DOI 10.1002/prot.24768

Systematic solution to homo-oligomeric structures determined by NMR

Jeffrey W. Martin[1] Pei Zhou[2] Bruce R. Donald[1,2,*]

December 8, 2014

Short title: Systematic solution to homo-oligomers

Keywords: nuclear magnetic resonance spectroscopy, protein homo-oligomers, simulated annealing, structure determination, distance restraint

[1] Department of Computer Science, Duke University, Durham NC 27708
[2] Department of Biochemistry, Duke University Medical Center, Durham NC 27710
[*] Corresponding Author: Bruce R. Donald, brd+proteins14@cs.duke.edu

Received: Mar 28, 2014; Revised: Dec 12, 2014; Accepted: Jan 12, 2015

Abstract

Protein structure determination by NMR has predominantly relied on simulated annealing-based conformational search for a converged fold using primarily distance constraints, including constraints derived from nuclear Overhauser effects (NOEs), paramagnetic relaxation enhancement (PRE), and cysteine crosslinkings. Although there is no guarantee that the converged fold represents the global minimum of the conformational space, it is generally accepted that good convergence is synonymous to the global minimum. Here, we show such a criterion breaks down in the presence of large numbers of ambiguous constraints from NMR experiments on homo-oligomeric protein complexes. A systematic evaluation of the conformational solutions that satisfy the NMR constraints of a trimeric membrane protein, DAGK, reveals 9 distinct folds, including the reported NMR and crystal structures. This result highlights the fundamental limitation of global fold determination for homo-oligomeric proteins using ambiguous distance constraints and provides a systematic solution for exhaustive enumeration of all satisfying solutions.

Introduction

Simulated annealing is a primary method for structure determination of proteins by nuclear magnetic resonance (NMR) spectroscopy [1, 2]. NMR restraints and biophysical principles are encoded into an energy function whose minimization results in models of the protein structure that satisfy the restraints. If the method consistently returns similar structures that adequately satisfy the restraints, the structural ensemble is considered well-converged and the structure determination successful, although the low restraint violation and convergence does not necessarily mean the structure is accurate [3]. The main strength of simulated annealing is its ability to transform a coarse structural model into a more refined structure with improved restraint satisfaction. Where the method falls short is its inability to exhaustively sample topologically distinct structural models. Therefore, it can become trapped in the local minima of the energy landscape, thus missing the genuine fold(s) with similar or lower energies. Further complicating the situation, even if the global minimum structure of the energy function could be obtained, small inaccuracies in the energy function (e.g. due to approximation of complex physical phenomena or misinterpretation of even a few experimental distance constraints) could cause a genuine fold to be incorrectly ranked with a higher energy than the erroneous folds. Although such a situation is considered rare when all distance constraints are uniquely assigned, the odds increase significantly in the presence of ambiguous distance restraints for structure determination of homo-oligomeric protein complexes.

Ambiguous distance restraints (ADRs) [4] refer to distance information (such as NOEs) that cannot be uniquely attributed to a single pair of atoms. Since the chemical shifts of equivalent atoms in all subunits in a homo-oligomeric complex are identical and thus indistinguishable, ADRs are unavoidable for distance measurements in trimers and higher-order homo-oligomers. We refer to this phenomenon as *subunit ambiguity* [5, 6, 7, 8]. For dimers, separating intra- vs inter-subunit NOEs using X-filtered NOESY [9] is sufficient to resolve subunit ambiguity. For trimers and higher-order oligomers, even after a distance restraint has been classified as inter-subunit, it still has at least two possible assignments and is still ambiguous. ADRs consider degenerate atom pairs by using an average function derived from a mean field approximation. Although it has been demonstrated that genuine interactions can be extracted from ADRs, these methods are prone to becoming trapped in local minima since they rely heavily on the initial fold to remove assignment ambiguity. The energy landscapes for homo-oligomers contain a large number of minima with similarly low energy, so when simulated annealing methods using ADRs become trapped in local minima, these methods can fail to report satisfying folds from other minima.

This situation is further exacerbated in the case of homo-oligomeric membrane proteins, for which dense restraint collection is often impractical [10, 11, 12, 13, 8]. In the case of Diacylglycerol Kinase from *Escherichia coli* (henceforth, simply DAGK), a membrane-associated homo-trimer, two different structures have been published. The solution NMR structure [14] of DAGK, determined using ambiguously-assigned distance restraints, possesses a domain-swapped subunit interface, while the crystal structure [15] has a subunit with a more compact conformation and without domain-swapping.

Here we show that the difference between the two structures is due to the local minimum limitations of current methodology for NMR structure determination. We demonstrate that this limitation can be mitigated by searching over topologically distinct folds using a systematic approach called *fold-operator theory.* Once an initial satisfying fold is discovered, mathematical operators transform the fold into alternate folds. The operators define a group action on the configuration space of protein folds. These alternative folds can be subsequently refined using traditional simulated annealing methods and evaluated for restraint satisfaction. Using this systematic approach, we found 48 distinct folds of DAGK, among which 9, including the published NMR and crystal folds, upon energy minimization, satisfied experimental restraints.

Methods

Schematic representation of three-dimensional structure exposes helical packing

DAGK is a transmembrane protein consisting of four helices in each of its three identical subunits (Figure 1, right). Helices 1, 2, and 3 are all roughly parallel, span the membrane, and pack together into a helical bundle. The amphiphilic SH helix floats on the cytoplasmic surface of the membrane. To clearly show the differences in helical packing between the NMR and crystal structures (PDB IDs, respectively: 2KDC, 3ZE4), we reduced the three-dimensional structures of DAGK to two-dimensional *fold schematics* (Figure 1, middle). From these schematic representations of the folds, it is easy to visualize the domain-swapped configuration of the NMR structure relative to the compact subunits of the crystal structure.

Of the deposited restraints collected for DAGK in solution, there are no inter-subunit NOEs, nor long range ($i-j > 4$) NOEs within the same subunit. Hence, the NOEs, hydrogen bond restraints, dihedral angle restraints, and RDCs primarily constrain secondary structures within each subunit. The helices SH, H1, H2, and H3 are well-restrained individually, but the inter-helical linkers are relatively unrestrained, giving little long-range information to pack the quaternary structure. The helical packing of DAGK, and hence the overall fold, is largely defined by the inter-subunit restraints: cysteine cross-linking via disulfide bonds, and restraints from paramagnetic relaxation enhancement (PRE).

Since the PREs are plagued by intra/inter ambiguity [7] as well as subunit ambiguity, we focused on the effect of cysteine crosslinking restraints (which are only complicated by subunit ambiguity) to predict satisfying folds. The absence of a possible intra-subunit assignment makes the disulfide bond restraints much simpler to interpret, so our computational approach will initially focus solely on these restraints. Therefore, our goal will to be to find all possible topologically distinct folds that satisfy the disulfide bond restraints. The PRE restraints will be used later as a filter to eliminate the erroneous predictions.

