue-residue couplings from multiple alignments of structurally related protein sequences (Lapedes, Giraud, & Jarzynski, 2002; Marks et al., 2011; Morcos et al., 2011; Jones, Buchan, Cozzetto, & Pontil, 2012; Sulkowska, Morcos, Weigt, Hwa, & Onuchic, 2012; Ekeberg, Lovkvist, Lan, Weigt, & Aurell, 2013; Kamisetty, Ovchinnikov, & Baker, 2013). These evolutionary couplings (ECs) can provide accurate information about residue pair contacts present in the 3D structures of proteins (Marks et al., 2011; Morcos et al., 2011; Hopf et al., 2012; Jones et al., 2012; Marks, Hopf, & Sander, 2012; Sulkowska et al., 2012; Kamisetty et al., 2013). These predicted contacts can then be used, together with traditional spatial-restraint modeling methods, restrained molecular dynamics, knowledge-based, and/or energy minimization methods to model the native

structure of the protein, often providing correct identification of the protein fold (Marks et al., 2011; Hopf et al., 2012; Sulkowska et al., 2012; Ovchinnikov et al., 2015; Sheridan et al., 2015; Ovchinnikov et al., 2016; Ovchinnikov et al., 2017). ECs have also been successfully used to identify residue pair contacts at protein-protein interfaces (Hopf et al., 2014; dos Santos, Morcos, Jana, Andricopulo, & Onuchic, 2015), to predict the structures of protein-protein complexes (Hopf et al., 2014; Ovchinnikov, Kamisetty, & Baker, 2014), and to model tertiary structures of RNA molecules (Weinreb et al., 2016). The ability to identify direct co-variance couplings from multiple sequence alignments is a significant advance, which is having a huge impact on the accuracy of template-free protein structure predictions, as assessed by the Critical Assessment of Protein Structure Prediction (CASP) (Ovchinnikov et al., 2015; Abriata, Tamo, Monastyrskyy, Kryshchak, & Dal Peraro, 2018; Schaarschmidt et al., 2018). For these reasons, contact prediction has evolved into a very important area of structural bioinformatics research (Simkovic, Ovchinnikov, Baker, & Rigden, 2017; Abriata et al., 2018; Adhikari, Hou, & Cheng, 2018; Buchan & Jones, 2018; Kassem, Christoffersen, Cavalli, & Lindorff-Larsen, 2018; Schaarschmidt et al., 2018; Wang, Sun, & Xu, 2018).

In addition to their role in modeling the native structures of proteins, protein complexes, and RNA, it is important to also recognize that high-confidence ECs may also reflect transient protein-protein interactions (Toth-Petroczy et al., 2016), alternative conformational or allosteric states (Morcos, Jana, Hwa, & Onuchic, 2013; Toth-Petroczy et al., 2016), and/or more subtle features of protein structural dynamics. Co-variance pairs that are “false positives” with respect to the lowest energy conformational state of the protein structure may alternatively provide critical information about protein dynamics and/or function.

As outlined in our recent monograph (Huang, Brock, Sander, Marks, & Montelione, in press), while a breakthrough in the field of protein structure prediction, the modeling of 3D structures from evolutionary coupling data alone has a number of limitations in generating accurate protein models. ECs are derived from multiple sequence alignments, providing information about contacts that are conserved across a protein or protein domain family. These contacts may not all be present in the particular member of the protein family under investigation. Hence, “false positive” contact predictions may result from “structural drift” through the course of evolution. Some predicted contacts which do indeed occur in homologues may be inconsistent with the native structure of the subject protein, and if satisfied will result in an erroneous model. In addition, even highly-ranked ECs generated from extensive multiple sequence alignment data may not reflect direct structural contacts in the predominantly populated native protein structure, as sequence covariance may result from contacts in alternative conformational states adopted by the target protein, allosteric networks, homo-oligomerization, and/or indirect residue interactions that are mediated through substrates or binding partners. They may also result from simple false positives in the parameter inference computation, especially when insufficiently diverse sequences are available across the entire protein structure of interest, or when the sequence coverage is only sufficient for EC prediction over a segment of the target protein sequence. As a result, EC-derived models of proteins derived exclusively from EC data may be distorted, or differ in details, from the predominant native structure.

Solution-state NMR can generally provide accurate three-dimensional (3D) structures of small (MW < ~ 20 kDa) proteins (Mao, Guan, & Montelione, 2011; Mao, Tejero, Baker, & Montelione, 2014). However, for larger proteins efficient ^1H - ^1H transverse spin relaxation 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 , ^{15}N , and ^{13}C nuclei, increasing the sensitivity and feasibility of NMR for larger proteins (Gardner, Rosen, & Kay, 1997). However, perdeuteration also reduces the number of ^1H 's providing ^1H - ^1H NOEs, providing much fewer structural restraints. This incompleteness of NOE data can be compensated to some degree using dihedral restraints based on chemical shift and orientation restraints from residual dipolar coupling (RDC) data. Protein structure models based on such "sparse NMR data" can also be improved using advanced knowledge-based molecular modeling methods (Raman et al., 2010; Lange et al., 2012; Sgourakis et al., 2014). However, the protein structures derived from perdeuterated samples with limited sidechain resonance assignments are generally less accurate and precise than those obtained for smaller, fully-protonated proteins with complete sidechain resonance assignments.

In recent work (Tang et al., 2015), we have demonstrated that spatial contact information derived from sparse NMR data and from ECs can provide highly complementary information. This creates the opportunity to combine NMR and EC data for more reliable structure determination than can be achieved using either data type alone. This hybrid approach is referred to as "EC-NMR". The method is particularly useful for proteins which cannot be modeled from a homologous template structure, and for which only sparse NMR data can be obtained.

