Center for Biotechnology and Interdisciplinary Studies
Rensselaer Polytechnic Institute

Abstract

In structurally modeling proteins many models in an ensemble are typically generated which represent in a sense both the internal and bulk dynamics. Certain regions of proteins are physically flexible and some regions are not able to be consistently modeled in an ensemble due to dynamics. Software is available to determine, with assumptions, the regions of the structure which are 'well-defined' in the ensemble and which are not. These algorithms include Expanded FindCore, Cyrange, FindCore, and the use of dihedral angle order parameters (DAOP). These are reviewed. A new Residual Dipolar Coupling (RDC) database of 90 proteins in one to three media has been developed and is used to further quantify the relevance of using well-defined regions in protein validation and refinement using RDC's.

Section 1: Introduction

Section 2: Well-defined definition and algorithms

NMR structures are typically reported as an ensemble of stable energy minimized models [1]. However, these static models in the ensemble vary between each other. This geometrical variation is representative of the conformational states and dynamics of the protein in solution. It is of interest to find a representative structure from the ensemble, and the variation of the models can be problematic. For example, dihedral angles may vary in the ensemble. This variation is simplified by using a notion of 'well-defined' regions of the protein across the ensemble in residues, atoms, and their coordinates. If there is little variation of the models in a region of residues, then that set of residues is well-defined. There are several methods to quantify this geometrical variation in the context of an ensemble of models. These methods are described in this section.

From the point of view of a molecular dynamics trajectory the ensemble in the PDB structure represents snapshots of a molecule in motion. The ensemble is representative of the dynamics and potentially contains information about the internal motion and also conformational states. The population of these conformations is difficult to obtain even in MD simulations, however; experimental input and dissected trajectory fragments can to be used to obtain information of this population.

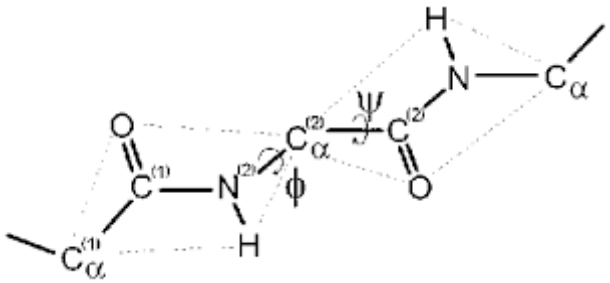
Well-defined regions as found by an ensemble indicate confidence in the molecular modeling due to the repetition of the geometrical coordinate structure in different models. This repetition is used to find a convergence of the model ensemble. After well-defined atoms, residues, and regions of the protein structure are defined, a reasonable superposition of N models can be more accurately obtained and then a representative structure of the protein can be determined.

There are several commonly used algorithms to determine well-defined residues and atoms from protein PDB files. The simplest is the dihedral angle order parameter (DAOP), which is used to quantify the variation of backbone dihedral angles [2]. A threshold value sets a limit on what variation is acceptable for a well-defined linkage. More complicated is the atomic-distance variation of atoms in the set of models in the ensemble. This variation can be quantified using an inter-atomic distance variance matrix [3,4], also with a threshold defining what the acceptable variation is. In this method an atom is considered in the set of all other atoms, and as a whole set of atoms, the variation is examined across models. This method is formulated in the methods Expanded FindCore [5,6] and Cyrange [4] restriction. Both techniques, dihedral angle and variance of inter-atomic distances, are in softwares available to the public at several sites: https://montelionelab.chem.rpi.edu/software/pdbstat.htm (PDBStat:DAOP, Extended FindCore) and http://www.bpc.uni-frankfurt.de/cyrange.html (Cyrange).

Dihedral angle order parameter restriction:

Inter-residue dihedral 3-bond angles characterize the linkage of backbone amino acids [2]. There are two relevant angles that give the orientation about the $C\alpha$ angle of two linked amino acids, ϕ_i and ψ_i, illustrated in Figure 1.

Figure 1: Illustration of dihedral angles of a peptide bond.



These two angles vary in the ensemble models for all of the backbone residues (also related 3-bond torsion angles of the sidechains). Dihedral order parameter (DAOP) estimates are a typical means for finding well-defined residues. The order parameter for each of the angles is defined in Eqn. 1. $\phi_{i,j}$ is a dihedral angle j for a model i, with N models [7].

$$S(\phi_j) = \frac{1}{N}\sqrt{\left(\sum_{i=0}^{N} sin\,\phi_{i,j}\right)^2 + \left(\sum_{i=0}^{N} cos\,\phi_{i,j}\right)^2} \qquad \text{Eqn. (1)}$$

The dihedral order parameter is a measure of the angle variations from the different models. The amount of fluctuation among the models is quantified by adding this for the two dihedral angles. If $S(\phi) + S(\psi)$ is greater than some value, in Eqn. 2, then that backbone residue is considered well-defined.

$$S(\phi_j) + S(\psi_j) \geq 1.8 \qquad \text{Eqn. (2)}$$

Heuristically this cutoff value is chosen as 1.8, corresponding to individual cutoffs of .9 for both angles. This number was conventionally chosen to represent at most a 24% fluctuation in the orientation of the linked amino acids in the models [7]. RDCs can be used to analyze this heuristic quantitatively.

Although this DAOP quantity is a measure of the well-defined backbone regions, there are two faults with this method. First, it lacks as a measure of disjoint regions of the protein. For example, there could be two separated domains of the protein which are both found to be entirely well-defined by DAOP. The total region of both domains may not superimpose well from the ensemble which would give an inaccurate set of atoms in the determination of the well-defined set. ϕ, ψ dihedral order parameters will not give information of both separated regions being well-defined as a whole. In other words, the DAOP approach does not quantify long-range order in the protein. Second, the DAOPs are typically restricted to backbone residues, and not atoms in the sidechain as found by similarity in the model ensemble.