Fold-operator theory finds alternative folds allowed by restraints

Since the restraint provided by the disulfide bonds is ambiguous and rather loose ($\mathrm{d}_{\mathrm{C}_\alpha}(i,j) \leq$ 10 Å), there are ways that the fold of the NMR and crystal structures for DAGK can be significantly changed without violating any disulfide bond restraints. For example, Figure 2 shows a sequence of changes that transform the crystal fold into the NMR fold, where the start fold, the end fold, and the intermediate fold all satisfy at least one assignment of each disulfide bond restraint.

The two changes described in Figure 2 can be decomposed into sequences of smaller changes called *operators*. These operators describe small changes to the folds that always result in a three-helical H2 bundle in the core of DAGK, and a maximal number of pairs of adjacent helices (i.e., the helical packing produced doesn't have holes in it), but don't necessarily produce only folds that satisfy the disulfide bond restraints. These operators are a mechanism to search the space of possible helical packings for DAGK to produce a set of folds which can be subsequently filtered against the disulfide bond restraints to return satisfying structures.

Only two operators, *roll* and *swap*, are needed to describe all the changes that can be made to the folds (Figure 3), and the application of all possible sequences of these operators to the original NMR fold results in 48 unique possible folds for DAGK (Figure 4). The fold changes shown in Figure 2 are examples of these operators applied to folds. The first change is the roll operator applied twice. The second second change is the swap operator applied once. Therefore, to transform the fold of the crystal structure into the NMR fold, one needs to apply the operator sequence `RRS` to the crystal fold where `R` is the roll operator, and `S` is the swap operator. These operators can be applied in any order and the result is the same. Consequently, `R` and `S` form the basis of a finite Abelian group of order 36. The mathematical structure of this group is discussed in the Supplementary Information (SI), Section 1.

Fold-based assignment of disulfide bonds

We used the folds predicted by our fold-operator theory to determine subunit assignments for the disulfide bonds. Since the upper distance of the C^α disulfide bond restraints is 10 Å, a disulfide bond assignment was considered satisfied by the fold if its two restrained helices were adjacent in the fold schematic. Using this simple criterion, we eliminated disulfide bond assignments that were inconsistent with the topology of each fold by eliminating assignments where the restrained helices were not adjacent in that fold. Each disulfide bond restraint has two possible assignments each due to the subunit ambiguity. Since the elimination step can potentially eliminate zero, one, or both assignments for each restraint, folds having a restraint with no remaining assignments can be excluded from further consideration. Our fold-based assignment excluded all but 26 of the predicted folds for DAGK (blue region in Figure 4). 18 of these folds had unique (i.e., unambiguous) assignments and 8 of these folds had ambiguous assignments.

Fold-to-structure protocol

For each of the 26 folds predicted by the fold-operator theory to satisfy the disulfide bond restraints, we constructed a crude atomic-resolution model of DAGK so its structure matched the fold. Each crude model was constructed using the following protocol.

1. Using PyMOL [16], we created a reduced model of the DAGK subunit by deleting all but residues 6–12, 32–44, 50, 57–77, 85, and 94–117 from the PDB structure 2KDC, model 1. These residues are, respectively, fragments of the SH helix, the H1 helix, the H1-H2 linker, the H2 helix, the H2-H3 linker, and the H3 helix.

2. For the chosen fold, we translated and rotated the fragments from step 1 so they aligned with one subunit of the fold. This step created a template structure for the subunit of DAGK. Since the SH helix was not modeled by the fold schematics, the SH helix fragment was oriented so it pointed away from the core of DAGK.

3. Using Xplor-NIH [17], we annealed an extended (i.e., unfolded) model of a single DAGK subunit using the intra-subunit NMR restraints: NOEs, hydrogen bonds, dihedral restraints, and RDCs. We configured the refinement to penalize differences between the backbones of the refined model and the template structure created in step 2. The result was a structure of the DAGK subunit that simultaneously matched the chosen fold and satisfied the NMR restraints.

4. Using PyMOL again, we made three copies of the subunit structure created in step 3. We rotated and translated the subunit structures until they matched the trimeric conformation of the chosen fold. The result here was a trimeric "seed" structure for DAGK to be used in later refinements.

We used the crude structures constructed using this protocol as seed structures for further refinement in Xplor-NIH. The refinement included all the experimental restraints: NOEs, hydrogen bonds, dihedral restraints, RDCs, disulfide bonds, and PREs. The disulfide bond restraints assignments used for the simulation were determined according to the fold-based assignment protocol above which resulted in either ambiguous or unambiguous assignments for each fold. For the PRE restraints, we used the deposited ambiguous assignments in the simulation. Unlike the subunit refinement, this trimeric refinement did not use a template structure to restrain the backbone of the refined structure. Without a backbone template, the trimeric refinement was free to change the fold of the structure when such a change resulted in a lower energy. Further details of the Xplor refinements are described in the SI, Section 2. The refinements were repeated 64 times for each of the 26 folds which resulted in 26 structural ensembles.

Results

Predicted folds refined to satisfying structures

The 26 satisfying folds predicted by our fold-operator theory were based on the disulfide bond restraints and the published NMR structure for DAGK [14]. The subsequent refinements in Xplor-NIH used all the deposited restraints, including NOEs, PRE restraints, disulfide bond restraints, dihedral restraints, hydrogen bonds, and RDCs. We analyzed the resulting 26 ensembles for satisfaction of the restraints (Figure 5). To simplify comparisons between the 26 different ensembles, we only report statistics on the lowest energy structure from each ensemble.

Structures were evaluated using two measures. The first entails the *Xplor total energy* as the value of the energy function returned by Xplor-NIH after refinement of individual structures, including the published NMR structure. Since all structures were refined using the same script, Xplor total energies are comparable across different structures.

The second scoring measure, the *RMS violation index*, is an RMS function of individual violation indices. Each violation index quantifies the satisfaction of a structure with respect to a class of restraints: NOEs, hydrogen bond restraints, RDCs, dihedral angle restraints, disulfide bond restraints, and PREs. Each violation index V reports the magnitude of the worst violation among the restraints in the class:

$$V = \frac{1}{N} \max_r \min_a v(r, a) \qquad (1)$$

where $v(r, a)$ is the violation of assignment a for restraint r, and N is a normalization constant. N transforms the violation onto a scale where zero indicates perfect satisfaction of the restraints and one indicates the worst violation is within acceptable limits. The normalization constants chosen for the violation indices in this study were: 0.5 Å for NOEs, 0.5 Å for hydrogen bonds, 1.0 Hz for RDCs, 5° for dihedral angle restraints, 2.0 Å for disulfide bond restraints, and 2.0 Å for PREs. Therefore, an NOE violation index of one or less indicates the worst NOE violation is 0.5 Å or less. The normalization constants can thus be chosen intuitively and allow violation indices for different restraint classes to be combined via the RMS function into a single statistic that reports the overall restraint satisfaction for a structure. The main benefit of this measure is it provides a natural cutoff at 1 that is based on commonly acceptable violation magnitudes. Figure 6 shows the RMS violation index and the Xplor total energy for the 26 determined structures. Figure S3 shows the Xplor total energies and RMS violation indices for the structures organized by distance to the crystal structure.