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 the dynamic states accessible to the protein structure at the temperature of study. Predicted contacts, if reliable, can complement this spectroscopic information to provide missing contact information, and more accurate models. However, the ECs potentially predict 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 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 in the solution conditions under investigation* (Tang et al., 2015).

2. The EC-NMR Algorithm and Implementation

2.1. General Approach.

The general EC-NMR approach as proposed by Tang et al. (2015) is outlined in Fig. 1. The overall process can be divided into three stages. Stage 1 involves generation of ECs from multiple sequence alignments using maximum entropy or pseudolikelihood models of the protein sequence (Marks et al., 2011; Morcos et al., 2011; Jones et al., 2012; Ekeberg et al., 2013; Kamisetty et al., 2013). In Stage 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 methyl groups (Rosen et al., 1996; Gardner et al., 1997; Tugarinov, Kanelis, & Kay, 2006). These data are used to determine backbone $^1\text{H}^{\text{N}}$, ^{13}C , and ^{15}N , as well as sidechain amide $^1\text{H}^{\text{N}}$ - ^{15}N and some methyl $^{13}\text{CH}_3$ resonance assignments. Backbone dihedral angle restraints may also be generated from the $^{13}\text{C}^{\alpha}$ and $^{13}\text{C}^{\beta}$ chemical shift data using programs like TALOS-N (Shen & Bax, 2015). Unassigned NOESY peak lists are then compiled from simultaneous 3D ^{15}N , ^{13}C -NOESY spectra. In some cases, ^{15}N - ^1H residual dipolar coupling (RDC) data are also measured using one or more RDC alignment media. In Stage 3, automated NOESY analysis software is used to identify and iteratively refine a list of distance restraints using both EC and NMR data simultaneously. These formatted distance restraints are then used to generate a small ensemble of 3D structures. In this process, 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/or intermediate 3D structures, and to generate initial 3D models of the protein.

A key component of this process is the generation of accurate multiple sequence alignments (MSAs) for large numbers of protein sequences. In generating these MSAs, it is important to choose an appropriate range of evolutionary neighbors: not too distant, so as to optimize specificity of structural constraints to the target of interest, and not too restricted, so as to retrieve as many sequences as possible at maximum sequence diversity (Marks et al., 2012; Huang et al., in press). It is also critical to suppress transitive (indirect) correlations using methods such as global entropy maximization, pseudolikelihood scoring, or machine learning (Lapedes et al., 2002; Marks et al., 2011; Morcos et al., 2011; Jones et al., 2012; Sulkowska et al., 2012; Ekeberg et al., 2013; Kamisetty et al., 2013). In our published implementation of EC-NMR (Tang et al., 2015; Huang et al., in press), evolutionary residue-residue couplings were computed using a pseudo-likelihood maximization algorithm (Ekeberg et al., 2013) implemented in the computer program *plmc*, part of the *EVcouplings* software suite (<https://github.com/debbiemarkslab/plmc>). However, other methods of contact prediction may also be used to generate ECs (Marks et al., 2011; Morcos et al., 2011; Jones et al., 2012; Sulkowska et al., 2012; Kamisetty et al., 2013; Adhikari et al., 2018; Buchan & Jones, 2018; Kassem et al., 2018; Wang et al., 2018).

Although alternative implementations may be anticipated, the automated combined analysis of NMR and EC data has been implemented in the NOESY assignment program *ASDP* (formally AutoStructure) (Huang, Tejero, Powers, & Montelione, 2006; Huang, Mao, Xu, & Montelione, 2015; Tang et al., 2015). NMR and EC data are combined to generate distance restraints. Intermediate 3D structures are then generated from these distance restraints provided by *ASDP* using the program *CYANA* (Herrmann, Guntert, & Wuthrich, 2002), although other programs (e.g. NIH-XPLOR (Schwieters, Kuszewski, & Clore, 2006), CHARMM (Brooks et al., 2009), CNS (Adams et al., 2013), Rosetta (Lange & Baker, 2012; Lange et al., 2012), or ARIA (Nilges, 1995)) could be used to generate 3D structures. The resulting residue-pair contacts, derived by the combined analysis of EC and NMR data, are then deconvoluted into atom-specific distance restraints, which are then used to refine the protein structure using restrained energy minimization. In our published implementation (Tang et al., 2015), the refinement step used a specific restrained energy minimization and

knowledge-based modeling protocol with the program *Rosetta*, as described by Mao et al. (Mao et al., 2014), but alternative energy refinement protocols could also be used.

2.2. Perdeuterated Protein Sample Preparation and NMR Data Collection

Sparse NMR data (i.e. involving only backbone and some sidechain resonances) 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). In our laboratory, isotope-enriched proteins, and particularly perdeuterated ^{13}C , ^{15}N -enriched proteins, are generally produced using the condensed-phase *E. coli* Single Protein Production (SPP) system. SPP provides high yields of perdeuterated soluble and membrane proteins with Ile(δ 1), Leu, Val, and Ala methyl labeling in small volumes of defined media at reduced cost (Suzuki et al. 2007; Schneider et al. 2009; Schneider et al. 2010; Vaiphei et al. 2011). Backbone resonance assignments are determined using standard triple-resonance method with deuterium decoupling. NOESY data collection includes standard 3D N-NOESY, 3D C-NOESY (or 3D simultaneous N,C-NOESY), and for larger proteins HNCH and HCCH pseudo 4D NOESY (Diercks, Coles, & Kessler, 1999; Lange et al., 2012) data.

