Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Magnetic Resonance 167 (2004) 228–241
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

REDCAT: a residual dipolar coupling analysis tool

Homayoun Valafar[a] and James H. Prestegard[b,*]

[a] Southeast Collaboratory for Structural Genomics, Department of Biochemistry and Molecular Biology, University of Georgia, GA 30602, USA
[b] Complex Carbohydrate Research Center, University of Georgia, 220 Riverbend Road, Athens, GA 30602, USA

Received 2 September 2003; revised 16 December 2003

Abstract

Recent advancements in the utilization of residual dipolar couplings (RDCs) as a means of structure validation and elucidation have demonstrated the need for, not only a more user friendly, but also a more powerful RDC analysis tool. In this paper, we introduce a software package named REsidual Dipolar Coupling Analysis Tool (REDCAT) designed to address the above issues. REDCAT is a user-friendly program with its graphical-user-interface developed in Tcl/Tk, which is highly portable. Furthermore, the computational engine behind this GUI is written in C/C++ and its computational performance is therefore excellent. The modular implementation of REDCAT's algorithms, with separation of the computational engine from the graphical engine allows for flexible and easy command line interaction. This feature can be utilized for the design of automated data analysis sessions. Furthermore, this software package is portable to Linux clusters for high throughput applications. In addition to basic utilities to solve for order tensors and back calculate couplings from a given order tensor and proposed structure, a number of improved algorithms have been incorporated. These include the proper sampling of the Null-space (when the system of linear equations is under-determined), more sophisticated filters for invalid order-tensor identification, error analysis for the identification of the problematic measurements and simulation of the effects of dynamic averaging processes.

Keywords: SVD; Order tensor; RDC; Dipolar coupling

1. Introduction

The utility of residual dipolar couplings (RDCs) has increased over the past few years [1,2]. They have been used to refine or directly determine the structure of proteins [3–10], nucleic acids [11,12], and carbohydrates [13–15]. They have also been used to deduce relationships of sub-units in multi domain proteins and describe bound ligand geometry [16–19]. This emergence of a broad range of applications of RDCs has revealed the need for better acquisition techniques and the need for more user friendly and powerful analysis tools. Significant advances have been made in both of these areas [4,20–26], but need for new analysis tools remains high. In part this is because RDCs represent a fundamental change from distance dependent to orientation dependent data. We present here a program called REDCAT, for Residual Dipolar Coupling Analysis Tool, which builds on orientational principles to allow the efficient interactive calculation and analysis of RDCs.

The orientational dependence of RDCs stems from a contribution to couplings that is proportional to the average of $(3\cos^2(\theta)-1)/2$ where θ is the angle between the external magnetic field and the vector of interest. Once coupling contributions are collected, they can be decomposed to give information regarding the strength of the alignment as well as orientation of a molecular fragment. The extracted information can then be used to perform more complex tasks such as orienting different rigid components of a molecule with respect to each other, or studying the relative orientation of components in complexes such as a ligand bound to a protein. This decomposition and extraction of the information is not however trivial and analysis is greatly facilitated by more powerful and user-friendly software analysis tools.

One previously described program named order-ten_svd [27] utilizes singular value decomposition

* Corresponding author. Fax: 1-706-542-4412.
E-mail address: jpresteg@ccrc.uga.edu (J.H. Prestegard).

1090-7807/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2003.12.012

(SVD) and Monte Carlo sampling as the core methods for solving a system of linear equations relating measured couplings to elements of an order tensor. However this package suffers from several shortcomings such as the lack of a graphical user interface (GUI), the use of a number of libraries originated in both C and Fortran, and limited tools for presentation and analysis of the data. The lack of a GUI can very easily reduce the usability of any program, and libraries originated in different languages always make code less comprehensible. Finally, this program only provided order tensor solutions without any other analysis tools. REDCAT addresses most of these deficiencies. This program and its documentation and tutorial are available via our web site at tesla.ccrc.uga.edu.

REDCAT's graphical-user-interface has been implemented within the Tcl/Tk environment. Tcl/Tk is highly portable and requires no compilation of code. Even though interpreted languages suffer from slow execution time, overall, the effect of a slow GUI is negligible since a small fraction of the total execution time is spent on the GUI. The back-end computational engine of REDCAT is implemented in C/C++ with modest optimization efforts and is, therefore, reasonably fast. The computational engines receive their inputs from the command line and can be pipelined for easy automation and batch processing. Since this program has been developed and tested on Linux systems, and because of its modularity, it is highly amenable to parallel implementation on cluster like environments.

REDCAT integrates a number of additional analysis tools in comparison to its predecessor orderten_svd. The following is the list of added analysis features:

1. Easy manipulation of the data during an analysis session via the GUI implementation.
2. Easy extraction of principal order parameters, GDO [1], S_{zz}, η, and the Euler angles that allow the transformation of a structure into its principle alignment frame (PAF).
3. More meaningful methods of screening for valid order tensor solutions.
4. Back-calculation of RDCs with any given tensor and structure.
5. Calculation of the rmsd between back calculated and measured RDCs.
6. Report of the best solution tensor (in the rmsd sense).
7. Error analysis that allows the identification and isolation of problematic measurements.
8. Computation of error values that will produce solutions for the given coordinates and RDC data.
9. Dynamic averaging of RDCs subject to systematic motion.

The above topics are introduced and illustrated in the following sections.

2. Residual dipolar coupling

The program is based on a now well-established description of residual dipolar couplings. Residual dipolar couplings arise from the interaction of two magnetically active nuclei in the presence of the external magnetic field of a NMR instrument [1,2]. Eq. (1) describes the average angular dependence of the RDC between a pair of spin 1/2 nuclei.

$$D_{ij} = \frac{-\mu_0 \gamma_i \gamma_j h}{(2\pi r_{ij})^3} \left\langle \frac{3\cos^2(\theta_{ij}(t)) - 1}{2} \right\rangle. \tag{1}$$

Here, D_{ij} is the residual dipolar coupling in Hz between nuclei i and j, γ_i and γ_j are nuclear magnetogyric ratios, r_{ij} is the internuclear distance (assumed fixed), and $\theta_{ij}(t)$ is the time dependent angle of the inter-nuclear vector with respect to the external magnetic field. The brackets signify the time average of the quantity. Normally, the random, isotropic sampling of angles by a molecule tumbling in solution reduces the RDC to zero. This isotropic sampling may be made anisotropic by a magnetically induced alignment or with the aid of various types of liquid crystalline media [28]. This anisotropic sampling will result in a measurable RDC that is indicative of the average orientation of an inter-nuclear vector.

When RDCs can be measured for several vectors within a rigid molecular fragment, a description of the orientational preference and level of order of the fragment can be obtained. The dipolar couplings can be written in terms of elements of an order tensor containing the orientation and order information s_{kl}, and direction cosines relating various vectors to the arbitrarily chosen fragment frame (Eq. (2)). $D_{\max}$ in Eq. (2) is the nucleus specific collection of constants in Eq. (1) that corresponds to the splitting of resonance for a pair of nuclei separated by unit distance and perfectly aligned along the magnetic field. Coordinates of molecular fragments are used to determine the values for r and describe the direction cosines for each vector. Solution of a set of equations larger than the number of independent variables s_{ij} (5) yields a complete order matrix S.

$$D_{ij} = \frac{D_{\max_{ij}}}{r_{ij}^3} \sum_{k,l} s_{kl} \cos(\theta_k) \cos(\theta_l). \tag{2}$$

Linear algebraic analysis of the order tensor can reveal the strength of alignment along each of the principle axes of alignment (S_{xx}, S_{yy}, S_{zz}), the combined strength of alignment (generalized degree of order or GDO) and the alignment orientation with respect to an arbitrary molecular frame (Euler angles α, β, and γ) [1,29]. The derived information can be used in structure determination as well as in other applications. Therefore rapid and accurate extraction of this information

becomes very essential in the analysis and application of RDC data.

3. Description of the algorithm

The core of the algorithm employed by REDCAT is very similar to that of its predecessor orderten_svd; it utilizes singular value decomposition to provide a best solution to an imperfect system of linear equations. In addition it uses Monte Carlo sampling to generate possible input data sets consistent with the errors for input data. The general outline of the algorithm is illustrated in Fig. 1 as a flowchart.

The input to this program is very comparable to the one for orderten_svd. Each line of entry consist of six coordinates (x, y, and z for two nuclei), the value of a measured RDC, the maximum RDC for the type of nuclei at 1 Å distance, the allowed sampling range and a comment field. The underlying assumption in REDCAT is that a particular RDC reported as $D \pm \varepsilon$ has an equal likelihood of falling anywhere in that range. Therefore, the Monte Carlo sampling of the dipolar space is conducted in a uniformly distributed fashion.