One might have expected the PRE restraints to act as a filter to remove unfavorable folds. However the PRE restraints did very little to discriminate between the predicted folds. Counterintuitively, RDCs did most of the discrimination between folds since the violation indices for the RDCs were in many cases above 1. Since RDCs are not directly sensitive to differences in translation, the sensitivity of the RDCs to different folds must be due to changes in helix shape caused by the stresses of other restraints during the Xplor simulation. In the cases where all restraints were not simultaneously satisfiable, the RDCs were the first restraints to accumulate violations due to their sensitivity. Even though our violation index results (Figure 6) show that RDCs are actually responsible for the bulk of discrimination between folds, we believe this is an indirect effect that is likely an artifact of the Xplor simulation and the chosen potential weights.

In some cases, structures designed from one fold changed to another fold during refinement since we configured Xplor-NIH to perform full simulated annealing instead of just local energy minimization. There are eight such switches in total, which are shown with brown arrows in Figure 5. When viewed as a dynamical system, the network of fold switches has two prominent attractors. One is at fold O (the NMR fold) and the other is at fold B, which is not related to any published structure. See the blue letters in Figure 5 to find the names of the folds. Six out of the top seven structures by Xplor total energy and six out of the top seven structures by RMS violation index were either seeded from, or switched to, one of these two attractor folds.

Interestingly, the structure seeded with fold P converted to fold O (the NMR fold) during refinement, but its RMS violation index and Xplor total energy scores were better than the structure originally seeded with fold O. One might have suspected that the seed structure closest to fold O would have performed the best, but these results counter that intuition. The only difference between the two refinements was the starting fold and whatever random moves were used during the simulation. Since both structures ended in fold O, the fact that the structure seeded with fold P performed better than the structure seeded with fold O, shows that simulated annealing can indeed become trapped in local minima even when the starting structure is relatively close to a better minimum.

Of the 26 folds predicted by the fold-operator theory for DAGK to be satisfying, 9 of these folds yielded at least one structure that met the expectations (on average) for restraint satisfaction by having an RMS violation index of 1 or lower. 8 of the 26 folds yielded structures that switched to different folds during refinement, so it is not known from these results if these 8 folds describe satisfying structures or not. 9 folds resulted in structures with RMS violation indices of greater than 1, and hence these structures did not meet expectations for restraint satisfaction. Figure S1 shows all the structures grouped by their post-refinement fold.

Interestingly, the best structure with the crystal fold (fold M) scored similarly to the structures with the NMR fold (P, O, N). We found no systematic difference in the restraint satisfaction statistics between these four folds. A full listing of the violation indices for each structure is given in Table S3, and Table S4 shows additional restraint satisfaction statistics for each structure.

Figure S2 shows the differences between the best structure with fold M and the published crystal structure. The transmembrane helices of the structure with fold M appear bent in comparison to the helices of the crystal structure. Indeed, all the transmembrane helices refined from the NMR restraints (including those in both the published NMR structure and our refined structures) show these distortions. Since the NMR and crystal structures were solved in different environments using different detergents, one might expect such differences between the NMR and crystal structures. In the case of the NMR structures, the size of the detergent micelles may influence the helix shape. Such a detergent-caused deformation of protein conformation has been previously observed for NMR structures in detergent micelles and nanodiscs [18].

Post-refinement disulfide bond assignments

We also looked at which disulfide bond assignments were best satisfied by the structures computed by Xplor. For just the eight disulfide bond restraints between C^{α} atoms in the H1–H3 and H2–H3 helix pairs, we categorized the combination of assignments as *synchronized*, *unsynchronized*, or *ambiguous* (see Figure 7). Folds A, B, E–I, L–P, and S–Z had synchronized assignment combinations, folds J, K, Q had unsynchronized assignment combinations, and folds C, D, and R had ambiguous assignment combinations after refinement. Since DAGK is composed of mainly parallel helices, it was expected that most of the assignment combinations would be synchronized. Indeed, the best nine structures by both Xplor energy and RMS violation index had either synchronized assignment combinations or ambiguous assignment combinations, which are supersets of synchronized assignment combinations.

Surprisingly, the best fold by Xplor total energy was neither the fold of the NMR structure nor the fold of the crystal structure. Fold B has the lowest Xplor total energy, and the second lowest RMS violation index. It is topologically distinct from both the NMR and the crystal folds and its three refined structures differ by 12.31–12.87 Å transmembrane helical backbone RMSD from the published NMR structure and by 12.77–12.83 Å from the published crystal structure. It also satisfies different subunit assignments of the disulfide bond restraints than either published structure, which shows fold-operator theory was able to find previously unknown solutions to the restraint satisfaction problem for DAGK.

From our results, we cannot claim that this new putative fold B has biological significance for DAGK. We conjecture that if and when more experimental restraints can be measured for the membrane-bound solution structure, fold B will be ruled out. However, it must be emphasized that currently, based on all NMR measurements to date, (1) fold B is vastly different from the published structures, (2) it cannot be excluded as a possible structure, and, moreover, (3) it fits the NMR restraints as well or better than the two published folds.

Discussion

In many respects, the 2D schematic representation used in the fold-operator theory for DAGK is an oversimplification. Condensing the full three-dimensional structure of DAGK into a flat projection ignores some important structural details of DAGK. For instance, the transmembrane helices need not be strictly parallel, or even straight. Modeling changes to helix shape with operators could potentially enable the discovery of more satisfying folds, but simulated annealing methods likely already adequately search over such changes in helix shape. Since simulated annealing is prone to becoming stuck in local minima (like all local minimization methods) and therefore might miss genuine solutions, the goal is to choose operators that complement simulated annealing and overcome its local minimum limitations rather than use operators to model small changes to helix shape. Indeed, despite the simple representation of structure used by the fold schematics, the fold-operator theory predicted 24 distinct folds for DAGK that satisfied the disulfide bond restraints (in addition to the two published folds), of which 9 folds yielded structures that met stringent expectations for NMR restraint satisfaction.

One drawback to the fold-to-structure protocol presented here is that unrestrained degrees of freedom are not necessarily sampled by the final ensemble. For instance, the SH helix in our ensembles appeared more converged than was suggested by the NMR restraints and as a result, the ensembles for the SH helix were falsely precise. Normally, unrestrained degrees of freedom are searched by the random structure generation used in the beginning of most annealing protocols. For small modes of variability, the random structural sampling is able to report a variety of structures, but has difficulty searching topologically distinct folds. The fold-operator theory presented here completely replaces random structural sampling as a mechanism to search alternate folds, so one must take care to ensure that all degrees of freedom are captured by the operators. In our case, variability in the SH helix had little impact on the fold of DAGK, so we chose not to model it using the operators.

After noticing that low-energy structures for DAGK correspond only to synchronized disulfide bond assignment combinations, it may be tempting to dispense with the procedure of predicting folds, and instead exhaustively search all the synchronized assignment combinations. In general, there are vastly fewer synchronized assignment combinations than unsynchronized ones. In particular, DAGK has four synchronized assignment combinations and 124 unsynchronized ones. Each synchronized assignment combination could be fed into Xplor and the simulated annealing computation itself could search for the satisfying folds without having to deal with assignment ambiguity. This might work well for DAGK specifically, since its best folds happened to correspond to synchronized assignment combinations, but the procedure does not generalize to all proteins. If the native fold of the protein only satisfies unsynchronized assignment combinations, searching only the synchronized assignment combinations will never find the native fold. While this is probably not the case for DAGK since its helices are likely all parallel, another protein could have rigid fragments that lie in the plane induced by the symmetry (see Figure 8 for an example). In this case, a single fragment might make contacts to multiple instances of the same fragment from different subunits. Searching only the synchronized assignments would fail to find such a conformation.