FindCore, Expanded FindCore, Cyrange - Variance matrix approach:

An extension of this set of well-defined (backbone) residues from DAOP can be obtained by using an inter-atomic distance variance matrix. FindCore2 and Cyrange use both of these classifications, DAOP and variance matrix, to find a stable set of residues in the protein structure. The dihedral angle order parameter cutoff is used to find an initial set of residues, but this is not necessary as is typical in many search problems.

An improved method for finding the set of atoms or residues in the protein which are considered well-defined is found by using all the models and quantifying the fluctuations of distances using an inter-atomic variance matrix. This uses all the desired atoms in the input: atom 1 is compared to all the others, atom two is compared to all the others including atom one, etc… These distances are collectively represented in a variance matrix. The variance method uses the coordinates of an atom in relation to the coordinates of all the other atoms in the PDB structure. Well-defined is found by examining the atom's distance variation with other atoms in the ensemble of models across models. Different models in the ensemble may have different distances of an atom with others, considered as a fluctuation from a dynamical perspective in the different models. The matrix aspect comes from treating this atom with all of the other atoms in the structure in a pairwise correlation sense. Given a variance cutoff for fluctuations, μ, similar to the DAOP threshold 1.8 parameter, the atoms can be partitioned into well-defined and not well-defined [5]. Once the well-defined atoms are found in this approach, residues can be found as well-defined if a certain type and number of atoms in it are selected as well-defined. This method is FindCore [5], implemented in PDBStat [6]. An improvement using a different iteration with a larger variation parameter is Expanded FindCore (FindCore2) [5], implemented in PDBStat, and Cyrange [8].

The limitation with FindCore, however, was improved upon in the Expanded FindCore algorithm. After several studies of ensemble model superposition, it was found that its 'core set' of well-defined atoms is too small [5]; the requirements for atoms to be in the well-defined set are too stringent. The 'core set' of atoms from FindCore does lead to a minimized RMSD in the superposition of the models in the structure's ensemble. However, in these studies it was found that atoms not in the 'core set' also superimposed well.

Expanded FindCore iteratively improves upon on the initial assessment of the well-defined atoms by including more atoms within the allowed range of the mean square fluctuation of an atom's distance to the mean structure [7,8,9]. This is defined in Eqn. 3,

$$\mu^2 = ([x, y, z] - \langle [x, y, z] \rangle)^2 \qquad \text{Eqn. (3)}$$

which are the atomic coordinates [x,y,z] of an atom to its median analog, with an average over all atoms. This is also used in the FindCore algorithm. The $\langle \mu^2 \rangle$ distribution of this quantity was found not to be normal in FindCore [5], but log-normal, i.e. $log(\langle \mu^2 \rangle)$is normal. This leads to a consideration of increasing the value of $\langle \mu^2 \rangle$ as a cutoff to be used for the atom to be included in the 'core set' because the log of $\langle \mu^2 \rangle$ is Gaussian distributed about the mean. After this is done, and an increase of the threshold is used, an expanded core set of atoms is found.

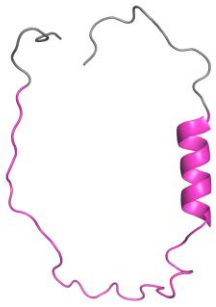
In the FindCore2 and Cyrange algorithm, a superposition of the models is used to find the mean of the superimposed models, which has the average x,y,z of the well-defined coordinates. After expanding the set of well-defined atoms, by changing the $\langle \mu^2 \rangle$threshold, the mediod model is recalculated, and the process is redone. Typically 2 or 3 iterations is required for a convergence of the atomic core set.

The 'core set' of atoms is used in the superposition of the models in the ensemble to evaluate the structure. More atoms are included in the Expanded FindCore2 or Cyrange set than in just using a DAOP or FindCore algorithm. Sidechain superposition is improved in using the variance matrix, in the Extended FindCore and Cyrange calculations.

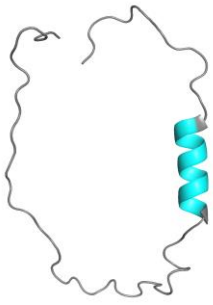
It is found by inspecting the output of Expanded FindCore and Cyrange that very similar ranges for the residues contained in the well-defined set are generated. Typically the regions from FindCore2 and Cyrange are almost the same in the collection of 90 unique proteins in the RDC database. Figure 2(a-d) shows an example, PDB ID 2K5K, where the 2 programs substantially differ. Figure 2(a,b) highlight these residues from the two programs in the 1st model. Figure 2(c,d) show the same but for all 10 models. In this case the helix is considered well-defined by both programs. However, there are 31 more residues from FindCore2. It notably superimposes well 2 segments and a helix, whereas Cyrange limited its output to only the helix.

Figure 2: 3d image of PDB ID 2K5K protein with colored residue regions of FindCore2 and Cyrange. (a) and (b) show in highlight the regions of well-defined for the 1st model, residues 14-54 and 15-24. (c) and (d) show the same but for all 10 models in the PDB ensemble.