2.3. Calculation of Evolutionary Couplings and Confidence Scores

In our published work, evolutionary couplings (ECs) were calculated using the *EVfold-plm* pipeline available at evfold.org, as described previously (Hopf et al., 2012). For structure modeling using ECs alone, EC pairs that are inconsistent with secondary structure prediction were removed (Hopf et al., 2012).

To assess the confidence of EC pairs, Hopf et al. (Hopf et al., 2014) have defined a score Q that measures how much the corresponding EC score is an outlier from the approximately Gaussian distribution expected from non-informative background coupling between the majority of positions. This measure depends on the shape of the EC scores distribution, which can be approximated by a mixed Gaussian-lognormal model, and has been shown to be a strong predictor of the accuracy of ECs (Hopf et al., 2014). Toth-Petroczy et al. (Toth-Petroczy et al., 2016) have defined a related metric for defining “significant” ECs, the posterior probability p of membership in lognormal component of a Gaussian-lognormal mixed model fit to the distribution. In this measure, ECs with $p \geq 0.9$ (i.e., 90% probability to be in the lognormal tail of the distribution) are considered to be significant. In our published work (Tang et al., 2015), we defined *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; while for the contact maps presented here we plot ECs with $p \geq 0.9$.

2.4. Implementation in the Automated NOESY Assignment Program ASDP

The EC-NMR method is implemented within the automated NOESY cross peak assignment program *ASDP* (Huang et al., 2006) (version 2.0), available from <http://ec-nmr.nesg.org/>. The following steps of the iterative EC-NMR analysis process (Tang et al., 2015) are outlined in Fig. 2. Additional details are provided in the original paper.

Step 1. Generate de novo initial NOE-based distance constraints and EC-based ambiguous Residue Pair Contact (RPC) distance restraints. Here, initial NOE-based distance restraints are generated from NOESY peak list and chemical shift data using algorithms encoded in the *ASDP* program (Huang et al., 2006). Secondary structures, including beta-strand alignments, are identified using previously described algorithms (Huang et al., 2006) that utilize chemical shift data (Wishart & Sykes, 1994) together with characteristic NOE patterns within and between secondary structure elements (Wuthrich, 1986). Additional NOE assignments are then ruled in and ruled out by *ASDP* based on uniqueness relative to the chemical shift list, NOESY cross-peak symmetry patterns, and the extensive network anchoring algorithms of *ASDP* (Huang et al., 2006). For perdeuterated proteins, a deuterium correction to the ^{13}C chemical shifts (Maltsev, Ying, & Bax, 2012) 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. For larger (> 20 kDa) perdeuterated proteins, the NMR data also include 3D H^{N} -Me and Me-Me NOESY peak list data, which provide NOEs involving Val $^{13}\text{C}^{\gamma}\text{H}_3$, Leu $^{13}\text{C}^{\delta}\text{H}_3$, and Ile $^{13}\text{C}^{\delta 1}\text{H}_3$ methyl groups.

EC-based RPC ambiguous restraints are also generated in Step 1 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. ECs are ranked based on EC reliability scores (Sheridan et al., 2015). Only ECs with reliability score ≥ 0.5 were included in these calculations, with weights (w) initially set equal to their reliability scores. These parameters may be optimized differently for specific EC data sets.

Step 2. Initial 3D structures are generated from these initial RPC restraints. One hundred decoy models are generated with distance restraints output by *ASDP* using the program *CYANA* (Herrmann et al., 2002). Specifically, input to *CYANA* (or an alternative structure generation program) includes ^1H - ^{15}N RDC data (if available), dihedral angle restraints generated from backbone chemical shift (Shen, Delaglio, Cornilescu, & Bax, 2009), together with unique NOE-based and EC-based ambiguous RPC distance restraints, identified by *ASDP* in Step 1 above.

The standard protocol of *Talos+* or *TalosN* (Shen et al., 2009) is used to generate backbone dihedral angle restraints based on $^{13}\text{C}^{\alpha}$ and $^{13}\text{C}^{\beta}$ chemical shifts, and residues with “good” *Talos* scores (e.g., *Talos+* reliability score of 10) are restrained to phi and psi ranges of $\pm 20^\circ$. A deuterium correction to ^{13}C chemical shifts is also applied in *Talos* analysis for perdeuterated proteins (Shen et al., 2009).

Step 3. Initial 3D structures are used to rule in / rule out NOESY peak assignments, identify inconsistent EC contacts, and to define additional Residue Contact Pairs (RCPs). The “top” 20 models generated by *CYANA* are identified using a combined score comprised of the *NMR RPF* Recall (Huang, Powers, & Montelione, 2005) 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 “best scoring” 3D decoy structures are then used by *ASDP* to rule in and rule out potential NOE and EC assignments.

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 restraint list. Structurally inconsistent NOE assignments from Step 1 are excluded using network anchoring algorithms describe in original *AutoStructure* paper (Huang et al., 2006). Structurally inconsistent RPCs (SI-RPCs) are also identified when ambiguous RPC distance constraints are violated by $> 0.5 \text{ \AA}$ in more than 60% (e.g. 12 of 20) conformers. These SI-RPCs are excluded from the next cycle of structure generation calculations. Ambiguous RPC constraints that are satisfied by all 20 decoy conformers (i.e. no violation $> 0.5 \text{ \AA}$) 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 this process, the *ASDP* software also identifies new RPCs, which are long-range residue pairs (i.e. $|i-j| \geq 5$) that have at least one inter-atom (i.e. any H, N, or C atom with a resonance assignment) distance $\geq 5 \text{ \AA}$ 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 contact predictions with low ranking scores in covariance analysis.

Steps 2–3 are then repeated two more cycles, resulting in an ensemble of 20 protein structure models.

Step 4. The protein structure models from Cycle 3 are 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. 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 $> 10 \text{ \AA}$ in all 20 conformers from *Cycle 3*. “Noise ECs” are initial ECs from Step 1 for which the corresponding ambiguous restraint is violated with distance $> 10 \text{ \AA}$ in all 20 conformers from *Cycle 3*. These “noise” NOESY cross peaks and ECs are removed from the EC-NMR restraint list.

Step 5. The resulting 20 NMR structure models are further energy refined using a standardized distance-restrained Rosetta refinement protocol (Mao et al., 2014). Specific atom-atom Rosetta distance restraints are generated for each atom pair in residue pairs of the EC list which have interatomic distance $\geq 5 \text{ \AA}$ in all 20 models. Upper-bound restraints of $7 \text{ \AA}$ are used for all of these specific atom-atom constraints, in order to allow the Rosetta force field to attain low-energy structures and to avoid generating overly constrained structures.

Variables *Cycle* and *Run* (Fig. 2) are used to control the repeated analyses of Steps 1–3. 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.

2.5. Refinement of EC NMR Structures with Rosetta.

Each of the resulting NMR structure models are then further energy refined using the restrained Rosetta refinement protocol outlined in Mao et al (Mao et al., 2014). Detailed protocols for Restrained Rosetta refinement are available on line at http://www.nmr2.buffalo.edu/nescg.wiki/Rosetta_High_Resolution_Protein_Structure_Refinement_Protocol

Atom-atom distance restraints are used by this Rosetta refinement protocol. The script *getCC.pl* from the *ASDP-2.0* package is used to generate specific atom-atom Rosetta distance restraints for each heavy atom (N, C, O, S) pair in RPCs which have interatomic distance $\leq 5 \text{ \AA}$ in all 20 models. These atom pairs are then assigned upper-bound restraints of $7 \text{ \AA}$. NOE-based upper bound distance restraints are loosened by 30% before converting to the Rosetta restraint format. Restrained Rosetta refinement (Mao et al., 2014) is then done using either a stand-alone version of *Rosetta*, or using our Restrained Rosetta Refinement Server available on line at: http://psvs-1_4-dev.nescg.org/consRosetta.html

2.6. Access to ASDP Software and Tutorial for EC-NMR Calculations.

Access to the *ASDP* software, together with a short tutorial for its use, are available on line at: http://www-nmr.cabm.rutgers.edu/NMRsoftware/asdp/Quick_Starts.html Additional instructions for using *ASDP* are available at: http://www.nmr2.buffalo.edu/nescg.wiki/AutoStructure_Structure_Determination_Program. A web-based tutorial for running EC-NMR calculations with *ASDP* is available on line at <http://ec-nmr.nescg.org/tutorial.html>. The on-line tutorial includes sample input and output data files (see On Line Methods of Tang et al, 2015).

3. Quality Control and Structure Validation

3.1. Assessment of Overall Structure Accuracy

In testing and developing the EC-NMR protocol, the primary assessment of structure accuracy was done by comparison with the corresponding X-ray crystal structure or high-quality conventional NMR structure determined using essentially complete backbone and sidechain resonance assignments. When X-ray crystal structures are used as the reference structure, hydrogens are added using the *Reduce* program of the *Molprobity* software package (Word, Lovell, Richardson, & Richardson, 1999). Backbone (defined as N, C $^{\alpha}$, C', and O atoms) and all-heavy-atom (N, C, O, S) Root Mean Square Deviations (RMSDs) are computed using the *fit* command as implemented in the *PyMol* software (DeLano, 2002), or using the molecular structure superimposition tools of the *PDBStat* software package (Tejero, Snyder, Mao, Aramini, & Montelione, 2013). Note that these two software packages use somewhat different criteria for defining the subset of atoms to use in superimpositions, and hence generally provide slightly different RMSD values for the same coordinate pairs.

Structure quality evaluation also includes a model vs data assessment. NMR DP scores (Huang et al., 2005; Huang, Rosato, Singh, & Montelione, 2012), which are crystallographic R-factor-like metrics for NMR structures, provide a global measure of how well a protein structure model fits with the NMR NOE and chemical shift data. Several calibration studies

using decoys distorted from the correct structure, or inaccurate NMR structures compared with reference structures, have demonstrated that accurate structure models generally have DP scores > 0.70 (Huang et al., 2005; Huang et al., 2012; Rosato et al., 2012). NMR DP scores are computed by the *ASDP* program, or can be assessed independently 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 normalized Structure Quality scores (Bhattacharya, Tejero, & Montelione, 2007) $Z > -2$ for *Procheck*(backbone), *Procheck*(all dihedral), *Verify3D*, *MolProbity*, and *Prosa II* knowledge-based structure quality assessment metrics. Structure quality Z scores are 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 also available at <http://www.nmr2.buffalo.edu/nesg.wiki/PSVS>.

3.2. Assessment of Core Sidechain Accuracy

In assessing NMR-derived model accuracy relative to reference structures, it is also useful to compare sidechain conformations. One useful metric is the comparison of χ_1 rotamer states. In this analysis, χ_1 rotamers for all residues in each reference X-ray crystal structure are assigned to the nearest g^+ , t or g^- conformational state. Generally, we limit this analysis to “buried” side chains, with solvent accessible surface area (SASA) less than $40 \text{ \AA}^2$ in the reference X-ray crystal structure [calculated using the *Molmol* program (Koradi, Billeter, & Wuthrich, 1996)], as surface sidechains are generally not limited to a single conformation. In considering NMR structure ensembles (e.g., the EC-NMR structure or a NMR reference structure obtained from the PDB), side chains whose χ_1 dihedral angle values have standard deviations of < 30 degrees are considered as “converged side chains”. To assess sidechain conformation agreement between model and reference structures, rotamer states for residues with both buried and converged side chains are compared between the reference X-ray crystal (or the ‘representative’ NMR conformer) and each member of the ensemble of NMR structures. The representative model is defined as the medoid conformer in the NMR ensemble; that is the one conformer most like all of the others (Montelione et al., 2013; Tejero et al., 2013)]. This representative conformer of the ensemble can be selected using *PDBStat* (Tejero et al., 2013).

3.3. Assessment of False Positive Rates in Contact Predictions

The Precision (P), Recall (R), and Performance (F) statistics for contact prediction (ECs or RPCs) are computed relative to reference X-ray or NMR structures as:

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

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

$$\text{F-measure (F)} = (2 \times \text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision}) \quad \text{Eqn. 3}$$

In this analysis a TP contact is defined for residue pair (i,j) if any atom of residue *i* is $\geq 5 \text{ \AA}$ 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.

4. Representative EC-NMR Results

4.1. Assessment on Eight Reference Structures

The EC-NMR method was initially tested using experimental NMR (chemical shift, NOESY, and RDC data) for eight proteins ranging in size from 6 to 41 kDa (Tang et al., 2015). The resulting EC-NMR structures were compared with “reference structures”, determined either by X-ray crystallography or by NMR using essentially complete backbone and sidechain resonance assignments. The EC-NMR structures of smaller ($< 20 \text{ kDa}$) 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. Larger proteins, including human p21 H-Ras (21 kDa) and maltose binding protein (41 kDa) 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). The percentages of χ_1 states for buried and converged sidechains that are consistent between the representative NMR structure selected from the ensemble [defined as the medoid conformer in the NMR ensemble (Montelione et al., 2013; Tejero et al., 2013)] and the reference X-ray crystal were also assessed. Averaged over all 8 EC-NMR structures, ~80% of buried, well-defined side chains have χ_1 rotamers that match the corresponding reference structures (Tang et al., 2015). For the 3 largest proteins studied, 85%, 81%, and 65% of the buried, well-defined side chains have χ_1 rotamers that match the corresponding X-ray crystal structures. Generally speaking, the EC-NMR structures determined using HN only, or HN and Me, NOEs are quite accurate, based on comparisons with the corresponding reference structures (Tang et al., 2015).

4.2. Impact of Updated Database of Protein Sequences

Recently, five of the EC-NMR structures originally described by Tang et al. (2015) were redetermined using the same archived NMR data, but an updated database of protein sequences (Huang et al., in press). For the four smaller protein targets, with molecular weights of 6 to 15 kDa, the NMR data include only $\text{H}^{\text{N}}\text{-H}^{\text{N}}$ NOE data, along with restraints on backbone dihedral angles computed from $\text{C}^{\alpha}/\text{C}^{\beta}$ chemical shifts using Talos-N (Shen & Bax, 2015). For two of these four proteins, $^{15}\text{N}\text{-}^1\text{H}$ RDCs were available in two different molecular alignment conditions, for a third $^{15}\text{N}\text{-}^1\text{H}$ RDCs were available only one alignment condition, and for the fourth no RDC data are available. EC's were generated using the *EVfold* software suite. Contact maps for the most confident ECs, with posterior probability $p > 0.9$, are shown in Fig. 3 together with the contact maps of the corresponding reference

structures. EC-NMR models were generated with the *ASDP* / *Cyana* / *Rosetta* pipeline described above in Sections 2.4 and 2.5 above, and compared with the reference NMR structures determined with complete sidechain proton assignments and much more extensive NOESY data (Fig. 4). They were assessed based on (i) accuracy of atomic positions 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. 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, using *PDBStat* (Tejero et al., 2013). The backbone RMSD's between these EC-NMR structures and the corresponding reference 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 (Huang et al. in press). The χ_1 values of well-defined buried sidechain, 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. These results are similar to those reported for 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. These EC-NMR structures determined using only H^N - H^N NOE data together with ECs (and no methyl resonance assignments) have accuracies that compare with high quality NMR structures determined with complete backbone and sidechain resonance assignments, demonstrating 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.

4.3. EC-NMR Structure of 41 kDa Maltose Binding Protein

Tang et al. (2015) and Huang et al. (in press) also generated EC-NMR structures 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($\delta 1$), Leu, and Val methyl proton assignments, providing also Me-Me and HN-Me NOEs, along with restraints on backbone dihedral angles computed from C^α/C^β chemical shifts using Talos-N (Shen & Bax, 2015) and H^N - ^{15}N RDC data in a single alignment media (Lange et al., 2012). These results (Fig. 5) demonstrate high-quality EC-NMR structures of MBP 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) 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 (Tang et al., 2015). The core sidechains of the EC-NMR structure are significantly more accurate than PDB ID 1EZP, 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. 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.

4.4. Sensitivity to Numbers of Sequence Homologs

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 $2 \times L$ 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; Kamisetty et al., 2013; Hopf et al., 2014; Michel et al., 2014; Ovchinnikov et al., 2014; 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.

For EC-NMR, the 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. In order to test the robustness of structure accuracy to the numbers of sequences used in the MSA, Tang et al. (2015) analyzed a series of MSAs, each with numbers of sequences ranging from $N_{\text{eff}}/L \sim 150$ down to $N_{\text{eff}}/L < 0.1$. This study is reproduced from the original publication in Fig. 6. In this analysis, using good quality NMR data for a perdeuterated, Ile($\delta 1$), 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 .

More recently we have explored other real EC-NMR data sets with N_{eff}/L ranging from 6 to 0.1. Using our current implementation of the EC-NMR software pipeline, we confirmed the production of accurate models for data with $N_{\text{eff}}/L \sim 6$, but failure of the method for $N_{\text{eff}}/L < 1$. Although we anticipate continued growth of sequence data bases, and the quality of results will also depend on the quality and extent of the sparse NMR data, it remains an important benchmark for the EC-NMR to provide accurate models for targets with low numbers (e.g., $N_{\text{eff}}/L < 5$) of sequence homologs.

4.5. Structural Restraints Can Guide Identification of False Positive Contact Predictions

EC-based contacts can complement sparse NMR data to provide more contact information, and more accurate models. However, EC lists will generally include predicted interactions that are false positive contacts (FPs) for the predominant low-energy structure of the target protein. The requirement that the structure model be consistent with all (or most) of the

experimental NMR data, however, provides “hard” constraints on the interpretation of ECs, allowing identification and removal of FP structurally-inconsistent residue pair contacts (SI-RPCs) (Tang et al., 2015). In all of the calculations we carried out, the EC-NMR protocol increased the number of ECs that could be used reliably in structure calculations and reduced the FP contact rates (Tang et al., 2015; Huang et al., in press). For example, for MBP, the number of FPs was reduced from 127 to 82 and the Precision of the ECs (relative to the reference X-ray crystal structure) was increased from 0.66 to 0.73 (see Tang et al, 2015, Supplementary Table 4B). The remaining false-positive EC contacts remaining at the end of the EC-NMR protocol are nearby in the 3D structures to true-positive inter-residue contacts, and have little effect on the accuracy of the final structure. Similarly, for the other EC-NMR structures that were calculated, all of the FPs that are significantly inconsistent with the reference structures were eliminated. Hence structural constraints that evolve in course of EC-NMR protocol are very effective in removing structurally-inaccurate FPs from the EC lists. Similar results have been recently reported using a different algorithm for imposing structural consistency in identifying FPs in EC analysis, based on the automated NOESY peak assignment program YAPP (Kassem et al., 2018).

5. Future Prospects

Contact predictions 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 predicted EC-based contacts are consistent with internal structural constraints and experimental NMR data collected for one member of the protein family under specific conditions. This requirement eliminates potentially important, but confounding, EC-derived contact restraints that may arise from structural drift across the protein family, or from transient oligomeric interactions, allosteric networks and/or alternative states which may also be detected as evolutionary sequence co-variation. 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 requires large multiple sequence alignments, which are currently available only for a fraction of 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.

The EC-NMR method is **largely automated**, and provides high-quality 3D structures with accurate backbone and core sidechain conformations (Tang et al., 2015; Huang et al., in press). **For small proteins** and domains up to 150 residues (< ~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 impossible with current methods, 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 solution 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). The EC-NMR method should be particularly powerful for modeling structures of biomolecules using solid-state NMR data.

In our work to date, we have 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, more sophisticated spatial-restrained modeling methods, such as the Resolution Adapted Structural REcombination protocol (RASREC) developed by Lange and Baker (Raman et al., 2010; Lange & 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 with EC data (Braun, Koehler Leman, & Lange, 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 the structure 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 that are accessible to the protein, excited states, and potentially provide information on allosteric networks, as residue contacts in these alternative states may also co-vary in the course of evolution. 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

- Abriata LA, Tamo GE, Monastyrskyy B, Kryshchuk A, & Dal Peraro M (2018). Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins*, 86 Suppl 1, 97–112. doi:10.1002/prot.25423 [PubMed: 29139163]
- Adams PD, Baker D, Brunger AT, Das R, DiMaio F, Read RJ, ... Terwilliger TC (2013). Advances, interactions, and future developments in the CNS, Phenix, and Rosetta structural biology software systems. *Annual Rev. Biophysics*, 42, 265–287. doi:10.1146/annurev-biophys-083012-130253
- Adhikari B, Hou J, & Cheng J (2018). Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning. *Proteins*, 86 Suppl 1, 84–96. doi:10.1002/prot.25405 [PubMed: 29047157]

- Bhattacharya A, Tejero R, & Montelione GT (2007). Evaluating protein structures determined by structural genomics consortia. *Proteins*, 66(4), 778–795. doi:10.1002/prot.21165 [PubMed: 17186527]
- 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
- Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, ... Karplus M (2009). CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30, 1545–1614. [PubMed: 19444816]
- Buchan DWA, & Jones DT (2018). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins*, 86 Suppl 1, 78–83. doi:10.1002/prot.25379 [PubMed: 28901583]
- DeLano WL (2002). The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC The PyMOL molecular graphics system, Version 1.5.0.4 Schrödinger, LLC.
- Diercks T, Coles M, & Kessler H (1999). An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments. *J. Biomol. NMR*, 15(2), 177–180. doi:10.1023/A:1008367912535 [PubMed: 20872110]
- dos Santos RN, Morcos F, Jana B, Andricopulo AD, & Onuchic JN (2015). Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Reports*, 5, 13652. doi:10.1038/srep13652
- 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, Statistical, nonlinear, and soft matter physics*, 87, 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. *J. Mol. Biol.*, 309, 961–974. doi:10.1006/jmbi.2001.4695 [PubMed: 11399072]
- Gardner KH, Rosen MK, & Kay LE (1997). Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry*, 36, 1389–1401. [PubMed: 9063887]
- Gobel U, Sander C, Schneider R, & Valencia A (1994). Correlated mutations and residue contacts in proteins. *Proteins*, 18, 309–317. doi:10.1002/prot.340180402 [PubMed: 8208723]
- 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. *J. Mol. Biol.*, 319, 209–227. [PubMed: 12051947]
- Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, & Marks DS (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149, 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*, 3, e03430. doi: 10.7554/eLife.03430.
- Huang YJ, Brock K, Sander C, Marks DS, & Montelione GT (in press). A hybrid approach for protein structure determination combining sparse NMR with evolutionary coupling sequence data In JL Markley S Burley H Nakamura G Kleywegt (Eds.), *Integrative structural biology with hybrid methods*. Singapore: Springer.
- Huang YJ, Mao B, Xu F, & Montelione GT (2015). Guiding automated NMR structure determination using a global optimization metric, the NMR DP score. *J. Biomol. NMR*, 62, 439–451. doi: 10.1007/s10858-015-9955-2 [PubMed: 26081575]
- Huang YJ, Powers R, & Montelione GT (2005). Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J. Amer. Chem. Soc.*, 127, 1665–1674. doi:10.1021/ja047109h [PubMed: 15701001]
- Huang YJ, Rosato A, Singh G, & Montelione GT (2012). RPF: a quality assessment tool for protein NMR structures. *Nucleic Acids Res.*, 40, W542–546. doi:10.1093/nar/gks373 [PubMed: 22570414]
- Huang YJ, Tejero R, Powers R, & Montelione GT (2006). A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins*, 62, 587–603. 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, 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, 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. USA*, 110, 15674–15679. doi:10.1073/pnas.1314045110 [PubMed: 24009338]
- Kassem MM, Christoffersen LB, Cavalli A, & Lindorff-Larsen K (2018). Enhancing coevolution-based contact prediction by imposing structural self-consistency of the contacts. *Sci. Reports*, 8, 11112. doi:10.1038/s41598-018-29357-y
- Koradi R, Billeter M, & Wuthrich K (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics*, 14, 51–55, 29–32.
- Lange OF, & Baker D (2012). Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins*, 80, 884–895. [PubMed: 22423358]
- Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, ... 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. USA*, 109, 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. *arXiv*, 1207.2484.
- Laskowski RA, Moss DS, & Thornton JM (1993). Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.*, 231, 1049–1067. doi:10.1006/jmbi.1993.1351 [PubMed: 8515464]
- Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, ... Richardson DC (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*, 50, 437–450. doi:10.1002/prot.10286 [PubMed: 12557186]
- Luthy R, Bowie JU, & Eisenberg D (1992). Assessment of protein models with three-dimensional profiles. *Nature*, 356, 83–85. doi:10.1038/356083a0 [PubMed: 1538787]
- Maltsev AS, Ying J, & Bax A (2012). Deuterium isotope shifts for backbone (1)H, (1)(5)N and (1)(3)C nuclei in intrinsically disordered protein alpha-synuclein. *J. Biomol. NMR*, 54, 181–191. doi:10.1007/s10858-012-9666-x [PubMed: 22960996]
- Mao B, Guan R, & Montelione GT (2011). Improved technologies now routinely provide protein NMR structures useful for molecular replacement. *Structure*, 19, 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. Amer. Chem. Soc.*, 136, 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, 10.1371/journal.pone.0028766
- Marks DS, Hopf TA, & Sander C (2012). Protein structure prediction from sequence variation. *Nature Biotechnology*, 30, 1072–1080. doi:10.1038/nbt.2419
- Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, & Elofsson A (2014). PconsFold: improved contact predictions improve protein models. *Bioinformatics*, 30, i482–i488. doi:10.1093/bioinformatics/btu458 [PubMed: 25161237]
- Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, ... Markley JL (2013). Recommendations of the wwPDB NMR Validation Task Force. *Structure*, 21, 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. USA*, 110, 20533–20538. doi:10.1073/pnas.1315625110 [PubMed: 24297889]

- Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, ... Weigt M (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA*, 108, 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. *J. Mol. Biol.*, 300, 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. USA*, 91, 98–102. [PubMed: 8278414]
- Nilges M (1995). Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.*, 245, 645–660. [PubMed: 7844833]
- Ovchinnikov S, Kamisetty H, & Baker D (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, 3, 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, 67–75. doi:10.1002/prot.24974 [PubMed: 26677056]
- Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, ... Baker D (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife*, 4, 10.7554/eLife.09248
- Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, ... Baker D (2017). Protein structure determination using metagenome sequence data. *Science*, 355, 294–298. doi:10.1126/science.aah4043 [PubMed: 28104891]
- Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, ... Baker D (2010). NMR structure determination for larger proteins using backbone-only data. *Science*, 327, 1014–1018. doi:10.1126/science.1183649 [PubMed: 20133520]
- Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, ... Bonvin AM (2012). Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure*, 20, 227–236. doi:10.1016/j.str.2012.01.002 [PubMed: 22325772]
- Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, & Kay LE (1996). Selective methyl group protonation of perdeuterated proteins. *J. Mol. Biol.*, 263, 627–636. [PubMed: 8947563]
- Schaarschmidt J, Monastyrskyy B, Kryshchuk A, & Bonvin A (2018). Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins*, 86 Suppl 1, 51–66. doi:10.1002/prot.25407 [PubMed: 29071738]
- Schwieters CD, Kuszewski JJ, & Clore GM (2006). Using Xplor-NIH for NMR molecular structure determination. *Prog. NMR Spectroscopy*, 48, 47–62. doi:10.1016/J.Pnmrs.2005.10.001
- 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 f1r1 immune modulator family. *Structure*, 22, 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 Mol. Biol.*, 1260, 17–32. doi:10.1007/978-1-4939-2239-0_2 [PubMed: 25502373]
- Shen Y, Delaglio F, Cornilescu G, & Bax A (2009). TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR*, 44, 213–223. doi:10.1007/s10858-009-9333-z [PubMed: 19548092]
- Sheridan R, Fieldhouse RJ, Hayat S, Sun Y, Antipin Y, Yang L, ... Sander C (2015). EVfold.org: Evolutionary couplings and protein 3D structure prediction. *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 Engineering*, 7, 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

- Sippl MJ (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17, 355–362. doi:10.1002/prot.340170404 [PubMed: 8108378]
- Sulkowska JI, Morcos F, Weigt M, Hwa T, & Onuchic JN (2012). Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A.* 109, 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. *Nature Methods*, 12, 751–754. doi:10.1038/nmeth.3455 [PubMed: 26121406]
- Taylor WR, & Hatrick K (1994). Compensating changes in protein multiple sequence alignments. *Protein Engineering*, 7, 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*, 56:337–51. 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, 158–170. doi:10.1016/j.cell.2016.09.010 [PubMed: 27662088]
- 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, 749–754. doi:nprot.2006.101 [pii] 10.1038/nprot.2006.101 [PubMed: 17406304]
- Wang S, Sun S, & Xu J (2018). Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins*, 86 Suppl 1, 67–77. doi:10.1002/prot.25377 [PubMed: 28845538]
- Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, & Marks DS (2016). 3D RNA and functional interactions from evolutionary couplings. *Cell*, 165, 963–975. doi:10.1016/j.cell.2016.03.030 [PubMed: 27087444]
- Wishart DS, & Sykes BD (1994). The ¹³C chemical-shift index: a simple method for the identification of protein secondary structure using ¹³C chemical-shift data. *J. Biomol. NMR*, 4, 171–180. [PubMed: 8019132]
- Word JM, Lovell SC, Richardson JS, & Richardson DC (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol*, 285(4), 1735–1747. doi:10.1006/jmbi.1998.2401 [PubMed: 9917408]
- Wuthrich K (1986). *NMR of proteins and nucleic acids*. New York: Wiley.

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

The hybrid EC-NMR strategy combines Evolutionary Coupling (EC) information from multiple sequence alignments with sparse experimental nuclear magnetic resonance (NMR) data. Reproduced from Tang et al, 2015.

Fig. 2. Flow chart of iterative EC-NMR analysis process.

The EC-NMR analysis process has been fully automated by integration into the automated NOESY NMR data assignment program *ASDP* (Huang et al., 2005; Huang et al., 2006). NOESY cross peak, ^1H - ^{15}N resonance assignments, and ^1H - ^{13}C methyl resonance assignments (where available) are combined with high-scoring EC contacts predicted from sequence co-variance analysis. Distance restraints are identified by *ASDP* from these data, and used together with RDC data (where available) to generate structural models with the program *CYANA* (Herrmann et al., 2002). Initial structures are then used to assess false-positive EC contact predictions and false-NOESY peaks (i.e. noise peaks), and to identify consistent Residue Pair Contacts (RPCs). Once the analysis converges, these specific interatomic distance restraints are used for loop rebuilding and core repacking with the program *Rosetta* (Mao et al., 2014).

Fig. 3. Contact maps comparing ECs with X-ray crystal or conventional NMR structures. Contact maps (residue number vs residue number) are shown for the reference NMR structure determined using complete sidechain resonance assignments. Contacts in these reference structures (lt. blue dots) are shown for residue pairs with heavy atom (N, C, S or O) distances $< 5 \text{ \AA}$. The secondary structure (strands and helices) of the reference structure are shown on both the X-and Y-axes. ECs with posterior probabilities $p > 0.9$ of being in the skewed-tail of the EC score distribution are also shown as back circles; the size of these circles corresponds to the strength of the EC.

Fig. 4. EC-NMR structures determined using only H^N - H^N NOESY data superimposed on reference conventional NMR structures.

The representative (medoid) structure from the ensemble of conformers generated by the EC-NMR method (blue) is superimposed on a representative (medoid) 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. 5. EC-NMR structure of *E. coli* maltose binding protein superimposed on reference X-ray crystal structure.

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. Blue ribbon structures – final EC-NMR structure of MBP. Grey ribbons – reference X-ray crystal structure.

Fig. 6. Evaluation of structure quality statistics for 20 ensembles of EC-NMR structures. The 20 ensembles of EC NMR structures generated in the sensitivity analysis outlined in the text were assessed using structural quality metrics of the Protein Structure Validation Software Suite (PSVS) (Bhattacharya et al., 2007), including NMR DP scores (Huang et al., 2005). The top panel plots backbone RMSD to reference structures (black data points) and the Number of High-Confidence ECs (red data points), as defined in the main text, for each of the 20 EC-NMR analyses carried out using different amounts of sequence data (N_{eff}/L). The bottom panel plots DP (left axis) or knowledge-based Z score (right axis) computed with the *PSVS* software. The metrics include the NMR DP score (Huang et al., 2005) (red circles), *ProCheck* (Laskowski, Moss, & Thornton, 1993) analysis of backbone dihedral angle distributions (Procheck, green boxes) and distributions for all backbone and sidechain

dihedral angels (Procheck_all, blue diamonds), *Verify3D* (Luthy, Bowie, & Eisenberg, 1992) (magenta triangles), *Prosa II* (Sippl, 1993) (grey triangles), and *Molprobrity* atomic clash scores (Lovell et al., 2003) (black asterix). Z scores are normalized against a collection of high-resolution crystal structures (Bhattacharya et al., 2007). “Reliable” structures (backbone RMSD < 3.5 Å from X-ray crystal structure), obtained using multiple-sequence alignments of $N_{\text{eff}}/L > 5$, all have DP score > ~ 0.73 and knowledge-based Z scores > -2 (upper right quadrant), while less accurate structures all have DP < ~ 0.73 and Z-scores for Verify3D < -2. Reproduced from Tang et al, 2015.