The fold-operator theory presented here bears some similarity to methods in protein structure prediction. The *ideal forms* proposed by Taylor et al. [19] describe different protein folds using the “combinatorial approach” [20]. Under this regime, possible folds are enumerated from a space of choices governing the placement of α-helices and β-sheets and then structures are fit to these ideal forms, refined, and finally scored. While our fold-operator theory shares the combinatorial generate-and-test approach, where the methods differ is how the combinatorial space is defined. The ideal forms were curated from a database of structural information, while in the fold-operator theory, the different folds are algebraically defined by the initial satisfying fold and the group action of operators.

We have demonstrated our method on DAGK, showing how to find a remarkable variety of satisfying folds, but the method can also be applied to other homo-oligomeric proteins where ambiguous restraints necessarily hinder structure determination with simulated annealing. The only requirement is that a single atomic-resolution structure that satisfies the restraints be determined. Then that structure is analyzed via our fold-operator theory to search for alternate folds that might also satisfy the restraints. The application of the fold-operator theory to a new protein requires defining F, a set of folds, and G, a group of operators, analogously to our example with DAGK. This defines a group action on the configuration space of folds (see SI). The first step is to discover one fold $f \in F$ that satisfies the restraints, and (similarly to our example in Figure 2) search the changes to the structure that preserve restraint satisfaction. If relatively rigid backbone fragments can be determined (e.g., helices within each subunit), then restraints can be categorized as restraining pairs of rigid fragments and the number total number of assignment possibilities is vastly reduced. Therefore, changes to f that preserve inter-subunit restraint satisfaction for symmetric homo-oligomers will generally include substituting fragments in one subunit with identical fragments from other subunits.

The next step is to factor the satisfaction-preserving changes into a set of finer operators (e.g., Figure 3) that form the basis of an Abelian group G. The group structure is necessary to precisely model the symmetry inherent in many homo-oligomeric proteins, but the operators need not preserve restraint satisfaction. Removing this restriction was necessary to obtain the group structure in the case of DAGK, and, more generally, it allows the operators to hop between "islands" of satisfying folds. G and f are then used to construct F via the group action and therefore describe the possible folds. For DAGK, F was small and exhaustive search was a feasible method to find the low-energy folds. If F is large (which appears to require a larger protein than the $121 \times 3 = 363$ residue DAGK), more sophisticated algorithms may be needed, such as branch-and-bound pruning used in protein design [8].

Systematic approaches to NMR structure determination such as DISCO [7] and Fold-Operator Theory (this paper) constitute powerful techniques, and indeed we recently used DISCO to determine the solution structure of the membrane bound MPER trimer of HIV-1 gp41 [21]. Since these two algorithms address different aspects of the problem of structure determination for symmetric homo-oligomers, it is conceivable that in the future they could be combined to reap the benefits of both strategies.

Conclusion

We have presented a general method for structure determination of protein homo-oligomers and demonstrated the method on DAGK. We conclude that the differences in the published NMR and crystal structures are due to limitations of current NMR structure determination methodology. We overcame these limitations by using a new fold-operator theory to explicitly search the space of folds and predict distinct fold topologies for further investigation. These folds were used to reduce (and in some cases eliminate) ambiguity in restraint assignments which lessened the difficulty of subsequent refinement of seed structures in Xplor-NIH. By explicitly performing a search over topologically distinct folds, we avoided the implicit fold search performed by local minimization methods which can become trapped in local minima and therefore fail to report satisfying solutions. Using explicit fold-space search methods to address the limitations of local minimization techniques such as simulated annealing enables robust structure determination for difficult homo-oligomeric systems, particularly membrane associated systems hindered by the availability of only sparse and ambiguous restraints.

References

[1] Schwieters, C. D, Kuszewski, J. J, & Marius Clore, G. (2006) Using Xplor-NIH for NMR molecular structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy* **48**, 47–62. doi: DOI: 10.1016/j.pnmrs.2005.10.001.

[2] Herrmann, T, Güntert, P, & Wüthrich, 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**, 209–227.

[3] Rosato, A, Tejero, R, & Montelione, G. T. (2013) Quality assessment of protein nmr structures. *Current opinion in structural biology* **23**, 715–724.

[4] Nilges, M, Malliavin, T, & Bardiaux, B. (2010) Protein structure calculation using ambiguous restraints. *eMagRes*.

[5] Potluri, S, Yan, A. K, Chou, J. J, Donald, B. R, & Bailey-Kellogg, C. (2006) Structure determination of symmetric homo-oligomers by a complete search of symmetry configuration space, using NMR restraints and van der Waals packing. *Proteins: Structure, Function, and Bioinformatics* **65**, 203–219.

[6] Potluri, S, Yan, A. K, Donald, B. R, & Bailey-Kellogg, C. (2007) A complete algorithm to resolve ambiguity for intersubunit NOE assignment in structure determination of symmetric homo-oligomers. *Protein Science* **16**, 69–81.

[7] Martin, J. W, Yan, A. K, Bailey-Kellogg, C, Zhou, P, & Donald, B. R. (2011) A graphical method for analyzing distance restraints using residual dipolar couplings for structure determination of symmetric protein homo-oligomers. *Protein Science* **20**, 970–985.

[8] Donald, B. R. (2011) *Algorithms in Structural Molecular Biology.* (MIT Press, Cambridge, MA).

[9] Ikura, M & Bax, A. (1992) Isotope-filtered 2D NMR of a protein-peptide complex: study of a skeletal muscle myosin light chain kinase fragment bound to calmodulin. *Journal of the American Chemical Society* **114**, 2433–2440.

[10] Vinogradova, O, Sönnichsen, F, & Sanders, C. R. (1998) On choosing a detergent for solution NMR studies of membrane proteins. *J. Biomol. NMR* **11**, 381–6.

[11] Gautier, A. (2013) Structure determination of α-helical membrane proteins by solution-state NMR: Emphasis on retinal proteins. *Biochim. Biophys. Acta.*

[12] Bellot, G, McClintock, M. A, Chou, J. J, & Shih, W. M. (2013) DNA nanotubes for NMR structure determination of membrane proteins. *Nat Protoc* **8**, 755–70.

[13] Arora, A. (2013) Solution NMR spectroscopy for the determination of structures of membrane proteins in a lipid environment. *Methods Mol. Biol.* **974**, 389–413.

[14] Van Horn, W. D, Kim, H.-J, Ellis, C. D, Hadziselimovic, A, Sulistijo, E. S, Karra, M. D, Tian, C, Sönnichsen, F. D, & Sanders, C. R. (2009) Solution nuclear magnetic resonance structure of membrane-integral diacylglycerol kinase. *Science* **324**, 1726–9.