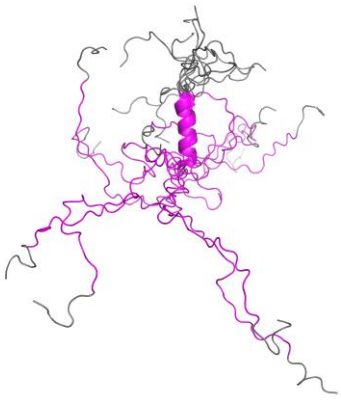
a



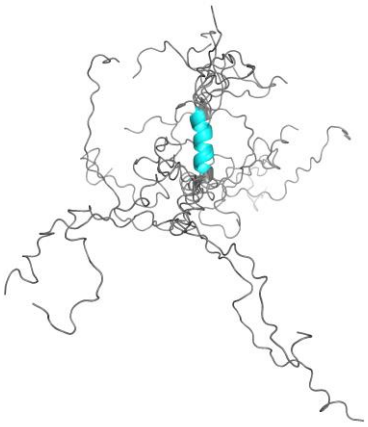
b



c



d

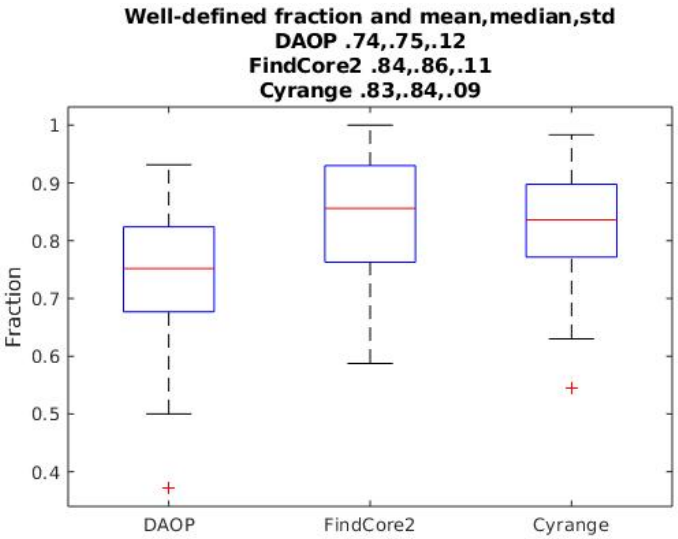


Expanded FindCore, DAOP, and Cyrange well-defined sets are compared using 90 pdb files in the RDC protein database. It is found by inspecting the output of Expanded FindCore and Cyrange very similar ranges for the residues contained in the well-defined set are generated. DAOP is more restrictive due to using just the constraints of separate dihedral angles. Expanded FindCore and Cyrange also use this restriction in the construction to find the initial set of residues, but include the inter-atomic distance variance in the calculations, giving more atoms and residues. From 90 proteins, of an average of approximately 100 amino acids each, the number of well-defined amino acids was calculated.

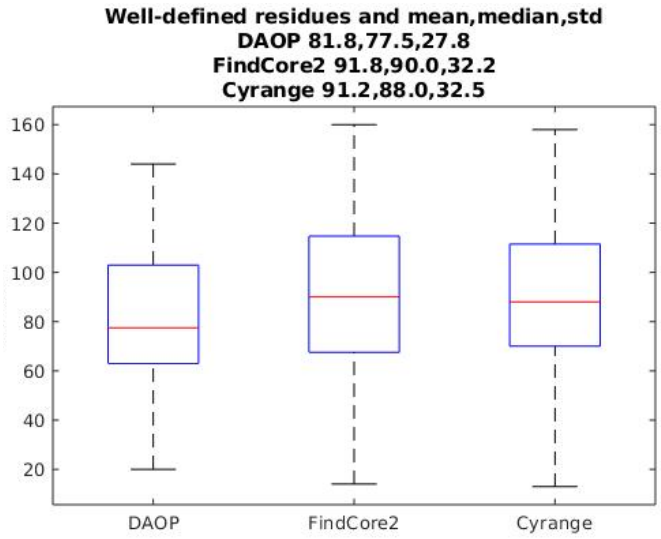
The downloadable Cyrange software was used with default parameters. FindCore2 and DAOP were used from the PdbStat software. Figure 3(a) shows the mean, median, and standard deviation of both the fraction of well-defined residues from the total and also the number in Figure 3(b). Typically, 74% of the residues are in this set from just dihedral angle constraints and 84% if the variance approach is used. In Figure 3, the DAOP constraint of 1.8 was used. There are approximately 10% more amino acids in the expanded set from the variance method. The ranges in the total residues plot could be misunderstood in the range; by inspection however 1ID6 has only 15 residues. This lessens the lower range limit.

FindCore2 and Cyrange usually differ in the number of well-defined residues by 1 or 2. There are some examples in which the difference is greater than 10. These are listed in Table 1.

Figure 3: Distributions of fraction (a) and total number of well-defined residues (b) as found from the 150 RDC database having 90 proteins in different media (peg, phage, and gels).



a



b

Table 1: Out of the 90 proteins, 9 protein structures differ by more than 10 well-defined residues. The 2nd column has the number of additional residues from FindCore2 not from Cyrange. The 3rd column has the number of additional Cyrange well-defined residues compared with FindCore2. These additional residues are unique.

PDB ID	FindCore2	Cyrange	Total residues
2HFD	0	13	132
2K5K	31	0	62
2KCU	12	0	166
2KI8	23	0	146
2KK8	10	0	84
2KKV	0	14	121
2KL1	12	0	87
2KWB	0	14	183
2KZV	0	14	92

Residual Dipolar Coupling (RDC) database and well-defined:

Residual dipolar couplings are not used directly in the previously discussed algorithms for defining well-defined residues. RDCs are very important in structure determination [10], especially in larger deuterated proteins in which there is a lack of NOEs (distance constraints) and a large number of RDCs [13,14]. Like the variance approach, which treats the entire set of well-defined atoms as a whole in calculating the mediod model and the threshold parameter, individual calculated RDCs also require all of the ^{15}N-H, ^{13}C-H, … coordinates of a protein in a collective calculation of a single RDC. RMSDs, Pearson coefficients, and Q-factors are used to compare different models in the protein structure. In the next section, these

measures are used in comparing FindCore2, Cyrange, and with no restriction to well-defined residues. The match of measured RDCs to the geometry of the models is shown to improve by using only well-defined residues in the RDC back calculation by redefining the calculation.

The recently developed protein RDC database, using only ^{15}N-H vectors, has 90 deposited PDB structures from the NESG [cite where], and 150 sets of protein RDCs in different media, each with typically 100 residues. The database is used to compare the different definitions of well-defined and not well-defined regions from the perspective of RDCs. This is similar to the inter-atomic distance variance in the sense that all atoms are used to calculate an RDC. The back calculation of an ^{15}N-H RDC from model coordinates and data requires all the ^{15}N-H pairs of coordinates in the data set in order to calculate an individual RDC. This back calculation is used in order to determine the best orientation of the protein (or domain), in turn, the orientations of the individual ^{15}N-H vectors in the protein model. All PDB's were analyzed in the context of RDCs. Software for back computing the RDCs from experimental data is available in the database, and in [15,6].

Section 3: Residual dipolar couplings in proteins, well-defined regions, and back calculation comparison

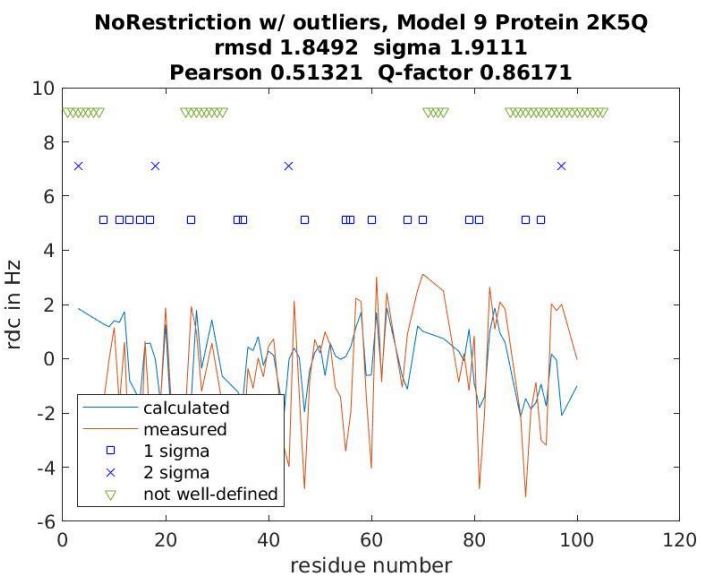
Residual dipolar couplings (RDC) measured in aligned media give much structural protein information, particularly in larger proteins, in which deuteration decreases the number of NOE measurements. The database presented in this work has a catalog of RDC data of 90 proteins, in 149 total aligned proteins, collected and determined by the NESG and collaborators. There are approximately 3000 structure models (3000 comparisons with measurements) and several hundred thousand RDCs, both measured and back calculated.

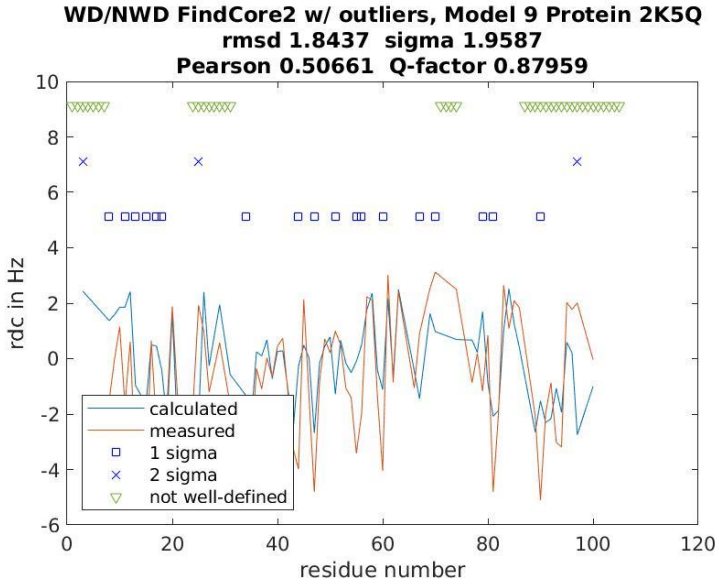
Using the measured RDC data and the deposited PDB coordinate information, the Saupe order tensor [11] is calculated and singular value decomposition [12] is used to back calculate the orientation of the protein model in order to calculate the RDCs. All but 5 of the proteins are single domain. A review of RDC back calculations is here [10]. Comparison plots of the measured and back calculated RDCs are shown in Figure 4 for a typical model, PDB ID 2K5Q, model 14. The root mean square deviation, the Pearson correlation coefficient, and Q-factor are given in the figure. These figures and additional information for each model are in the database.

Figure 4 has 2 examples: 4(a), NoRestriction on residues used in RDC back calculation, and 4(b) FindCore2 residues. In addition, any abs(measured-calculated)>2 standard deviations (std) in Hz of the back calculated and measured RDC/residue is marked by an 'x' to indicate an outlier. If the difference is between 1 and 2 sigma, then a square is used. The outliers of greater than 2 sigma are used in the RDC, RMSD, Pearson, and Q-factor coefficients, and the elimination of the outliers improves these quantifiers. These 'x's and boxes in Figure 4 point out difference in the comparison of back calculated to experimental measurement. The comparison plot is different between using all the residues and only well-defined as Figure 4(a,b) shows. The differences are usually not much. The not well-defined residues are marked in 4(a,b) by triangles and these residues are not used in the order tensor calculation of 4(b). FindCore2 and Cyrange give essentially the same results.

The use of RDC data can be used as a method to identify badly modeled regions of protein structures. This in turn can be used for further refinement and an improvement in the structure determination and validation. However, due to the non-local nature of the RDC back calculation, simply correcting the order tensor calculation by not using a region of the protein does not translate into correcting the back calculation in that region. Figure 4 is a conservative example. As shown in Figure 4(a) there are badly modeled 'x' coordinates if compared to Figure 4(b), which has one less 'x'. However, there could still be additional bad matches in the well-modeled residues, signified by the squares, which can be improved by refinement of the protein structure using RDCs. There are also fewer 'squares' of sigma 1-2 in 4(b) than in 4(a). The reduction of 'x'^{s} and squares is about 10% in not including the well-defined residues in the order tensor calculation, and this example contains the outliers in the RDC back calculation. This effect of eliminating the >2 outliers is examined in the next section, see Figure 5(a,b); this figure shows the comparison between NoRestriction and FindCore2 residues.