Given the input file as described above, REDCAT performs some manipulation of data in order to reformulate the problem as a linear system of equations with the least dimensionality. Eq. (2) describing the RDC between any two nuclei can be rewritten as in Eq. (5) using the two distinct relationships listed in Eqs. (3) and (4) The first relationship (Eq. (3)) expresses the traceless property of any valid order tensor matrix. The second relationship (Eq. (4)) defines the direction cosines of any vector in terms of its Cartesian coordinates (x, y, z) and length (r).

$$S_{zz} = -S_{xx} - S_{yy}, \tag{3}$$

$$\cos(\theta_x) = \frac{x}{r}, \quad \cos(\theta_y) = \frac{y}{r}, \quad \cos(\theta_z) = \frac{z}{r}, \tag{4}$$

$$D = \frac{D_{\max}}{r^5}\left[(y^2 - x^2)S_{yy} + (z^2 - x^2)S_{zz} + 2xyS_{xy} + 2xzS_{xz} + 2yzS_{yz}\right]. \tag{5}$$

Eq. (5) is linear in terms of the elements of the order tensor matrix; hence the equation for a set of dipolar couplings can be rewritten in the following matrix form:

$$A_{m\times 5}S_{5\times 1} = D_{m\times 1}. \tag{6}$$

Here the dimension m is the number of entries in the input file, A is the collection of coordinates as shown in Eq. (5) and S is the vector consisting of five independent elements of the order tensor matrix listed in Eq. (5) (S_{yy}, S_{zz}, S_{xy}, S_{xz}, S_{yz}). The individual elements of the order matrix that are obtained on solving Eq. (6) can be used to completely describe a 3×3 order tensor using its traceless and symmetric properties (described in the next section).

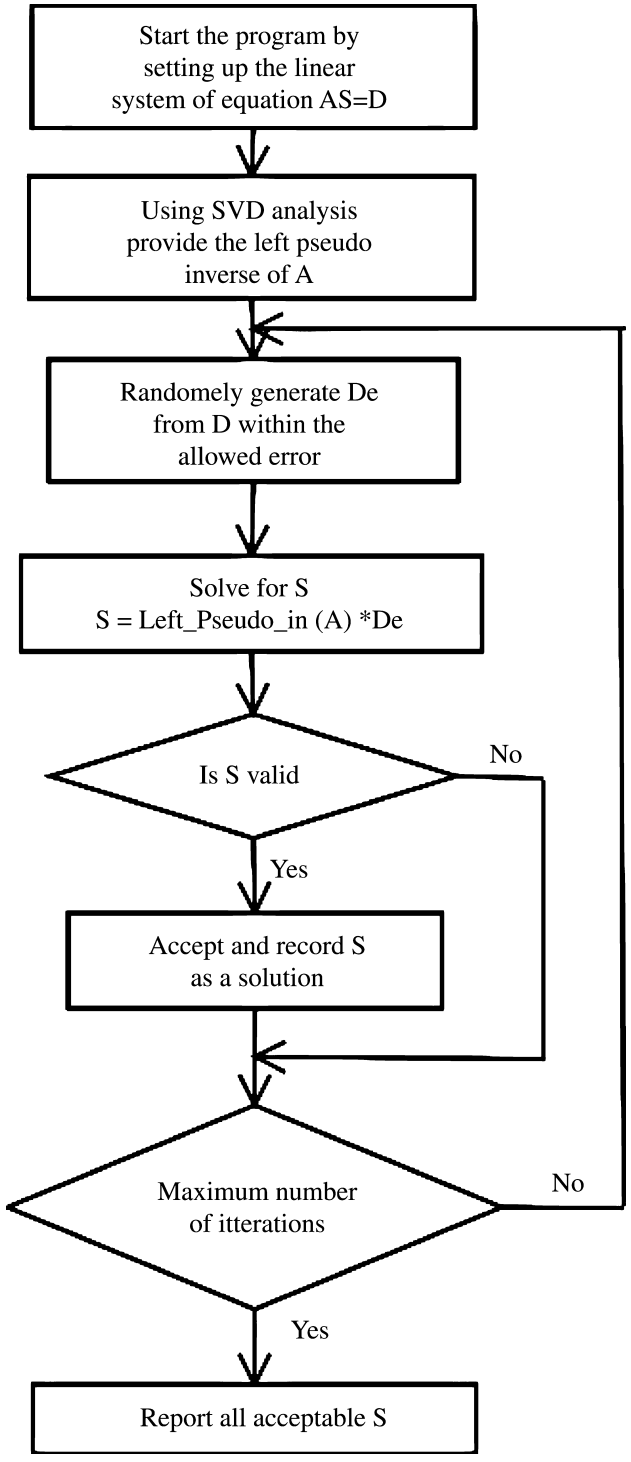

Fig. 1. Operational flowchart of REDCAT.

The contribution of SVD to this algorithm is very significant and clear. In general, under the conditions that the matrix A is ill conditioned, its inverse does not exist, and therefore, the solution to the above system of linear equations does not exist. When the matrix A is ill-conditioned (under determined or over-determined), SVD will provide a pseudo-left inverse for the matrix A. Therefore the following relationships will hold:

$$A^{-L}A = I, \tag{7}$$

$$AA^{-L} \neq I. \tag{8}$$

Using the first relationship, values for the unknown elements of vector S can be obtained by multiplying both sides of Eq. (6) by the left-pseudo-inverse of A as shown below:

$$S_{\text{optimal}} = A^{-L}D. \tag{9}$$

Multiplying by A and using Eq. (8), it is easy to show that the solution will not reproduce the original dipolar couplings. However the SVD algorithm will find the solution S that reproduces the original RDCs (D) in the best rmsd sense [30].

$$AS_{\text{optimal}} = AA^{-L}D = D' \neq D. \tag{10}$$

The above procedure does not guarantee that a particular solution S will produce RDCs within given error limits. Because of this phenomenon, the validation of the solution S for a given set of RDCs becomes necessary. Also, some inherent internal relationships within a valid order tensor need to be tested. REDCAT implements different filters for this validation.

4. Filtering of valid order tensors

Filters based on values of order parameters can be derived starting with the definition of the order parameters in Eq. (2). As a result of an intermediate step, the individual components of an order tensor matrix can be represented as the following:

$$s_{ij} = \left\langle \frac{3\cos(\theta'_i)\cos(\theta'_j) - \delta_{ij}}{2} \right\rangle, \tag{11}$$

where the $\cos(\theta'_i)$ are direction cosines relating different axes of the fragment frame to the magnetic field (these terms should not be confused with the direction cosines relating the inter-nuclear vector to the molecular frame used in Eq. (2)). Relationships among different elements of any given order tensor, and bounds on the elements, are derived from trigonometric relationships among the direction cosines and statistical properties of averages. In addition to the traceless and symmetric properties already incorporated in Eq. (5), REDCAT implements the following relationships as the means for determining validity of order tensors:

$$D - \varepsilon \leqslant AS = D' \leqslant D + \varepsilon, \tag{12}$$

$$-\frac{1}{2} \leqslant s_{ii} \leqslant 1, \tag{13}$$

$$-\frac{3}{4} \leqslant s_{ij} \leqslant \frac{3}{4}, \tag{14}$$

$$0 \leqslant \left(\frac{s_{ii}}{|S|}\right)^2 \leqslant \frac{2}{3}, \tag{15}$$

$$0 \leqslant \left(\frac{s_{ij}}{|S|}\right)^2 \leqslant 0.675, \tag{16}$$

$$0 \leqslant \eta = \frac{S_{xx} - S_{yy}}{S_{zz}} \leqslant 1. \tag{17}$$

The first criterion listed in Eq. (12) simply checks for the validity of the solution by comparing the back calculated RDCs to experimental values. The $|S|$ in Eqs. (15) and (16) refers to the total magnitude of the order tensor defined in Eq. (18) (note that this value is related to the previously defined GDO by a factor of $\sqrt{2/3}$). This is used to scale elements so that the boundary on possible values can be treated universally.

$$|S| = \sqrt{\sum_{i,j} s_{ij}^2}. \tag{18}$$

The quantity listed in Eq. (17) is designated with η and is proportional to the rhombicity of the alignment.

The above conditions will inherently be satisfied by the constraints imposed by the data when the system is over or fully determined (matrix A has rank of 5). However, these acceptance criteria are often very useful when the system is under determined and a number of order tensor solutions outside the normal ranges can be generated by the addition of null space components to the best solution. Generation of solutions for an under-determined system of equations is discussed in the following section.