[15] Li, D, Lyons, J. A, Pye, V. E, Vogeley, L, Aragão, D, Kenyon, C. P, Shah, S. T. A, Doherty, C, Aherne, M, & Caffrey, M. (2013) Crystal structure of the integral membrane diacylglycerol kinase. *Nature* **497**, 521–4.

[16] Schrödinger, L.L.C. (2012) The PyMOL Molecular Graphics System. Version 1.5.0.1.

[17] Schwieters, C. D, Kuszewski, J. J, Tjandra, N, & Clore, G. M. (2003) The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance* **160**, 65–73.

[18] Hagn, F, Etzkorn, M, Raschle, T, & Wagner, G. (2013) Optimized phospholipid bilayer nanodiscs facilitate high-resolution structure determination of membrane proteins. *J. Am. Chem. Soc.* **135**, 1919–1925.

[19] Taylor, W. R, Bartlett, G. J, Chelliah, V, Klose, D, Lin, K, Sheldon, T, & Jonassen, I. (2008) Prediction of protein structure from ideal forms. *Proteins* **70**, 1610–9.

[20] Cohen, F. E, Sternberg, M. J, & Taylor, W. R. (1980) Analysis and prediction of protein beta-sheet structures by a combinatorial approach. *Nature* **285**, 378–82.

[21] Reardon, P. N, Sage, H, Dennison, S. M, Martin, J. W, Donald, B. R, Alam, S. M, Haynes, B. F, & Spicer, L. D. (2014) Structure of an HIV-1 neutralizing antibody target, the lipid-bound gp41 envelope membrane proximal region trimer. *Proceedings of the National Academy of Sciences* **111**, 1391–1396.

[22] Mizuno, Y, Berenger, B, Moorhead, G. B, & Ng, K. K. (2007) Crystal structure of Arabidopsis PII reveals novel structural elements unique to plants. *Biochemistry* **46**, 1477–1483.

Acknowledgements

This work was supported by the following grants from the National Institutes of Health: R01 GM-65982 and R01 GM-78031 to B.R.D. and R01 AI-055588 to P.Z. We thank Swati Jain and Kyle Roberts for help with MolProbity. We also thank the anonymous reviewers for helpful comments that improved our manuscript.

Figure legends

Figure 1: Fold schematics clearly show helical packing for the NMR (top) and crystal (bottom) structures of DAGK. In the fold schematic, the helices are shown as colored discs (the amphiphilic surface helix SH is not shown), the loop regions are shown as black lines, and the position of the three-fold symmetry axis is shown as a small black circle. Individual subunits are distinguished with different shading. Right: schematic of the subunit structure shows the helix naming and color schemes.

Figure 2: The crystal structure can be transformed into the NMR structure by repositioning the transmembrane helices. The changes are indicated by arrows. Left: In the fold of the crystal structure, one set of disulfide bond assignments are satisfied. Center: Moving the H1 (red) and H3 (blue) helices as shown transforms the crystal fold into an intermediate fold that satisfies a different set of assignments. Right: Swapping the H1 and H3 helices transforms the intermediate fold to satisfy yet another set of assignments.

Figure 3: The two operators in the fold-operator theory for DAGK: The Roll operator moves the red and blue helices (H3 and H1 respectively) along the perimeter of the three-helix core (H2) in a counterclockwise direction. The Swap operator exchanges the position of the red helix (H3) with the blue helix (H1) that lies immediately counterclockwise adjacent to it. After six applications of either of the two operators, the ending fold is always the same as the starting fold.

Figure 4: The fold graph of 48 distinct folds predicted for DAGK by the fold-operator theory. Graph vertices are represented by fold schematics. The edges are represented in the lower right panel. Generally, the roll operator sends any fold horizontally to its right neighbor. The swap operator sends any fold diagonally to its lower-right neighbor. Since the fold graph is embedded on the 2-torus, the operators “wrap around” the sides of the figure. Of these folds, 26 were predicted to satisfy the disulfide bonds (blue region), and 22 were not. Each satisfying fold was given a single-letter name, shown in blue. The operator sequence `RRS` that transforms the crystal fold into the NMR fold (also described in Figure 2)

is shown with three grey arrows.

Figure 5: The 26 satisfying structures computed for DAGK. Each structure is shown using the schematic of the fold that was used to seed the refinement. Structures that changed folds during the refinement are shown with brown arrows between the fold schematics. [1]The RMS violation index scores satisfaction of all solution restraints without regard to force field energies. This score is described in the text.

Figure 6: Top Left: For DAGK, the Xplor total energy function does not have a single low-energy well. Even though each structure was refined from a single initial fold, a single fold can describe more than one structure when structures change folds during refinement. For example, two structures changed from their original folds to the NMR fold during refinement, giving the NMR fold three (albeit similar) structures. Bottom Left: The same is true of the RMS violation index, indicating the restraints do not define a unique structure. Structures with a RMS violation index of 1 (purple line) or lower indicate these structures met expectations (on average) for restraint satisfaction. Top Right: Structures with low Xplor total energies also have low RMS violation indices. Bottom Right: Violation indices for each restraint type. To simplify the bottom right plot, structures are filtered so that among structures sharing the same final fold, only the structure with the lowest RMS violation index is shown. All structural distances (x-axis) are backbone atom (N,C$^{\alpha}$,C$'$) RMSD values in Å computed for the helical residues 30-48, 51-83, and 90-119 only. Variations in the loop regions were not considered in this score.

Figure 7: A single subunit-ambiguous distance restraint between H2 (yellow) and H3 (red) has two possible assignments (blue lines, left). A set of restraints between H2 and H3 are *synchronized* when the assignments satisfied by a structure restrain only one pair of helices. If the assignments are unambiguous and restrain multiple pairs of helices, the combination is *unsynchronized*. Otherwise, the assignment combination is *ambiguous*.

Figure 8: The crystal structure of *Arabidopsis thaliana* PII [22] (PDB ID: 2O66) shows a group of three β-strands (residues 40–46) positioned end-to-end, shown here in red, green,

and blue. Hypothetical distance restraints (yellow) from the strand in subunit A could restrain the β-strand to its symmetric partners in subunits B and C. The fold of PII requires that the subunit assignments for these restraints be unsynchronized. Synchronized assignments would not be compatible with this fold.