Figure 4: Back calculated residues overlaid with measured RDCs for peg 2K5Q model 9. RMSD, Pearson coefficient, Q-factor, and std of differences is shown. The 1 and 2 sigma regions are marked with x's and squares. Also, the not-well-defined residues according to FindCore2 are shown with a triangle. 4(a) uses all residue in RDC back calculation and 4(b) uses FindCore2 well-defined residues found from the inter-atomic coordinate variance of 20 models. All back calculated RDCs are shown.





Section 4: Well-defined (WD) and not well-defined (NWD) - NoRestriction, FindCore2, Cyrange from RDCs

Previous comparisons of NoRestriction, FindCore2, and Cyrange were visualized case by case in examples of overlaid PDB structures in 3-d visual tools such as Chimera, VMD, or PyMol. Quantitative measures such as those from PSVS were used []. In this section additional measures are introduced from RDCs, which can be used to evaluate the use of well-defined (WD) residues. The database has 150 sets of RDC data, each of which with a few exceptions have 20 PDB models. The full analysis of this data and analysis/comparison of the back calculated RDCs from the PDB file and data is not presented in this work. The information is used in this work to globally examine and compare the use of well-defined (WD) and not well-defined (NWD) regions. This is very relevant in a comparison because a statistically large sample of proteins is being used.

There are three types of quantifiers used to statistically compare the notion of well-defined in this mini-review. All of the models in the database are used to back calculate the RDCs from data. These are compared to experiment, and as shown in Figure 4 an RMSD, a Pearson coefficient, and a Q-factor can be calculated for each. The RDC RMSD mean, median, and std is calculated over all of the models of the aligned proteins and data, of which there are approximately 3000.

A residue can be WD or NWD in accordance with FindCore2 or Cyrange (or other techniques). If it is NWD, should it be used in the order tensor calculation to find the best orientation of the protein to match with RDC measurements? Also, should it not be used in validating the structural models after the protein model is oriented. The protein structures from the Protein Data Bank are ensembles of models, typically 20, which represent the dynamics of the conformational states, and the WD notion of residues comes from examining an inter-atomic distance variance matrix the ensemble to quantify fluctuations of atoms. There will be good and bad fits as a result, and typically 4 to 5 of these models can model a 'best fit' of PDB structural models to measured RDCs, in a least squares sense, not shown here. It is very relevant to point out that reductions of the number of residues can influence the RMSDs of each model, just by the number of residues used, and in all cases the RMSDs used in this analysis were found with the correct normalization with the number of (RDC-back calculated RDC) differences in the RMSD equation; restricting away from outliers or to WD residues is not explained by the reduction in the number of residues.

The statistical figures in this section used single domain proteins, all of which are solution NMR structures. Proteins 2K01, 2KS0, 2NWT, 2DSM, and 2JUW were not included. There were also 7 PDBs with only a single model structure. Both FindCore2 and Cyrange require at least 2 models to make the superposition to find the WD residues; these are not included in any of the calculations.

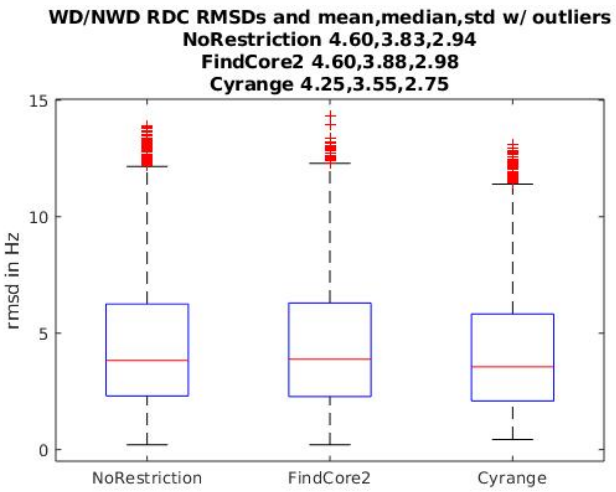
Figure 5(a) shows these distributions of the RDC RMSDs using no restriction on residues in the order tensor calculation, and a WD residue restriction from FindCore2 or Cyrange. Outliers of >2 sigma are used in the RMSD. The mean value without using any restriction gave 4.60 Hz. FindCore2 and Cyrange gave almost the same mean, 4.60 and 4.25 Hz. The order tensor is calculated using all residues in NoRestriction and only the WD residues in FindCore2 and Cyrange (this is typically 80% of the protein). Some models don't fit well in RDCs across all of the protein residues and some models fit extremely well (not shown); typically for a given structure, all the models agree about the same although there are exceptions in some protein ensembles. The 1st quantile in the box plot does have a large range, extending to 10 Hz, but this is largely reduced by excluding outliers. In all these cases in 5(a) the WD residues are used to calculate the order tensor, but all RDCs are calculated, from both WD and NWD (not well-defined). Both are included in the RMSD in Figure 5(a), but the outliers of >2*sigma are excluded in Figure 5(b).

Figure 5(b) shows the same as in Figure 5(a) but the 2 sigma outlier residue/measurements are not included in the model RMSDs. Outlier RDCs are in some sense another definition of WD residues. It only takes a few outliers to increase the RMSD of a particular model. These outliers indicate a deficiency in the molecular modeling from the point of view of RDCs coming from an ideal orientation of all of the ^{15}N-H vectors. The RMSD mean has decreased by 19% in eliminating outliers from NoRestriction and FindCore2, and 20% in Cyrange If the outliers were considered badly defined residues, the RDC's would statistically fit much better in the 3 cases (NoRestriction, FindCore2, Cyrange) of residues.

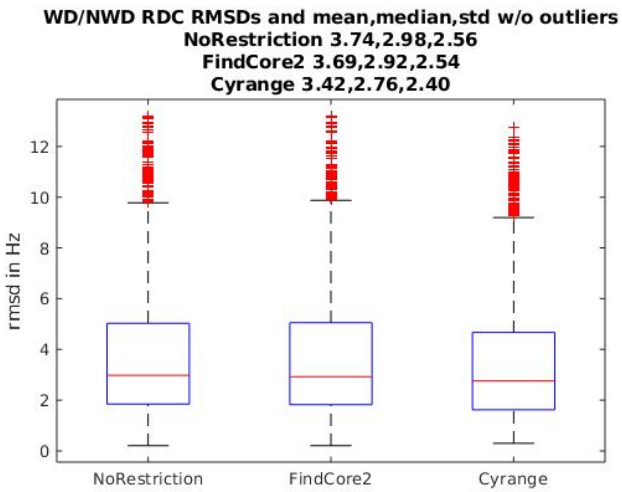
The next Figures 5(c) and 5(d) present the same analysis but restrict the RMSD to the WD residues; this is in addition to using only the WD residues in the order tensor. This point is relevant in the validation of the protein structure. There will be 80% of the residues typically from FindCore2 and Cyrange included in the RMSD due to their being on average 80% of the residues as WD. Note that the NoRestriction calculations didn't change because there is difference between 5(c,d) in NoRestriction except for eliminating outliers, which give the same NoRestriction results as in 5(a,b) (all residues are not WD restricted). In the cases of FindCore2 and Cyrange restricting to WD residues, the RMSD mean did decrease further, by 46% and 41% in 5(a)-5(c), which means including outliers from FindCore2 and Cyrange, but only using WD residues in the RMSD. The comparison of 5(b) and 5(d) gives a 35% and 31% reduction in the mean RMSD. The reduction is almost the same between including outliers and not including outliers, but only using WD residues in characterizing the quality of an ensemble of models. It isn't clear how this would happen by just reducing the number of residues and RDCs to the RMSD calculation. The choice of WD is structurally indirectly related to the ^{15}N-H vectors from the inter-atomic variance matrix.

Figure 5: Distributions of RDC RMSDs.

a

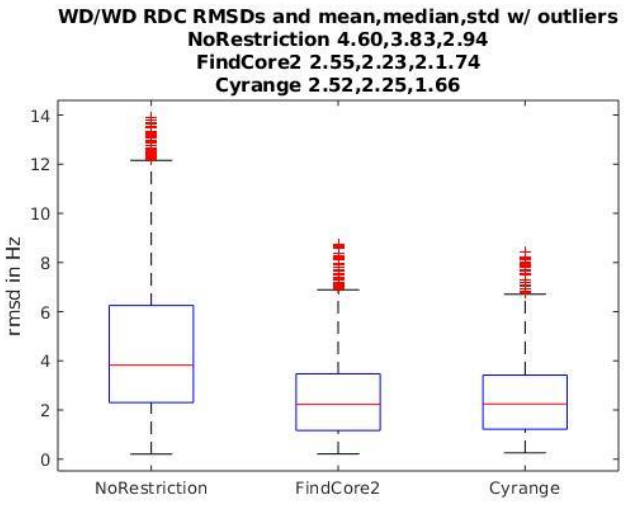


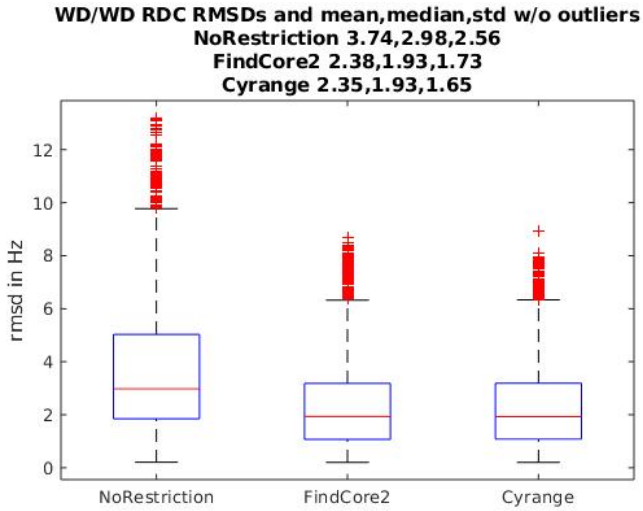
b



c

d





Two additional types of structure validation measures are also used to compare the different WD and NWD residues in the validation of protein structures, the Pearson and Q-factor coefficients. Figure 6(a-d) shows the mean, median, and std of the Pearson coefficient and Q-factors for the WD/NWD and WD/WD sets of back calculations. The drawback of the RMSD is that if the measured and back calculated RDCs are small, of the size 3 Hz or less, then the RMSD will be unnaturally small and it can give a misleading indication of the correctness of the model. This makes the RMSD a bad indicator in these examples. A set of small RDCs from an experiment can happen if the alignment media wasn't strongly concentrated enough [12]. A better quantitative measure of the comparison of RDCs is with a scale independent measure that takes into account fluctuations of the difference between measured and back calculated. The Pearson coefficient is appropriate for this, as it is scale free, and it quantifies the similarities of 2 vectors, such as those of residue RDCs. A Pearson coefficient between 2 vectors of 1 means that the vectors are identical (-1 means equal and opposite); A Q-factor of 0 means that the root mean square (RMS) difference of the 2 vectors is zero, but the Q-factor is normalized to the RMS of one of the vectors. The Q-factor is also a scale independent score, but it can also be misleading because it is a measure of the root mean square in the RDC comparisons without taking into account fluctuations.

In many cases from the point of view of RDCs, the Pearson coefficients and Q-factors are a better measure, although visual inspection of a plot such as Figure 4(a,b) is the best. Ideally a Pearson coefficient coefficient of .8-.9 and a Q-factor of .3-.4 indicates a good agreement between model to experiment. Four plots are shown in Figure 6(a-d) that are from the same set of models as in Figures 5(a) and 5(d). Figure 6(a,b) show the Pearson coefficient and Q-factor distribution from using the WD/NWD combination, that is, WD residues are used in the order tensor calculation and all residue RDCs are calculated. As Figure 5(a,b) showed in the RMSDs, these comparisons in Figure 6(a,b) are similar in NoRestriction, FindCore2, and Cyrange. Figure 5(a,b) included outliers.

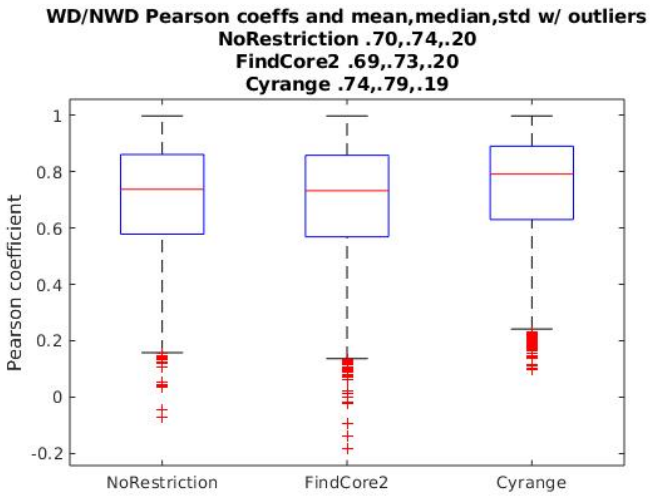
Figure 6(c,d) is the corresponding plot to Figure 5(d). This set of calculations used WD residues for the order tensor calculation and WD residues in Pearson and Q-factor calculations. Also, as in 5(d), 2 sigma outlier residues were eliminated from the calculations. The Pearson coefficient increased 29% if WD residues were used in FindCore2 or Cyrange, and 10% if no restriction was used on the calculations. The number of residues is not important in these coefficients, and the increase does show that the use of well-defined residues does increase the measure of quality between experimental and calculated RDCs. The Q-factor decreased by 41% and 35% by using WD residues in FindCore2 and Cyrange after excluding outliers, and the change was much smaller without using WD residues, 14%.

These three statistics can be used in either validation or as an indication of where refinement to a better structure can be made using the RDC data. Note also that some of the PDB structures did not use RDCs in making the structure or in refinement, and some of these structures agreed extremely well with RDC data. There is competition between different experiment types, NOESY, HSQC, RDC, Triple Resonance, etc…, in creating a structure. Some data can be more reliable

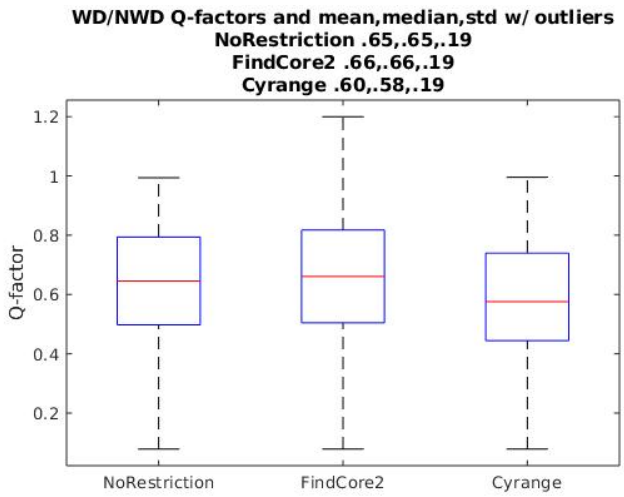
than other data depending on the environment of the protein and experiment, and the weights of each in the use of making the ensemble models or in refinement, including RDCs, should be taken into account.

Figure 6: This figure has Pearson coefficient and Q-factor distributions from 2 types of calculations. (a,b) use the WD/NWD set of RDC back calculations in which only WD residues are used in the order tensor. (c,d) restrict the coefficients to use only non-outliers in accordance with the differences of the measured to back calculated over all of the residues in the protein less than 2 sigma.

a

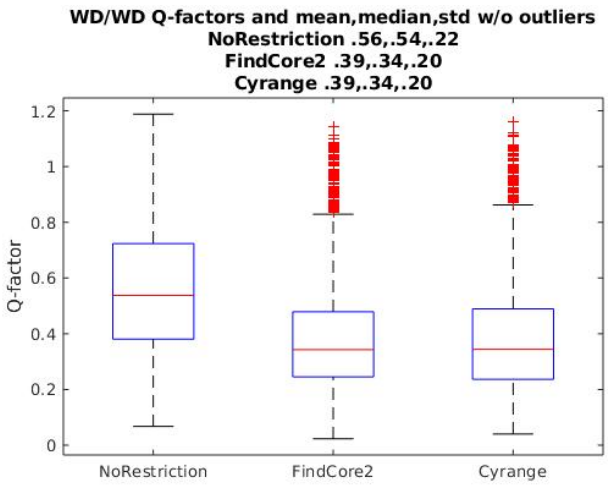


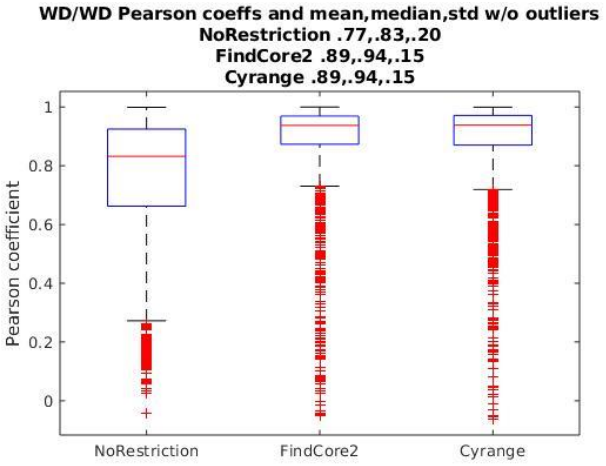
b



c

d





Section 5: Conclusions

The concept of well-defined (WD) regions of proteins is reviewed in this work. Several programs used to determine well-defined residues, FindCore2 and Cyrange, are compared. Next, WD regions of proteins are examined from the point of view of experimental and back calculated RDCs. WD residues from an ensemble of models in a PDB structure are necessary for an accurate superimposition of models, and can be relevant in protein structure validation in comparing with data. These definitions of WD residues and the algorithms to determine the residues are reviewed. In general, the fraction of WD residues of a protein are typically found to be 83-84% using an inter-atomic distance variance matrix method, such as in FindCore2 and Cyrange. The calculations are from an NESG database of 90 protein structures and 150 RDC datasets in 3 media. Differences between FindCore2 and Cyrange are pointed out.

A recently developed Residual Dipolar Coupling database is used in the analysis and review. This database has 90 NESG distinct proteins with 150 sets of RDC data from multiple alignments (peg, phage, gels). The typical size of the proteins in this database is 100 amino acids.

Residual dipolar couplings are used to compare the differences of back calculating with or without the restriction to WD residues. Back calculation of RDC's requires the coordinates of all the experimentally examined ^{15}N-H bonds (in this case ^{15}N-H's are used) in a model from the protein structure. The restriction of the protein residue to WD affects the back calculation, and from the recently made RDC database, it is shown that the outliers in the comparison of measured RDC's to back calculated are largely attributed to using non well-defined residues in the structure. Restricting to WD residues and non-outliers improves the match of experimental data to the structural calculations by generally 20 or more percent in the structural correctness quantifiers. This comparison was found using FindCore2 and Cyrange in the well-defined set of residues, and in comparison with no WD restriction. Different statistical quantifiers are used in the analysis: RMSDs of experimental RDC's with back calculated RDC's, Pearson coefficients of the protein's residues of the same, and Q-factors also of the same.

Unlike dihedral angles or an isolated spin pair approximation in interpreting NOE data, RDCs have to be interpreted non-locally due to all of the residue information in finding the best orientation of the protein or the domains of the protein. By using RDC's in the comparison of back calculated to measured, also with using only WD residues, validation and refinement of the structure can be made by either eliminating residues from the RDC calculation or by changing coordinates. This process can be as simple as taking a residue(s) and calling these non-ordered or by modifying the coordinates to make better agreement with experimental data. This process can be iterative in the inclusion or deleting of residues (i.e. ^{15}N-H vectors in the case of this database) by improving the order tensor and RDC back calculations from a set of improved RDC oriented WD residues of the protein. The use of RDC's is shown to be relevant in finding ordered and disordered regions in the protein from the 150 protein RDC database and well known software FindCore2 and Cyrange. Individually for these proteins these differences from using well-defined residues can be small, but overall from the approximately 3000 models, it is noticeable.

Acknowledgements:

This work was supported by a grant from the National Institutes of Health, …, and … The authors thank Khushboo Bafnak, James H. Prestegard and Roberto Tejero for useful discussions.

References:

[1] Protein Data Bank, https://www.rcsb.org/.
[2] Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology. 7: 95–9.
[3] Snyder DA, Montelione GT (2005) Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. PROTEINS: Struct Funct Bioinf. 2005;59:673–686.
[4] Kirchner DK, Guntert P (2011) Objective identification of residue ranges for the superposition of protein structures. BMC Bioinformatics 12:170-180.
[5] Snyder DA, Gullon J, Huang YJ, Tejero R, Montelione GT (2014) The expanded FindCore method for identification of a core atom set for assessment of protein structure prediction. Proteins 82(S2):219-230.
[6] Tejero R, Snyder D, Mao B, Aramini JM, Montelione GT (2013) PDBStat: A Universal Restraint Converter and Restraint Analysis Software Package for Protein NMR. J Biomol NMR 56(4): 337–351.
[7] Hyberts SG, Goldberg MS, Havel TF, Wagner G (1992) The Solution Structure of Eglin C Based on Measurements of Many NOEs and Coupling Constants and Its Comparison With X-ray Structures. Protein Sci. 1:736–751.
[8] Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A. 32:922–923.
[9] Kabsch W (1978) A discussion of the solution for the best rotation to related two sets of vectors. Acta Cryst. A34:827-828.
[10] Prestegard JH, Al-Hashimi HM, Tolman JR (2000) NMR structures of biomolecules using field oriented media and residual dipolar couplings. Quart Rev Biophys. 33(4):371–424.
[11] Saupe A, Englert G (1963) High-resolution nuclear magnetic resonance spectra of orientated molecules. Phys. Rev. Lett. 11:462–464; Saupe S. (1968) Recent results in the field of liquid crystals. Angew. Chem., Int. Ed. Engl. 7(2):97-112.
[12] Losonczi JA, Andrec M, Fisher MWF, Prestegard JH (1999) Order Matrix Analysis of Residual Dipolar Couplings Using Singular Value Decomposition. J. Magn. Res. 138,334 –342.
[13] Bax A, Kontaxis G, Tjandra N (2001) Dipolar couplings in macromolecular structure determination. Nucl Magn Reson Biol Macromol B 339:127–174.
[14] Schwieters CD, et al (2010) Solution structure of the 128 kDa enzyme I dimer from Escherichia coli and its 146 kDa complex with HPr using residual dipolar couplings and small- and wide-angle X-ray scattering. J Am Chem Soc 132(37):13026–13045.
[15] Valafar H, Prestegard JH (2004) REDCAT: a residual dipolar coupling analysis tool. J Magn Reson. 167(2):228–241. Tolman JR, et al (1995).