4.1. Null space sampling

The problem of null-space arises when working with an under-determined system of linear equations. The best way to describe this problem is to compare a system of linear equations of the form $A_{m\times5}S_{5\times1} = 0$ to a scalar equation of the form $ax = 0$. While the scalar equation has only one trivial solution of $x = 0$, a system of linear equations may have infinite number of non-zero solutions [30]. Therefore, when working with an under-determined system of equations of the form $A_{m\times5}S_{5\times1} = D_{m\times1}$, one single solution can give rise to an infinite number of solutions of the form listed below. Here S_0 is a solution to Eq. (6), S_N is the null-space solution to Eq. (6) and α is any scalar multiplier.

$$S = S_0 + \alpha \cdot S_N. \tag{19}$$

This new solution can be substituted in the original equation to confirm its validity as shown below.

$$A \cdot S = A(S_0 + \alpha \cdot S_N) = A \cdot S_0 + \alpha \cdot A \cdot S_N = D + 0 = D. \tag{20}$$

Despite our best efforts at acquiring RDCs, we are often forced to work with under-determined data sets. A system is considered to be under-determined if the number of independent experimental data is less than the number of unknowns (less than 5 in the study of RDCs). An under-determined system can arise either when an insufficient number of data are collected or when the collected data correspond to linearly dependent vectors. Analysis of RDCs to provide orientational alignment of individual peptide planes in a protein, or carbohydrate rings in an oligosaccharide can be cited as examples of the above two conditions. For any planar fragment, such as peptide, absent its α carbon substituents, it can be shown that only three independent RDCs can be collected. For carbohydrate rings in which most CH vectors are axial and nearly collinear, collected RDCs will correspond to dependent vectors. In a more general case, one normally does not expect this condition to occur when the number of RDCs measured is larger than 5. However, accidental degeneracy can lead to under-determined system of equations without any warning.

An under-determined system of linear equations, however, need not impede the analysis of the data. In the presence of any relevant supplementary information, the study of these systems can be very feasible. For example, one can obtain the principle order parameters of a degenerate system from a number of different sources. A well described powder-pattern [31,32], calculation based on molecular shape [33], the order tensor of an associated macromolecule, or values of magnetic anisotropic susceptibilities that have been reported in the literature constitute some of these sources. A priori knowledge of the principle order parameters in effect reduces the number of required independent RDCs to 3.

An under-determined system of linear equations will always have an infinite number of solutions. For a more detailed discussion of under-determined system of linear equations and null-space please refer to Press et al. [30]. Under this condition, SVD will simply return the solution with the smallest vectorial magnitude (almost smallest GDO). This solution may not only be irrelevant in the context of an order tensor solution, but it may not even be a valid one when considered in the context of constrains listed in Eqs. (12)–(17). REDCAT will alert the user when the system of study is under-determined and will provide the option of sampling of the null space in order to reconstruct a complete order tensor. For example, when the number of null-space samplings is set to 10, for every solution, 10 random linear combinations of the null-space vectors will be added to the solution. This will allow the reconstruction of valid order tensors by the addition of the influence of the null space. This feature can be useful in finding exact solutions by filtering of the solutions when the principal order parameters are known a priori. Eq. (21) below shows the construction of order tensor elements using the null-space vectors and the optimal solution provided by SVD.

$$S = S_0 + \sum_{i}^{|\mathrm{NS}|} \alpha_i \vec{N}_i. \tag{21}$$

Here the α_is are generated by a uniform random number generator [−1,1], N_is are the vectors that span the null-space and |NS| is the dimensionality of the null-space. Fig. 2A shows solutions in the absence of null space sampling for a dipeptide system in which only four dipolar couplings were available. Solutions are presented in the form of a Sauson–Flamsteed projection that plots the points at which axes of principal alignment frames of various solutions pierce the surface of a globe drawn in the frame of the dipeptide fragment. Solutions for the direction of the x-axis (labeled S_{xx}) appear fairly well clustered. Solutions for the directions of the y and z axes are also well clustered, but show some interchange of axis definitions (this can result when η is near 1 and is not necessarily an indication of an under-determined set). This picture is, however, deceiving; only the S_0 parts of the solution appear leaving out solutions defined by the addition of $\alpha_i N_i$ terms. Fig. 2B illustrates the order tensor solutions to the same problem with sampling of the null-space. It is easy to see that a large portion of valid solutions would not have been discovered in the absence of null-space addition to the analysis.

5. Graphical user interface

A Graphical user interface (GUI) is the most essential component of a user-friendly program. Fig. 3 below is a snapshot of the initial screen after having loaded an input RDC file. Coordinates for each pair of coupled atoms, a measured dipolar coupling and an estimated error are shown in rows from left to right. Not shown is a value of $D_{\max}$ provided in each line of the input file. Values displayed can be altered here manually if necessary. A comment field carried from the input helps to identify each entry. The select buttons shown on the left-hand side of this figure allow for the inclusion/exclusion of a particular entry in the analysis.

Results of any analysis are displayed within the "Message!" window as demonstrated in Fig. 4. Information such as the rejection status, order parameter solutions, the corresponding Euler angles, best solution and error analysis will be concatenated to the content of this window.

Fig. 5 gives an example of an RDC back calculation and rmsd analysis. In this window S_{xx}, S_{yy}, and S_{zz} are the principle order parameters and a, b, and c are the three Euler angles relating the alignment frame to the molecular frame that are to be used in the back

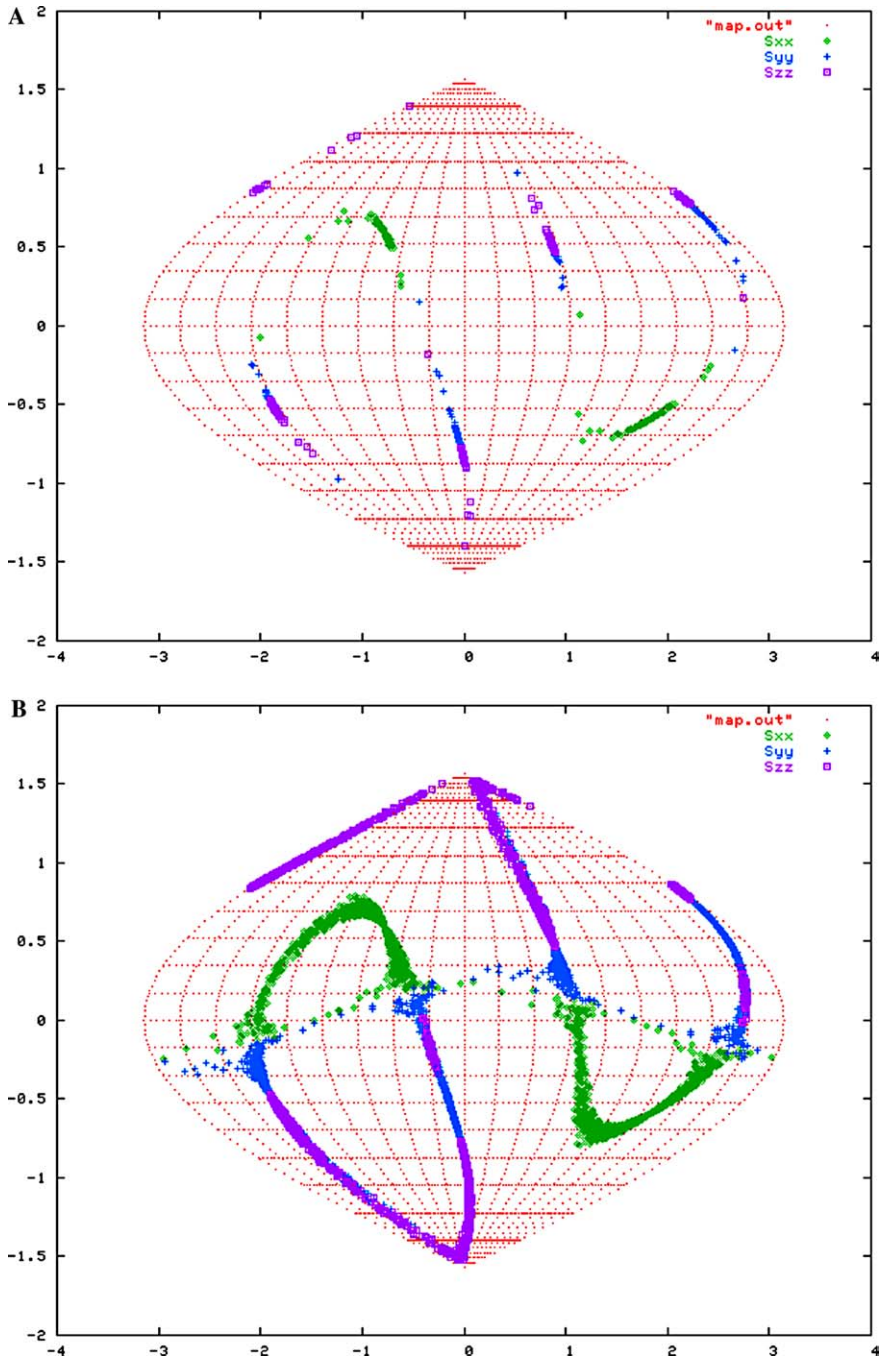

Fig. 2. (A) Solution space with only the best solutions. (B) Solution space with sampling of the null space.