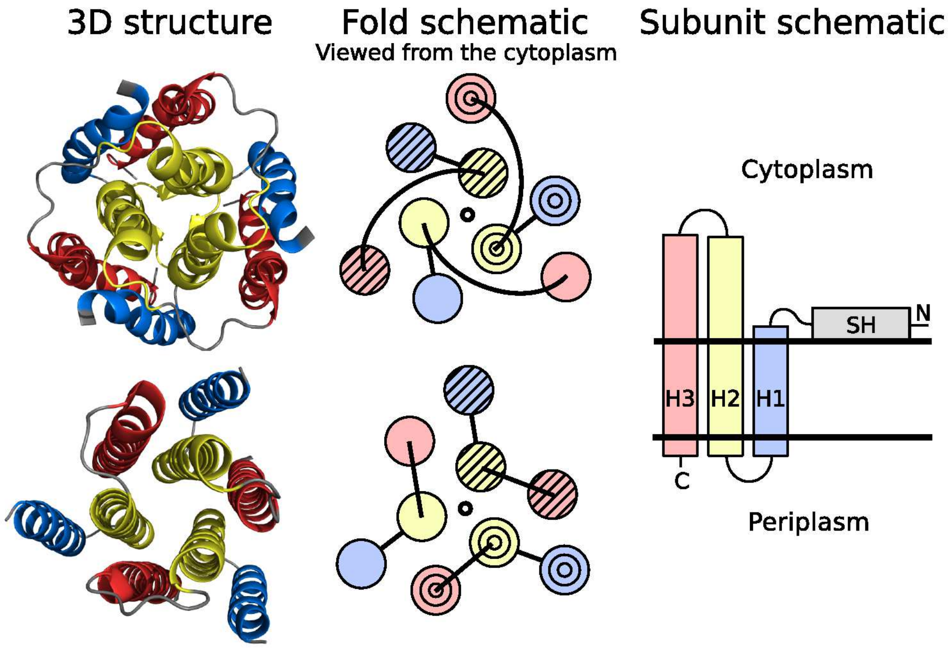

Figure 1:

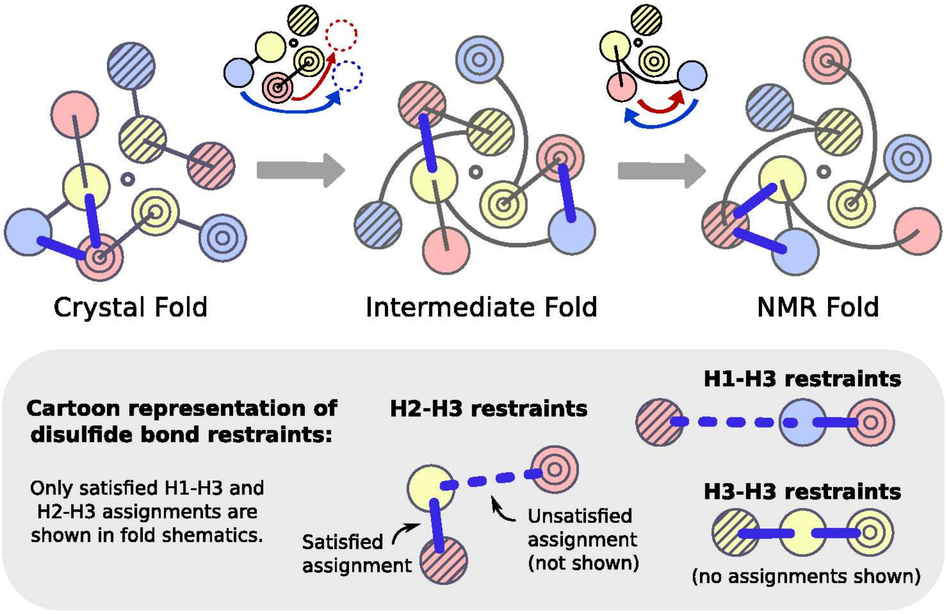
Crystal Fold
Intermediate Fold
NMR Fold
Cartoon representation of disulfide bond restraints:
Only satisfied H1-H3 and H2-H3 assignments are shown in fold shematics.
H2-H3 restraints
Satisfied assignment
Unsatisfied assignment (not shown)
H1-H3 restraints
H3-H3 restraints
(no assignments shown)

Figure 2:

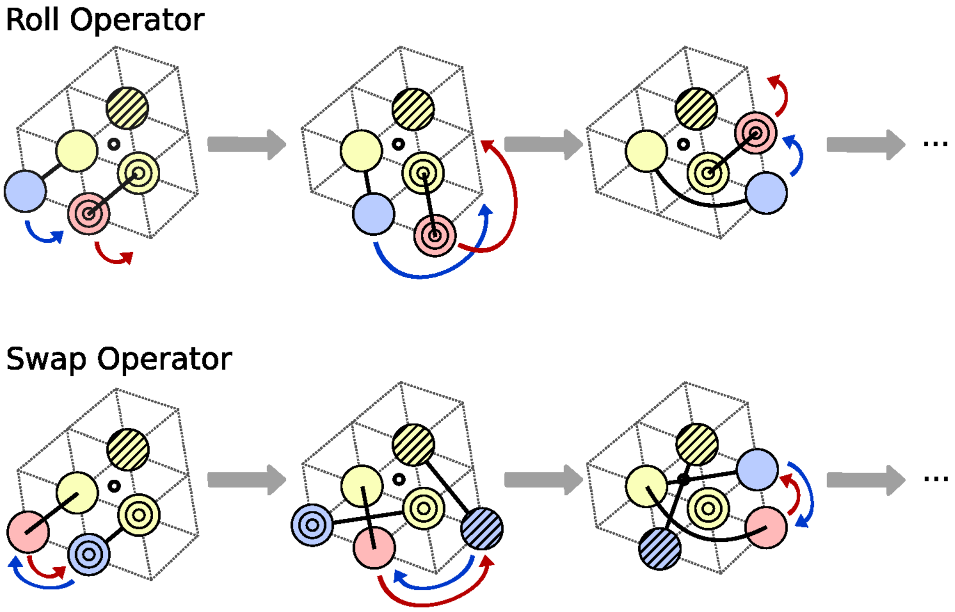

Figure 3:

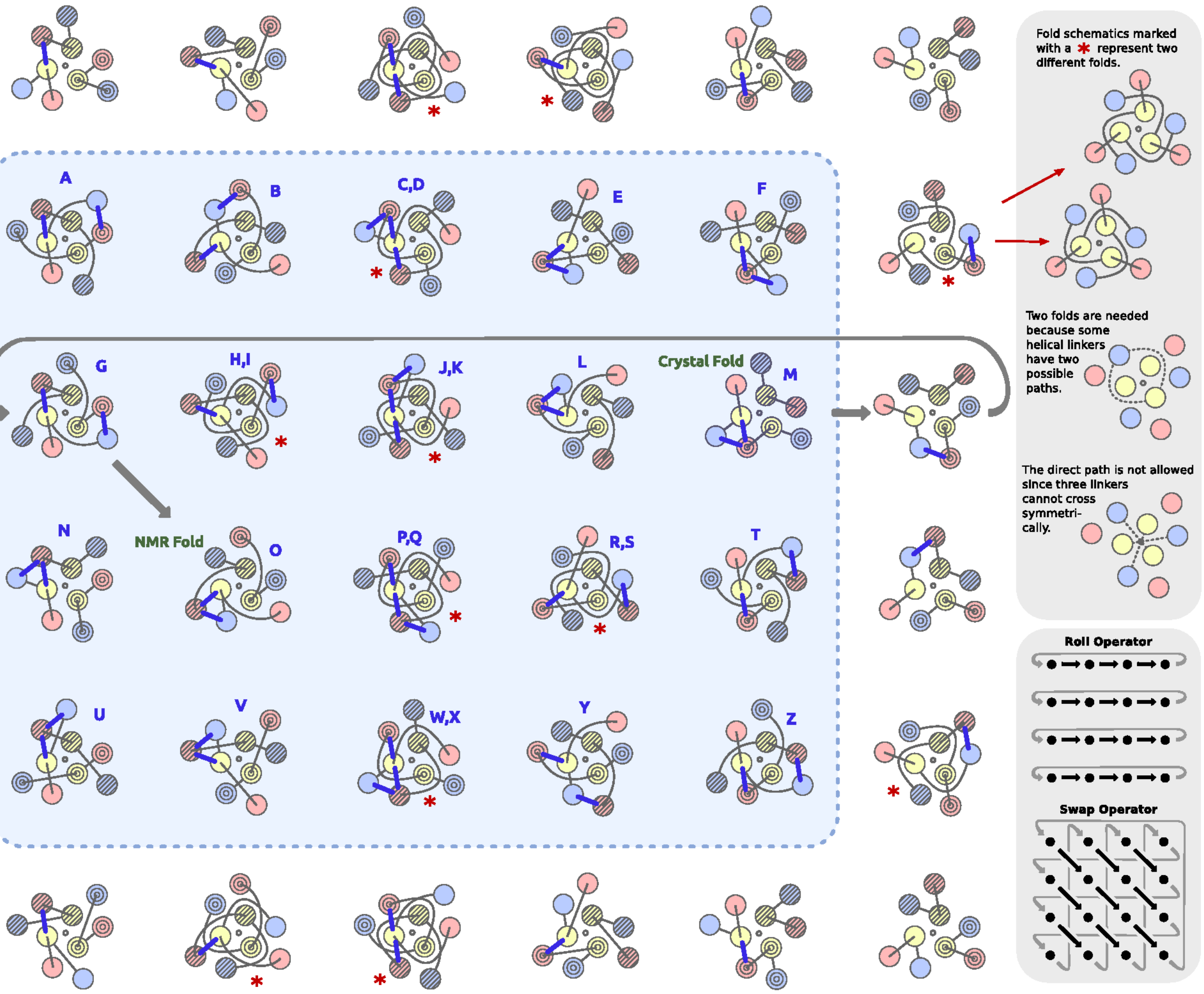
A
B
C,D
E
F
G
H,I
J,K
L
Crystal Fold
M
N
NMR Fold
O
P,Q
R,S
T
U
V
W,X
Y
Z
Fold schematics marked with a * represent two different folds.
Two folds are needed because some helical linkers have two possible paths.
The direct path is not allowed since three linkers cannot cross symmetri-cally.
Roll Operator
Swap Operator

Figure 4:

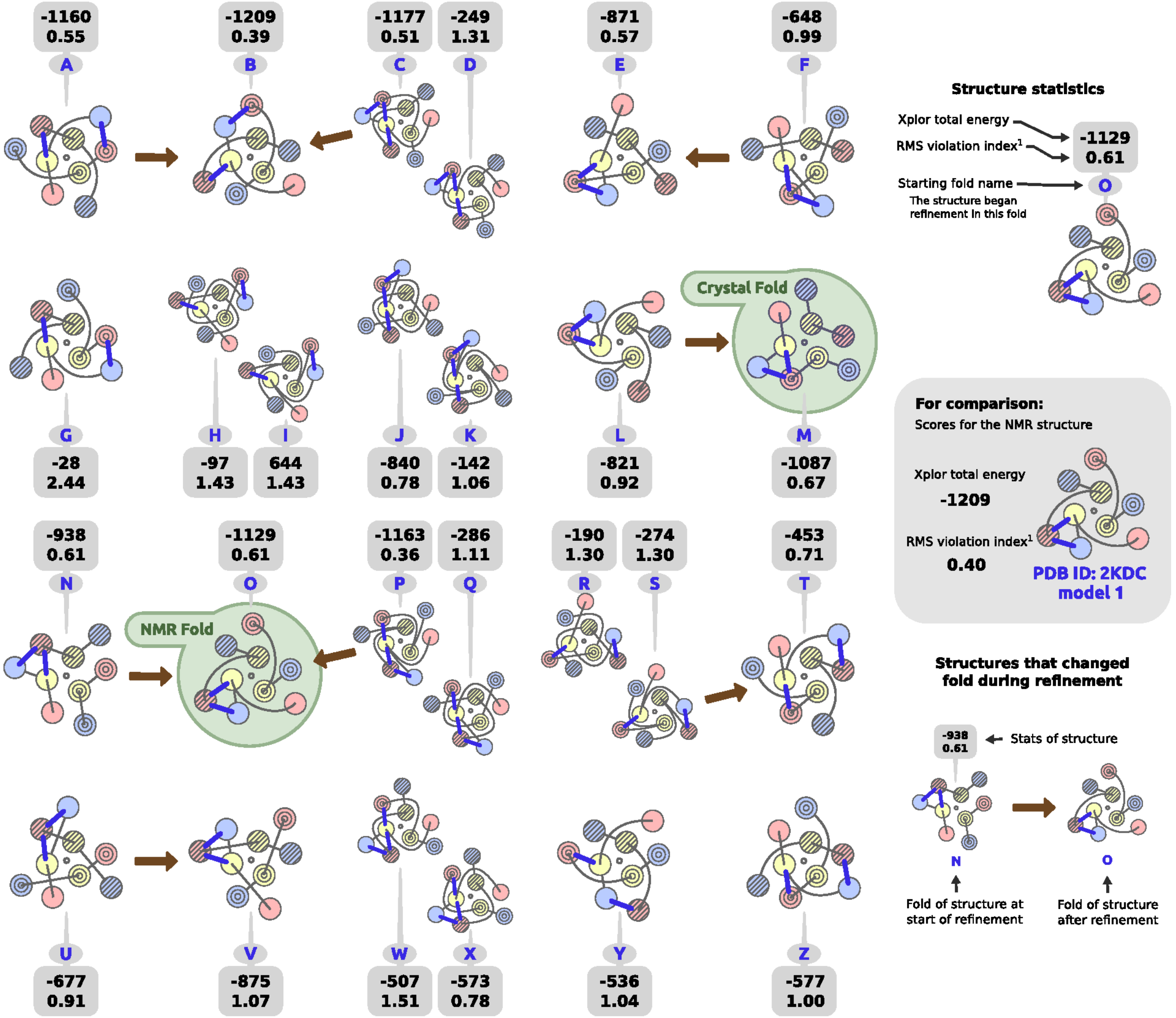
-1160
0.55
A
-1209
0.39
B
-1177
0.51
C
-249
1.31
D
-871
0.57
E
-648
0.99
F
Structure statistics
Xplor total energy
RMS violation index[1]
Starting fold name
The structure began refinement in this fold
-1129
0.61
O
G
-28
2.44
H
-97
1.43
I
644
1.43
J
-840
0.78
K
-142
1.06
L
-821
0.92
Crystal Fold
M
-1087
0.67
For comparison:
Scores for the NMR structure
Xplor total energy
-1209
RMS violation index[1]
0.40
PDB ID: 2KDC model 1
-938
0.61
N
-1129
0.61
O
NMR Fold
-1163
0.36
P
-286
1.11
Q
-190
1.30
R
-274
1.30
S
-453
0.71
T
Structures that changed fold during refinement
-938
0.61
Stats of structure
N
O
Fold of structure at start of refinement
Fold of structure after refinement
U
-677
0.91
V
-875
1.07
W
-507
1.51
X
-573
0.78
Y
-536
1.04
Z
-577
1.00