calculation. Error is the range of uniformly added random noise to the back-calculated RDCs (this is useful if they are to be used as input in a subsequent calculation). It is typical to obtain S_{xx}, S_{yy}, and S_{zz} and their corresponding Euler angles from a list of solutions such as shown in Fig. 4. The text box at the bottom of Fig. 5 lists the back-calculated RDCs and the rmsd between these values and the measured data. When the "Substitute RDC" button is checked, the back calculated RDCs will replace the currently listed experimental RDCs in the input window allowing easy association with the vectors giving rise to the couplings.

6. Best solution report

A successful data analysis session will provide a large number of possible solutions that satisfy all RDC constraints. However, for a number of reasons it is convenient to select only one among these large numbers of

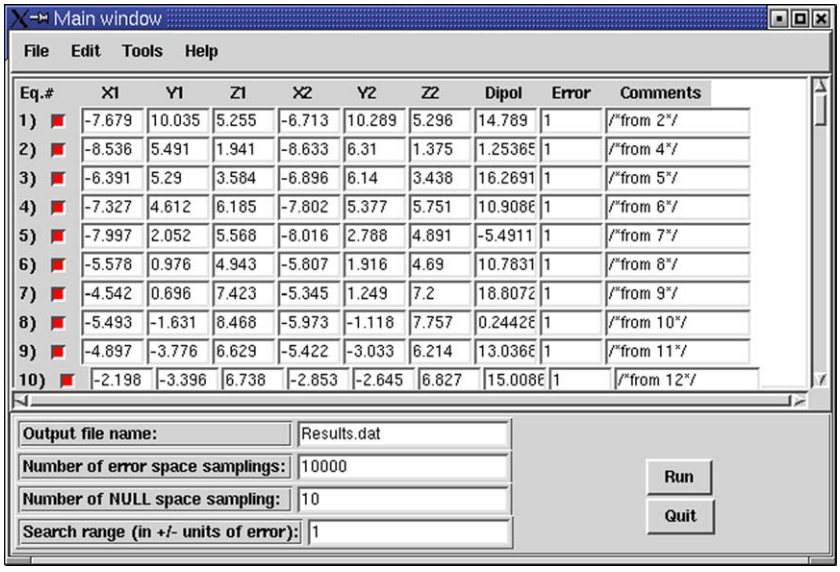

Fig. 3. Main screen of REDCAT.

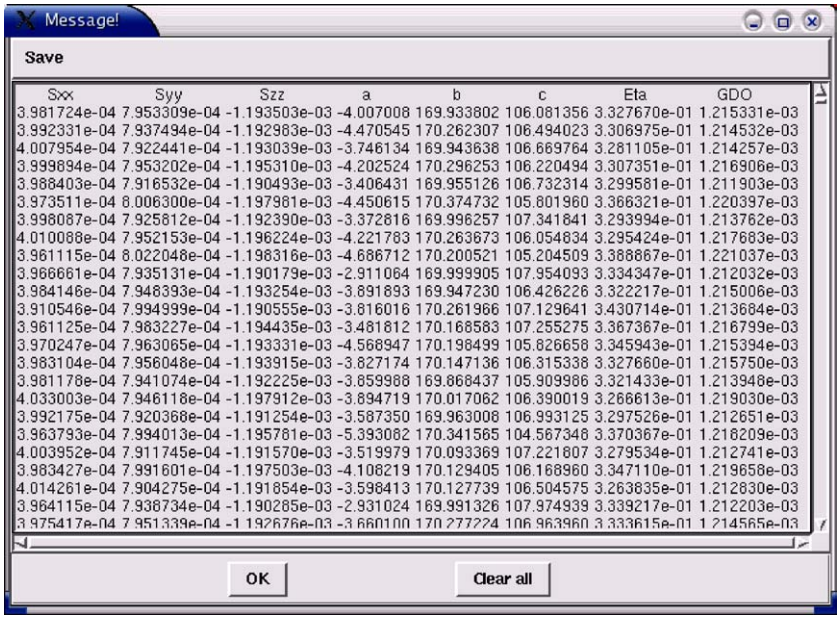

Fig. 4. Message window displaying analysis results.

solutions. This selected solution not only can be used to back calculate RDCs but also can be used to orient different fragments of the same complex with respect to one another. However, selection of different order tensor solutions can produce different results (significantly different results based on the range of errors). Therefore, it is useful to be able to isolate the best solution among a list of solutions. The "Best Solution Analysis" of REDCAT provides the best solution (in the rmsd sense). Note that the "best solution" in the rmsd sense may not satisfy error constraints on all RDC values and may therefore fall outside of the list of solutions. In this case REDCAT will not report the solution. This is only to prevent false interpretation of the best solution. The users can insist on viewing the best solution by expanding the error values. If the best order tensor solution is used for the back-calculation of the RDCs, the rmsd reported at the bottom of the back-calculation window will report the smallest rmsd value. Also, the best solution may or may not represent the median point of each cluster depending on the severity and exact nature of experimental error (systematic or random error). Figs. 6A and B below illustrate this phenomenon with synthesized data (Fig. 6A with S_{xx}, S_{yy}, and S_{zz} of 0.0008,

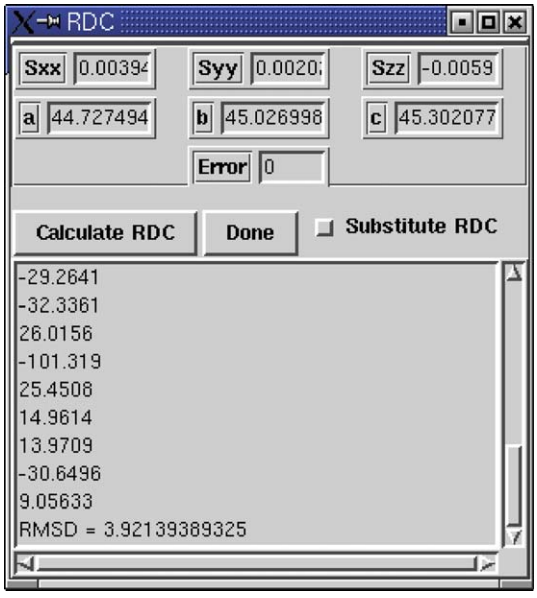

Fig. 5. Back calculation and rmsd analysis of the data.

0.0004, −0.0012, and 2 Hz uniformly added noise) and real data from a structural genomics initiative protein (PFU-1016054, Fig. 6B). As can be seen, the best solution may be in the center of a solution cluster or near the boundaries. Under some circumstances the best solution may even sit outside of the solution region. This situation may be indicative of the existence of systematic error in the measurement of the RDCs or systematic inconsistency of the data with the assumed model (such as bond length, planarity of peptide planes, etc.).

7. Error analysis

Analysis of RDC data is often complicated by the presence of measurements that are inconsistent with one another. Inconsistencies often arise because of error in measurement of the experimental data, miss-assigned spectral peaks, internal motions or models having poor bond lengths or local geometry. As is demonstrated below, a modest error in structure or value for even a single RDC can cause catastrophic rejection of sampled solutions by a number of equations. Occurrence of such a circumstance will force the user to alter the errors in a combinatorial fashion in hopes of isolating the inconsistent datum. The error analysis function of REDCAT will allow the identification of the problematic entries in a systematic fashion (as long as they are in the minority of a large set of entries). Furthermore, this analysis will produce a suggested range of errors in order to produce solutions.

Figs. 7A and B illustrate the rejection status (out of 10,000 samples) in the presence of a small amount of inconsistency. During the first experiment, the RDC

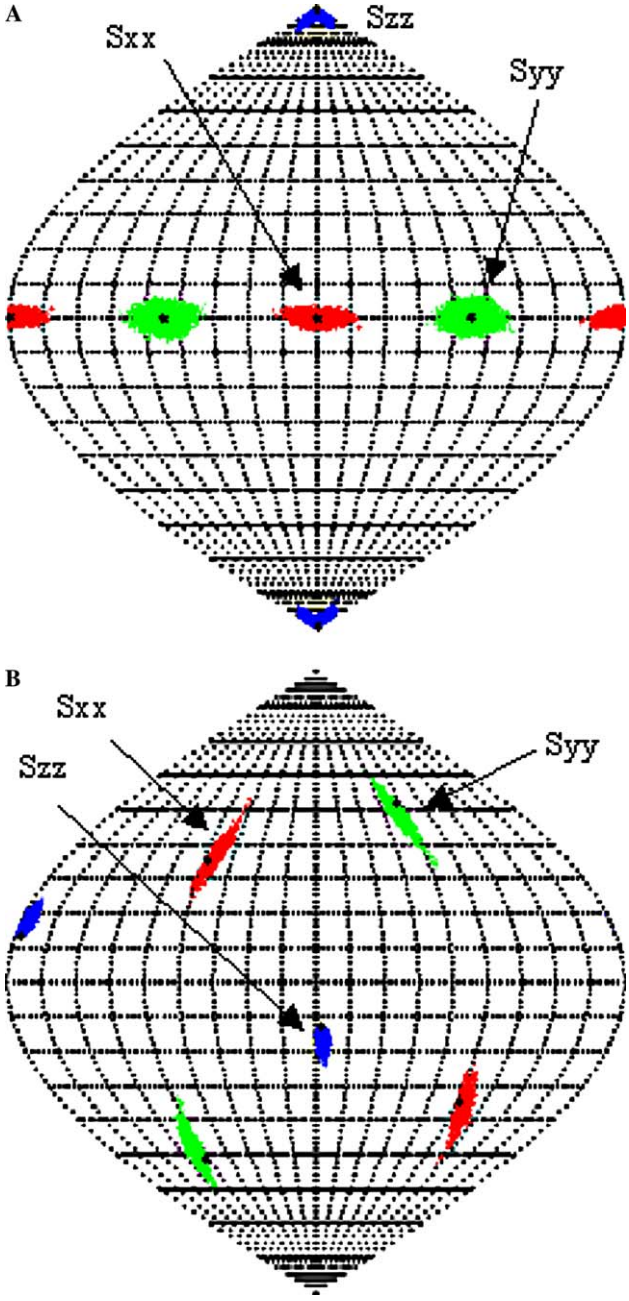

Fig. 6. (A) Solution space for couplings calculated from a dipeptide model. The best solution is indicated by *. (B) Solution space for experimental couplings from a dipeptide. The best solution is indicated by *.

of the first entry was changed from 1.73192 to 1.0. This minor change in RDC caused rejection of sampled solutions. The rejection status is shown in Fig. 7A. Based on the number of rejections, one can correctly isolate the problematic data entry (based on total number of rejections) in this particular instance. However when the error in the RDC (or the structure since structural error can be translated to RDC error) is moderate to severe, analysis of the number of rejections will not be very successful in the identification of the problematic data.

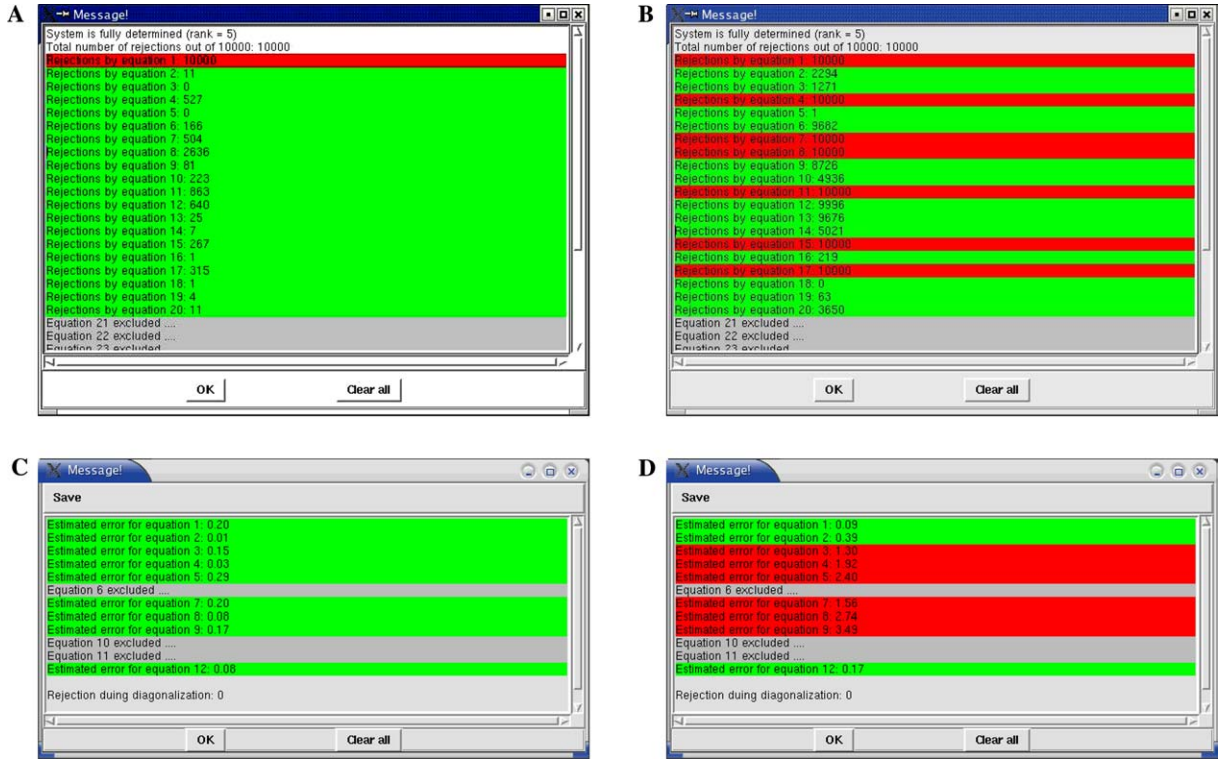

Fig. 7. (A) Rejection analysis with minor alteration of the first entry. (B) Rejection analysis with modest alteration of the first entry. (C) Error analysis applied to the right geometry of peptide planes 34 and 35 of protein Rubredoxin. (D) Error analysis applied to the wrong geometry of peptide planes 34 and 35 of protein Rubredoxin.

For example, alteration of only the first RDC from 1.73192 to −1.73192 will produce the results shown in Fig. 7B. Based on the data shown in Fig. 7B, entry numbers 1, 4, 7, 8, 11, 15, and 17 are all strong candidates for error since they produced a rejection 100% of the time.

Analysis of the above results shows that in the case of a minor error, the number of rejections is useful in the identification of the problematic entries, but when errors are moderate to large, as in the case of miss-assignment, utility decreases significantly. The error analysis function of REDCAT provides an alternative method of identifying the erroneous data. REDCAT accomplishes this error analysis by simply considering the best rmsd solution. As was mentioned before, the best solution is independent of the error limits. Therefore, comparison of back-calculated data to experimental data can show the changes required in the error boundaries in order to put the best solution inside the allowed RDC region.

The results of error-analysis are shown in Fig. 8 for the second problem mentioned above. The problematic entry has been identified clearly as the first entry. This figure suggests that the error boundary of the first entry needs to be set to 2.5 Hz to eliminate rejection by this entry. This number is meaningful when compared to the remaining errors, which are on the order of 0.5 Hz (a factor of five times smaller than the error reported for the first entry). The illustrated pattern becomes even more distinct as the error increases in magnitude. The success of this method in identification of the problematic entries will diminish with an increase in the number of entries that are in error. Alignment and the total number of entries, as well as distribution of these entries in the five dimensional order-parameter space will additionally affect the utility of this feature.

The error-analysis function of REDCAT can serve multiple purposes. For example under the condition of carefully collected and well scrutinized data, any significant error can be interpreted as deviations in structure and can therefore be used to accept or reject a certain proposed structure. This concept can easily be demonstrated by considering the problem of local torsion angle determination for residue 34 of the protein Rubredoxin [3]. Using the previously published RDC data, we can embark on the task of local structure determination by examining the fitness of the collected data with respect to possible torsion angles. Figs. 7C and D illustrate the results of error analysis applied to two different torsion angles. Fig. 7C lists the suggested errors for a dipeptide plane with torsion angles of −60°, −20°. Fig. 7D lists the results of the same analysis for an alternate geometry of −160°, 110°. Based on the suggested errors, one can conclude that the first structure is a much better fit to the experimental data than the

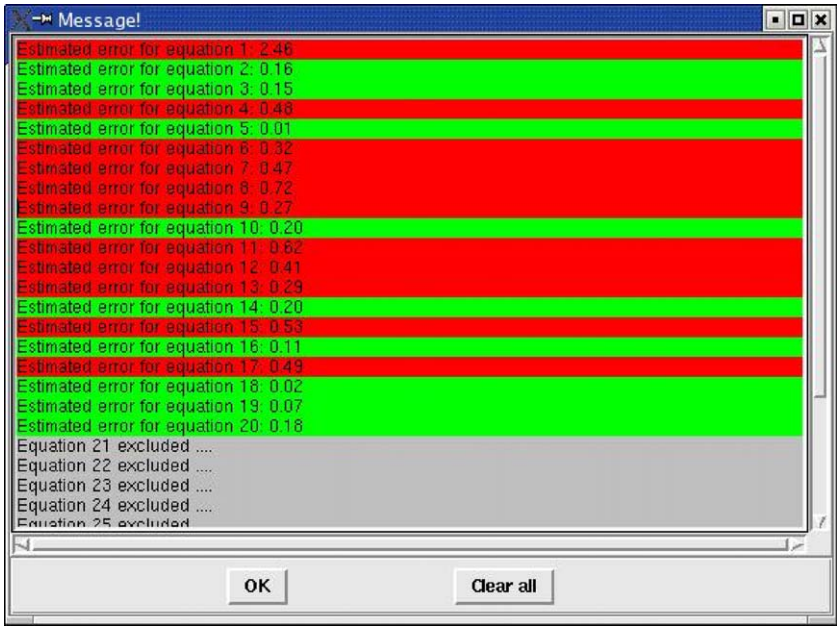

Fig. 8. The required expansion of error for each entry reported by error-analysis.

second one. On this basis the first structure can be accepted as the more likely structure. This conclusion is indeed in agreement with the torsion angles of $-57°$, $-31°$ obtained from the crystal structure of a nearly identical protein, 1BRF.

8. Rotation tools

Often, upon the completion of a successful analysis session, it is beneficial to rotate the domain or molecule of interest into its principal alignment frame. This rotational transformation is very useful in providing a set of structures having proper orientational relationships among multiple units of a complex, or different fragments of the same molecule. The information required for this transformation can be extracted by diagonalizing the order tensor solution. Once diagonalized, the rotation matrix consisting of the eigen vectors of the order tensor can be collected in a rotation matrix that relates the molecular frame to the alignment frame (or visa versa). This rotation matrix can be decomposed and described in terms of the three Euler angles that are provided by REDCAT. However for maters of convenience and consistency of definitions, rotation tools are incorporated into REDCAT. Two rotational tools namely "Rotate PDB" and "Rotate Coordinates" allow convenient rotation of either a given input PDB or the coordinates that are already loaded into REDCAT. Figs. 9A and B below illustrate the interfaces for these two functions. Coordinates x, y, and z that appear in Fig. 9A define the rotor axis. These coordinates can be obtained from the difference in coordinates of two points that define the axis.

9. Emulation of dynamics

It is frequently inappropriate to view an entire macromolecule as one completely rigid entity. A more likely situation is that parts of the molecule are rigid with respect to each other while other parts undergo internal reorientation. The dynamics experienced can be a random motion (such as the flopping of a loop) or a well-defined transformation between discrete states (such as the flipping of an aromatic ring of a protein sidechain). Such dynamic averaging can be used to explain inconsistent alignments reported by different parts of the same molecule. The dynamic averaging tool of REDCAT allows the calculation of RDCs subject to internal motions and can, therefore help resolve observed inconsistencies in alignment.

Although in general it is difficult to precisely describe a dynamic property of a molecule based on a limited set of experimentally collected RDCs, the validity of proposed motions can be tested by the use of the same set of data. This arises from the fact that the observed RDCs for the dynamical regions of a molecule are reduced in specific ways by the nature and extent of internal motion relative to the overall alignment frame of the molecule. Therefore, two different types of information are required for the emulation of dynamics. First, information on the overall alignment of the molecule is needed, and second, a description of the internal motion is needed. The overall alignment of the protein can be easily

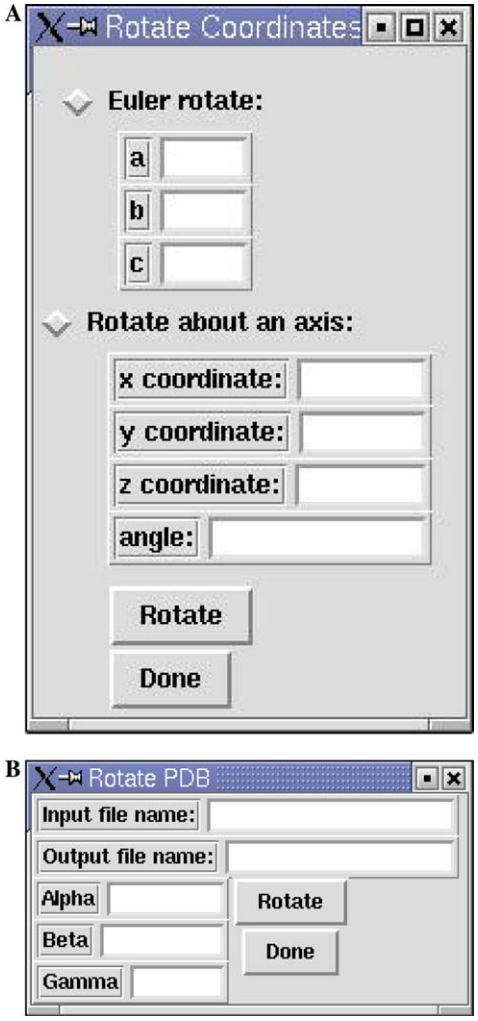

Fig. 9. (A) Interface for rotating already loaded coordinates. (B) Interface for rotating a PDB structure.

obtained by the analysis of RDCs from the static portion of the molecule using REDCAT (perhaps by the use of the best solution tool). The proposed motion can be obtained from molecular dynamic simulations or other more idealized models. The motion description is listed in the form of a series of entries indicating a distinct orientation in space and the weight of that state. Each orientation can be described in the form of three Euler angles or in the form of a rotation about a vector in space. The weight of each state can be interpreted as either the portion of time spent in each orientation or the fraction of population occupied in each state at any given time. If an ergodic process is assumed, then the two cases would be equivalent. The discrete representation of motion may not be a severe limitation to the utility of this function. Any continuous motion, for example, can be represented by a sequence of small discrete rotations. The accuracy of this representation is then dependent on the number of discrete steps describing a continuous motion.

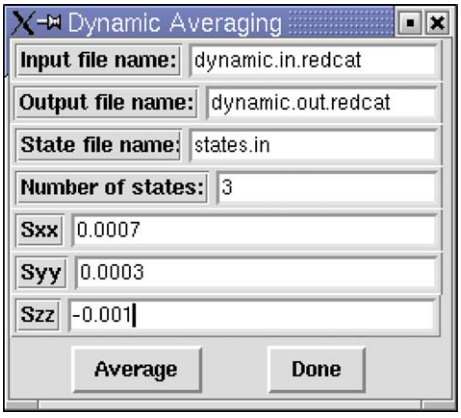

Fig. 10. Alignment of a terminal helix that undergoes C3 motion. $S_{xx} = 0.0007$, $S_{yy} = 0.0003$, and $S_{zz} = -0.001$ with the alignment frame coinciding with the molecular frame.

Fig. 10 illustrates the user interface to the dynamic averaging tool. The input fields correspond to the name of the input file (in REDCAT format), the name of the output file (produced by the dynamic analysis), the name of the file that contains the description of states, the number of states in that file and three principle order parameters. Note that since only three order parameters are used, all coordinates need to be expressed in the principal alignment frame. Fig. 11 below shows the alignment characteristics of a terminal helix that undergoes a three-state rotation in 120° steps about the vector (1, 2, and 3) with equal populations of each state. As illustrated in Fig. 11, even though the overall alignment

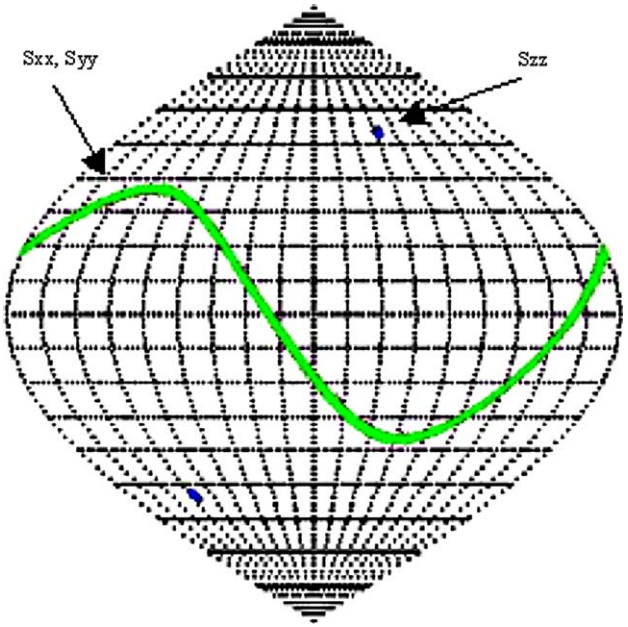

Fig. 11. Alignment of a terminal helix that undergoes C3 motion. $S_{xx} = 0.0007$, $S_{yy} = 0.0003$, and $S_{zz} = -0.001$ with the alignment frame coinciding with the molecular frame.

of the molecule is non-symmetric, the alignment reported by the terminal helix is axially symmetric with the axis of symmetry along the vector (1, 2, 3). This result has previously been shown analytically [34] and is observed in molecules with the required symmetry properties [35].

10. Inclusion of averaged data and CSA data in the calculation of order tensors

The above addresses emulation of the effects of dynamics on order tensors. A separate issue is the actual inclusion of averaged data in the initial calculation of an order tensor. It is not difficult to envision a situation where the only existing data from a particular region of a molecule are observed averages or a sum of contributions from discrete states. This clearly occurs when working with the RDCs from aromatic side chains of proteins. In most cases aromatic rings of tyrosines and phenylalanines undergo 180° flips on a sufficiently rapid time scale to average couplings. The existence of such averaging is reflected in the observation of a single set of resonances for symmetry related pairs of protons, and data likely to be affected by averaging are easily identified. A situation related to this averaging is seen where the sums of couplings may be easier to measure than individual couplings. The measurement of backbone C_α–H_α RDCs for glycines can be listed as one such example. Both H_α protons couple to the C_α giving, in principle, a doublet of doublets in the ^{13}C dimension of a spectrum. The positions of the outer lines at the sum of the couplings is well defined, but the position of the center lines is often not well defined due to overlap or second order effects. The sum of the RDCs is equivalent (except for a factor of two) to the average coupling of a H_α proton undergoing a two-state jump between H_α vector positions. Hence both cases can be treated using an averaging model. The following formulation can be utilized to include the RDCs originated from averaging regions of a molecule as an input to REDCAT. This formulation focuses on the representation of RDC in the five dimensional order parameter space (Eq. (22)). Here V_i indicates the ith position of the vector of interest. $D(V_i)$ is the average of all RDCs for that vector as a result of a discrete motion over n equally populated sites. $D_{\max}$ is adjusted up by a factor of 2 in cases such as the glycine example where the sum of couplings is measured. The resulting sum will constitute one single entry in our calculations. This functionality is implemented in REDCAT by placing the discrete vectors adjacent to each other in one block. An entry of "AVG" is placed for the RDC of all of the discrete vectors except the last one. The last entry then will have the value of the measured RDC (average observed value). An example of the input file is shown in Fig. 12

$$\vec{V}_i = (x_i, y_i, z_i),$$

$$D(\vec{V}_i) = (D_{\max}/n)\sum_i \frac{1}{r_i^5}\left[\left(y_i^2 - x_i^2\right)S_{yy} + \left(z_i^2 - x_i^2\right)S_{zz} + 2x_iy_iS_{xy} + 2x_iz_iS_{xz} + 2y_iz_iS_{yz}\right]. \quad (22)$$

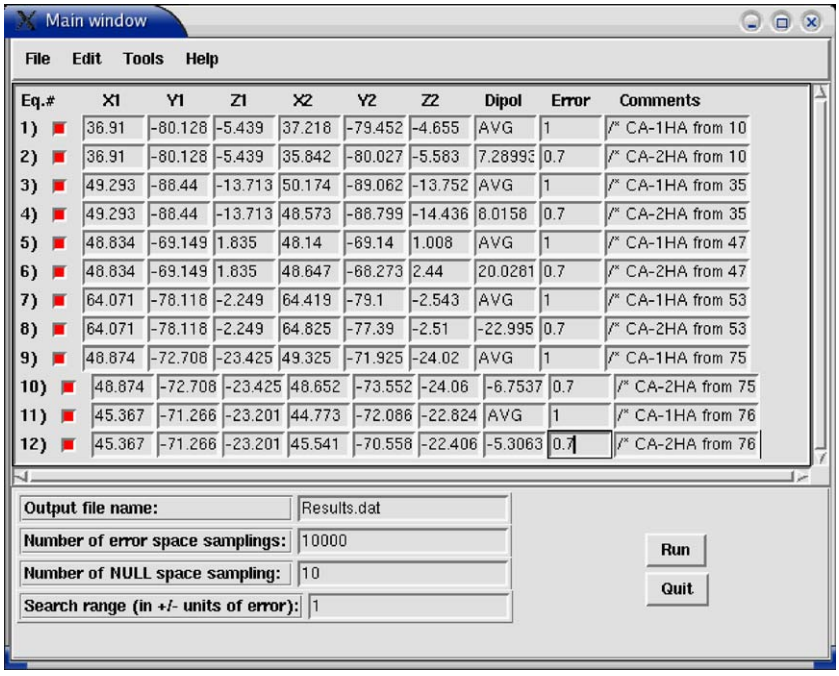

Fig. 12. An example of input file to REDCAT taking advantage of the averaging-analysis tool.

This module may be adapted for other types of data as well. One important type of data is chemical shielding anisotropy (CSA) offsets such as those that occur for carbonyl groups. It has recently been shown that CSA offsets can be rewritten in terms of a sum of two RDCs [36]. Appropriate vectors and a D_{max} can be deduced from known chemical tensors for the particular groups involved.

11. Discussion and conclusion

The GUI addition to the REDCAT program should significantly increase the usability of this software package while the integration of additional analysis tools in one package should increase its productivity. Tcl/Tk, which is the interpreted environment that provides the front end GUI of REDCAT, is highly portable and available for windows, Unix, Linux, and Mac operating systems. Furthermore availability of open-source visual programming packages such as visual Tcl (http://vtcl.sourceforge.net/) can facilitate the addition of custom build functions by individual users. This is also the same programming environment used for NMRDraw [37] and NMRView [38] minimizing the need to learn additional scripting languages.

Implementation of the computational engine of REDCAT in C/C++ ensures the best computational speeds. C and C++ are both very standard and well established programming languages with compilers available from the GNU web site (http://www.gnu.org) for a vast number of platforms. This separation of the computational engine from the graphical component provides a degree of flexibility that increases the utility of this package in a broad range of applications. The entire package can be installed on a single computer for provision of a user friendly analysis environment, while installation of the computational engine alone on a more powerful computer can provide automated batch processing of data. The computational engine can be executed independently in a pipelined fashion allowing easy porting of the package onto a Linux cluster allowing even larger number of processes.

Null-space sampling, best solution, error analysis, and dynamic averaging are anticipated to be the most useful additions provided by REDCAT. Null-space sampling can extend the range of utility of RDC measurements to cases where data are sparse. This addition will allow a meaningful analysis of systems with a small number of RDC data in conjunction with other information obtainable from independent sources. One can envision an extension of this algorithm to assist in the orienting of peptide planes and sugar rings with three or four data points. In addition, integration of information from alternate sources such as chemical shielding anisotropy (CSA) tensor or multiple alignment media pave the road to complete structure determination only based on orientational restraints.

Although in the current implementation of REDCAT the "best solution" is a tool of convenience, further understanding of the shifting of the best solution from the center of a cluster of solutions can be diagnostic in estimating systematic errors in the data. Furthermore, it is possible to correlate the systematic deviation with each of the vectors. Finally, error analysis cannot only be used for the modification of the estimated errors for the provision of solutions, but can also be used in identification of problematic vectors. Once the inconsistent data/vectors are identified, it is possible to engage in the task of structure refinement or reassignment in order to correct the erroneous components. In light of the recent advancements in the area of threading techniques in structure determination, one can use the results of a REDCAT analysis to confirm or reject a proposed structure easily.

Acknowledgments

Funding for this project was provided by grants from the National Institutes of Health, GM062407 and RR05351.

References

[1] J.H. Prestegard, H.M. Al-Hashimi, J.R. Tolman, NMR structures of biomolecules using field oriented media and residual dipolar couplings, Quarterly Reviews of Biophysics 33 (4) (2000) 371–424.
[2] A. Bax, G. Kontaxis, N. Tjandra, Dipolar couplings in macromolecular structure determination, Nuclear Magnetic Resonance of Biological Macromolecules, Pt. B (2001) 127–174.
[3] F. Tian, H. Valafar, J.H. Prestegard, A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones, Journal of the American Chemical Society 123 (47) (2001) 11791–11796.
[4] M. Andrec, P.C. Du, R.M. Levy, Protein backbone structure determination using only residual dipolar couplings from one ordering medium, Journal of Biomolecular NMR 21 (4) (2001) 335–347.
[5] H. Valafar, J.H. Prestegard, Rapid classification to a protein fold family using a statistical analysis of dipolar couplings, Bioinformatics 19 (2003) 1–8.
[6] G. Cornilescu, F. Delaglio, A. Bax, Protein backbone angle restraints from searching a database for chemical shift and sequence homology, Journal of Biomolecular NMR 13 (3) (1999) 289–302.
[7] C.A. Fowler, F. Tian, H.M. Al-Hashimi, J.H. Prestegard, Rapid determination of protein folds using residual dipolar couplings, Journal of Molecular Biology 304 (3) (2000) 447–460.
[8] G.M. Clore, C.A. Bewley, Using conjoined rigid body/torsion angle simulated annealing to determine the relative orientation of covalently linked protein domains from dipolar couplings, Journal of Magnetic Resonance 154 (2) (2002) 329–335.
[9] I. Bertini, C. Luchinat, P. Turano, G. Battaini, L. Casella, The magnetic properties of myoglobin as studied by NMR spectroscopy, Chemistry—A European Journal 9 (10) (2003) 2316–2322.

[10] M. Assfalg, I. Bertini, P. Turano, A.G. Mauk, J.R. Winkler, H.B. Gray, N-15-H-1 residual dipolar coupling analysis of native and alkaline-K79A *Saccharomyces cerevisiae* cytochrome *c*, Biophysical Journal 84 (6) (2003) 3917–3923.
[11] N. Tjandra, S. Tate, A. Ono, M. Kainosho, A. Bax, The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase, Journal of the American Chemical Society 122 (26) (2000) 6190–6200.
[12] A. Vermeulen, H.J. Zhou, A. Pardi, Determining DNA global structure and DNA bending by application of NMR residual dipolar couplings, Journal of the American Chemical Society 122 (40) (2000) 9638–9647.
[13] F. Tian, H.M. Al-Hashimi, J.L. Craighead, J.H. Prestegard, Conformational analysis of a flexible oligosaccharide using residual dipolar couplings, Journal of the American Chemical Society 123 (3) (2001) 485–492.
[14] H.F. Azurmendi, C.A. Bush, Conformational studies of blood group A and blood group B oligosaccharides using NMR residual dipolar couplings, Carbohydrate Research 337 (10) (2002) 905–915.
[15] H.F. Azurmendi, M. Martin-Pastor, C.A. Bush, Conformational studies of Lewis X and Lewis A trisaccharides using NMR residual dipolar couplings, Biopolymers 63 (2) (2002) 89–98.
[16] N.U. Jain, S. Noble, J.H. Prestegard, Structural characterization of a mannose-binding protein–trimannoside complex using residual dipolar couplings, Journal of Molecular Biology 328 (2) (2003) 451–462.
[17] N.R. Skrynnikov, N.K. Goto, D.W. Yang, W.Y. Choy, J.R. Tolman, G.A. Mueller, L.E. Kay, Orienting domains in proteins using dipolar couplings measured by liquid-state NMR: differences in solution and crystal forms of maltodextrin binding protein loaded with beta-cyclodextrin, Journal of Molecular Biology 295 (5) (2000) 1265–1273.
[18] C.A. Bewley, Rapid validation of the overall structure of an internal domain-swapped mutant of the anti-HIV protein cyanovirin-N using residual dipolar couplings, Journal of the American Chemical Society 123 (5) (2001) 1014–1015.
[19] G.M. Clore, C.D. Schwieters, Docking of protein–protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from H-1(N)/N-15 chemical shift mapping and backbone N-15-H-1 residual dipolar couplings using conjoined rigid body/torsion angle dynamics, Journal of the American Chemical Society 125 (10) (2003) 2902–2912.
[20] J.C. Hus, R. Bruschweiler, Principal component method for assessing structural heterogeneity across multiple alignment media, Journal of Biomolecular NMR 24 (2) (2002) 123–132.
[21] M. Andrec, P.C. Du, R.M. Levy, Protein structural motif recognition via NMR residual dipolar couplings, Journal of the American Chemical Society 123 (6) (2001) 1222–1229.
[22] F. Delaglio, G. Kontaxis, A. Bax, Protein structure determination using molecular fragment replacement and NMR dipolar couplings, Journal of the American Chemical Society 122 (9) (2000) 2142–2143.
[23] C.A. Rohl, D. Baker, De novo determination of protein backbone structure from residual dipolar couplings using rosetta, Journal of the American Chemical Society 124 (11) (2002) 2723–2729.
[24] J. Meiler, N. Blomberg, M. Nilges, C. Griesinger, A new approach for applying residual dipolar couplings as restraints in structure elucidation, Journal of Biomolecular NMR 16 (3) (2000) 245–252.
[25] C.D. Schwieters, J.J. Kuszewski, N. Tjandra, G.M. Clore, The Xplor-NIH NMR molecular structure determination package, Journal of Magnetic Resonance 160 (1) (2003) 65–73.
[26] P. Dosset, J.C. Hus, D. Marion, M. Blackledge, A novel interactive tool for rigid-body modeling of multi-domain macromolecules using residual dipolar couplings, Journal of Biomolecular NMR 20 (3) (2001) 223–231.
[27] J.A. Losonczi, M. Andrec, M.W.F. Fischer, J.H. Prestegard, Order matrix analysis of residual dipolar couplings using singular value decomposition, Journal of Magnetic Resonance 138 (2) (1999) 334–342.
[28] J.H. Prestegard, A.I. Kishore, Partial alignment of biomolecules: an aid to NMR characterization, Current Opinion in Structural Biology 5 (5) (2001) 584–590.
[29] J.R. Tolman, H.M. Al-Hashimi, L.E. Kay, J.H. Prestegard, Structural and dynamic analysis of residual dipolar coupling data for proteins, Journal of the American Chemical Society 123 (7) (2001) 1416–1424.
[30] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, The Art of Scientific Computing, second ed., Cambridge University Press, Cambridge, MA, 2002.
[31] S.J. Varner, R.L. Vold, G.L. Hoatson, An efficient method for calculating powder patterns, Journal of Magnetic Resonance Series A 123 (1) (1996) 72–80.
[32] G.M. Clore, A.M. Gronenborn, A. Bax, A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information, Journal of Magnetic Resonance 133 (1) (1998) 216–221.
[33] M. Zweckstetter, A. Bax, Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR, Journal of the American Chemical Society 122 (15) (2000) 3791–3792.
[34] H.M. Al-Hashimi, P.J. Bolon, J.H. Prestegard, Molecular symmetry as an aid to geometry determination in ligand–protein complexes, Journal of Magnetic Resonance 142 (1) (2000) 153–158.
[35] P.J. Bolon, H.M. Al-Hashimi, J.H. Prestegard, Residual dipolar coupling derived orientational constraints on ligand geometry in a 53 kDa protein–ligand complex, Journal of Molecular Biology 293 (1) (1999) 107–115.
[36] W.Y. Choy, M. Tollinger, G.A. Mueller, L.E. Kay, Direct structure refinement of high molecular weight proteins against residual dipolar couplings and carbonyl chemical shift changes upon alignment: an application to maltose binding protein, Journal of Biomolecular NMR 21 (1) (2001) 31–40.
[37] F. Delaglio, S. Grzesiek, G.W. Vuister, G. Zhu, J. Pfeifer, A. Bax, NMRpipe—a multidimensional spectral processing system based on unix pipes, Journal of Biomolecular NMR 6 (3) (1995) 277–293.
[38] B.A. Johnson, R.A. Blevins, NMR view—a computer-program for the visualization and analysis of NMR data, Journal of Biomolecular NMR 4 (5) (1994) 603–614.