Figure 5:

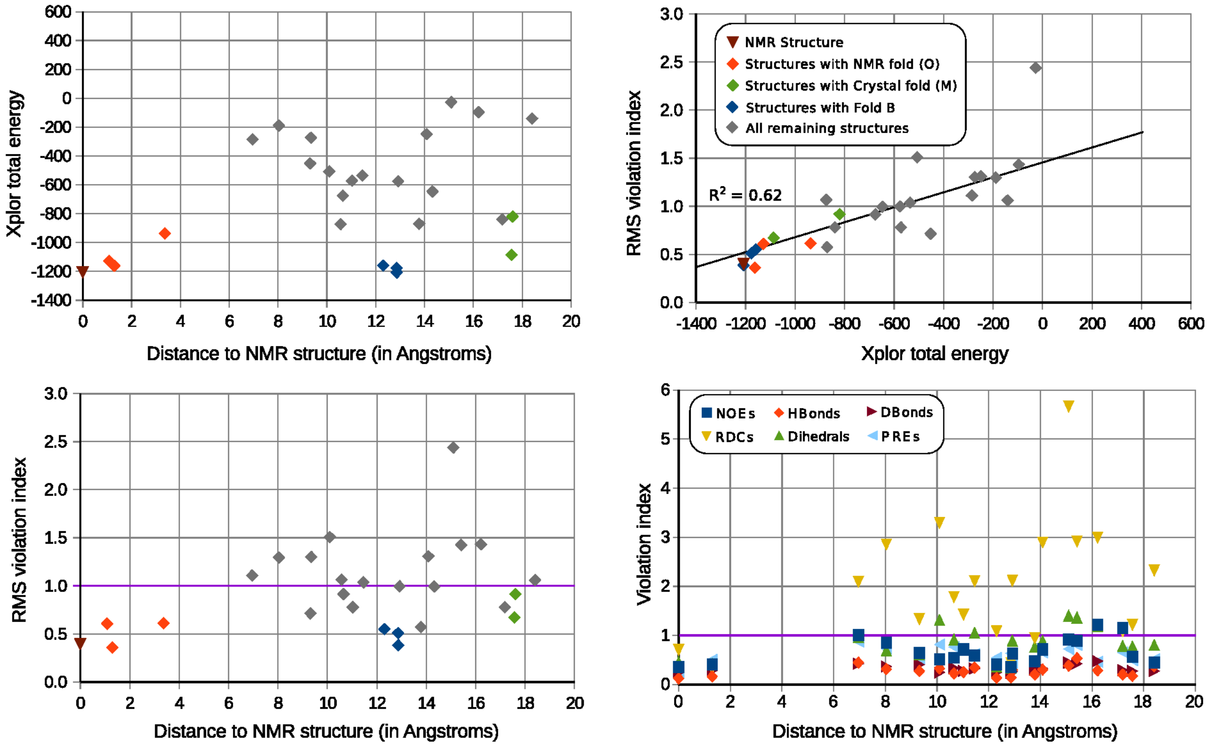

Figure 6:

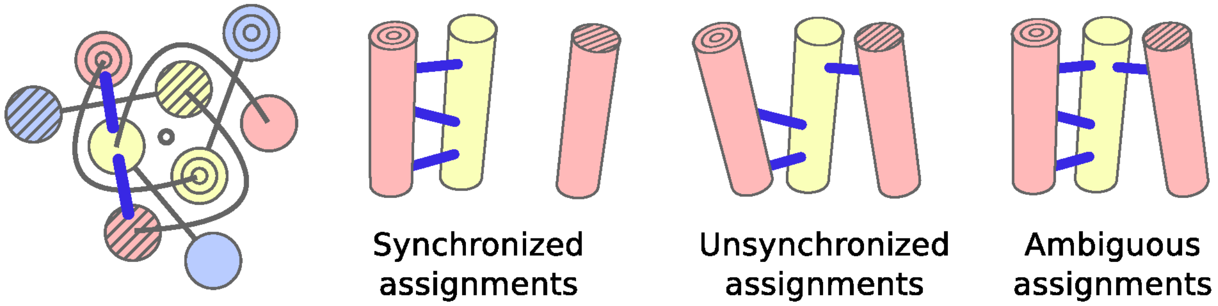

Figure 7:

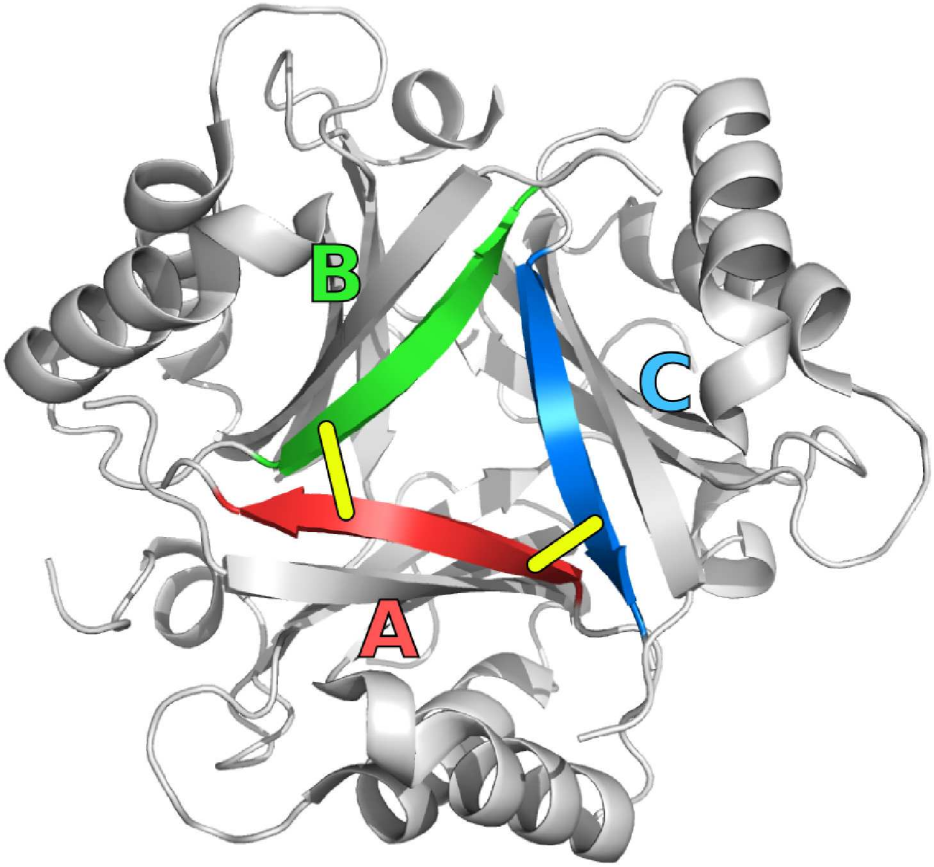

Figure 8: