Classifying Protein Structural Dynamics via Residual Dipolar Couplings

Ruben Valas[*] **Christopher James Langmead**[*†]

December 2004
CMU-CS-05-108

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[*]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA 15213.
[†] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA 15213

E-mail: cjl@cs.cmu.edu

This research is supported by a Young Pioneer Award to C.J.L. from the Pittsburgh Lifesciences Greenhouse.

Keywords: computational biology, structural biology, Nuclear Magnetic Resonance, NMR, Residual Dipolar Couplings, RDCs, dynamics

Abstract

Recent advances in Nuclear Magnetic Resonance (NMR) spectroscopy present new opportunities for investigating the conformational dynamics of proteins in solution. In particular, tensors for motions relevant to biological function can be obtained via experimental measurement of residual dipolar couplings (RDCs) between nuclei. These motion tensors have been used by others to characterize the magnitude and anisotropy of the dynamics of individual bond vectors. Here, we extend these results and demonstrate that RDCs can *also* be used to characterize the *global* nature of the protein's motion (e.g., hinge motions, shear motions, etc.). In particular, we introduce the first method for classifying protein motions from RDC data. Our classifier consists of a discriminative model trained on 2,454 different molecular dynamics trajectories spanning seven categories of motion. The classifier achieves precision and recall accuracy of 90.6% and 90.9%, respectively, using 10-fold cross-validation over these seven categories.

1 Introduction

A protein's three dimensional structure can be dynamic. Structural dynamics often play an important role in the functioning of the protein; mobility is essential to many processes including muscle contraction, transport, and catalysis [8, 17]. Qualitative functional characterizations are often mediated by specific, coordinated motions of many atoms. Examples include molecular "trapdoors", such as triosephosphate isomerase (e.g., [7, 10, 11]), lactate dehydrogenase (e.g., [25]), and tyrosyl tRNA synthetase (e.g., [22]); and molecular "switches", such as thymidylate synthase (e.g., [27]) and hexokinase (e.g., [16]).

Macromolecular motions relevant to biological function can occur on many timescales up to and including the microsecond to millisecond range (e.g., [13]). Motions may be localized, involving a few residues, or global, involving the reorganization of multiple domains. Nuclear Magnetic Resonance (NMR) spectroscopy is well suited to the study of molecular motions because the relaxation processes which give rise to the data are sensitive to the internal dynamics of the protein. NMR dynamics studies are also capable of probing motions over an extraordinary range: fluctuations from the femtosecond to millisecond timescales can be measured.

Until recently, however, these experimental measurements could only be used to quantify the magnitude of an individual atom's motion; it was not possible, for example, to obtain direct evidence for a) the orientation of the motion or b) correlated movements between groups of atoms. Recent theoretical and experimental breakthroughs in NMR have addressed the first problem, but not the second. It is difficult to obtain a coherent model of the molecule's dynamics if correlations between the motions of the individual atoms are unknown. Molecular dynamics simulations are one means for obtaining dynamic models (e.g., [19]), but such simulations are usually limited to motions in the nanosecond timescale.

This paper introduces a novel method for characterizing a protein's global dynamics from NMR data over arbitrary timescales. We believe our technique has two important applications. First, by characterizing the global nature of the dynamics, our technique can be used to derive constraints for molecular dynamics simulations. These constraints could potentially be used to explore motions on timescales that are presently out of reach to modern simulations. Second, we believe our technique may assist biologists in characterizing the function of proteins in terms of both structure and dynamics.

The outline of this paper is as follows. In section 2, we present a brief review of experimental techniques for studying molecular motions. In section 3 we discuss the specific class of NMR data for which our method was designed. In section 4, we describe the data set used in our experiments. We present our method in Section 5, and the results of our experiments in Section 6. Finally, we discuss our results and future plans in Section 7.

2 Background

A protein that lacks a single, well defined structure is said to be disordered. Disorder may be local, involving only a few residues, or global, involving many different parts of the protein. Moreover, the disorder can be either static or dynamic. A protein that is statically disordered will have multiple, discrete conformations present within a population. A protein that is dynamically disordered inter-converts between different conformations through time. The size of the energy barrier separating the different configurations largely determines whether the disorder is static or dynamic. We are concerned with dynamically disordered proteins in this paper.

It is possible to detect and characterize disorder experimentally in a variety of ways. Broadly speaking, these techniques can be classified by whether they reveal aggregate or atom-specific dynamics. For example, techniques based on optical absorbtion report aggregate behaviors while techniques based on fluorescence are residue-specific. It is worth noting, however, that fluorescence studies are limited to those residues which can act as fluorophores (i.e., aromatic residues). In this paper, we are interested in experimental techniques capable of probing the dynamics at an atomic scale for (essentially) all atoms. Given this criterion, X-ray crystallography and NMR are obvious candidates.

In X-ray crystallography, static disorder will result in two (or more) alternative regions of electron density for the same group of atoms. Occupancy percentages within a Protein Data Bank (PDB) [4] file, for example, reflect the fraction of the molecules in a crystal for which the position specified in a given ATOM record is valid. Temperature factors (also known as B-factors), which are calculated during the refining process in X-ray crystallography, can also be used to identify disorder. However, it is not possible, generally speaking, to classify the disorder as either static or dynamic (or both) from temperature factors alone. A variety of measurements can be made to probe dynamics via NMR. These measurements are made in addition to those required for structure determination. Amide-exchange rates, spin-lattice and spin-spin relaxation rates, and the Nuclear Overhauser effect (which is a special case of spin-lattice relaxation), can all be used to characterize the motion of individual atoms within the structure.

We wish to emphasize three important considerations when studying macromolecular motions experimentally. The first is that motions can occur on a variety of timescales. Ideally, an experimental method should be able to characterize motions on many different timescales. As previously mentioned, NMR has this property. The second consideration is how much information is revealed about the motion of a given atom. Here, NMR, and other techniques have historically been limited. The mathematical formalisms for measuring dynamics from NMR data assume *isotropic* motion. That is, the motion of the atom is assumed to be uniform within a sphere. In essence, what one obtains from from the data is the size of the sphere. In reality, however, the motion may, in fact, be *anisotropic*. Here, a more accurate model might use an ellipse to describe the motion, and not a sphere. The third consideration is whether any direct evidence for coordinated motions between atoms can be obtained. NMR and other techniques are not able to do so. In practice, given the descriptions of the motions of individual atoms, one must either infer, or demonstrate via molecular dynamics simulations, that a group of atoms are moving in concert. As previously mentioned, molecular dynamics simulations are usually limited to very fast timescales, depending on the size of the molecule. Thus, the full range of motions cannot be explored adequately via

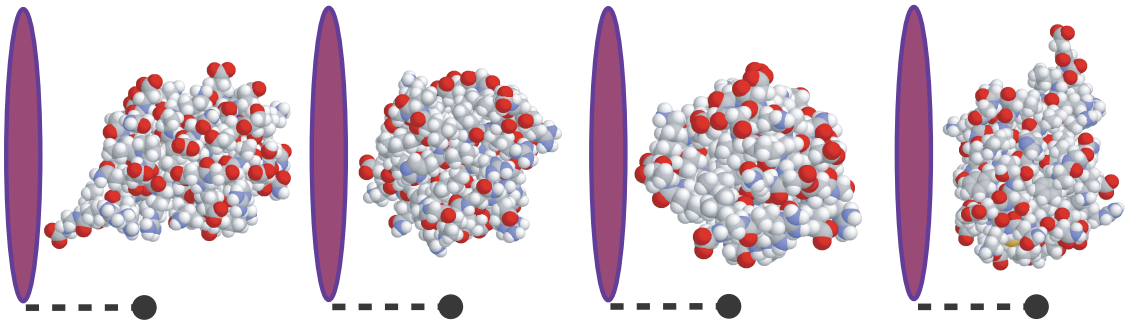

Figure 1: **Weak alignment in liquid crystalline phase**: Schematic of the interaction between an aligning agent, depicted as circular disks, and the protein ubiquitin. The surface of the aligning agent creates a steric barrier for the protein. The four panels show four possible orientations of the protein when the center of mass of the protein is at a fixed distance (depicted by the dashed line) from the surface of the aligning agent. As this distance decreases, the size of the set of possible orientations decreases. All NMR measurements are made over the entire ensemble of instances of the molecule. Thus, in an aligned medium, there is effectively an incomplete averaging over orientations, resulting in a dominant orientation of the molecule in solution.

molecular dynamics simulation. In summary, NMR is well suited to studying motions at a variety of timescales, but the measurements have traditionally been limited to the magnitudes of the motions of individual atoms.

3 Residual Dipolar Couplings

A *residual dipolar coupling* (RDC) is a measure of the dipolar interaction between pairs of atoms. The interaction is sensitive to the orientation of the vector between the two atoms, relative to the external magnetic field of the NMR spectrometer. Thus, RDCs carry important information about the orientation of inter-nuclear vectors. In solution NMR, RDCs are normally measured between atoms that are covalently bonded. Hence, RDCs are often used to determine the orientation of specific bond-vectors. The ability to record RDCs in solution has been perfected only recently [29, 30], but RDCs have already begun to play an important role in structure determination (e.g., [1, 15, 35, 36, 23]).

RDCs are only measurable in solution when the protein is in a so-called dilute crystalline form. Here, large molecules, often larger than the protein, are added to the solution. These agents align with the magnetic field of the spectrometer due to their own magnetic properties. The interactions between the aligning agent and the protein effectively reduce the number of possible orientations for the protein (Fig. 1) resulting in a "average" orientation to the protein. It is the presence of a dominant orientation that allows for RDCs to be measured.

The experimental measurements are scalar values. For each dipolar coupling, d, we have

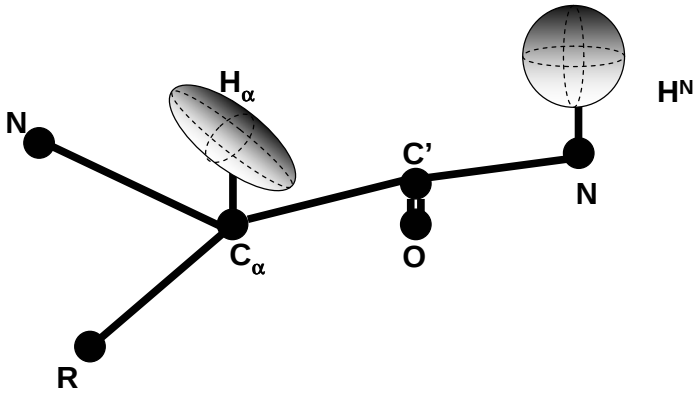

Figure 2: **Motion Tensors**: The motions of atoms can be characterized via RDCs. Here, the motions of H_α and H^N of adjacent residues are contrasted. 5 parameters describing the motion of each atom can be obtained (see text): S^2, η, α, β, and γ. S^2 corresponds to the magnitude of the motion. The radius of the sphere, and the length of the major axis of the ellipsoid are proportional to S^2. α and β are the polar coordinates describing the average orientation of the bond vector in 3D. η describes the anisotropy. Here, the H^N motion is isotropic (and thus depicted as a sphere) while the H_α is anisotropic (and thus depicted as an ellipsoid). When the motion is anisotropic, γ reveals the orientation of the movement. γ is proportional to the rotation about axis that goes through the bond vector.

$$d = -\frac{\mu_0 \gamma_a \gamma_b \hbar}{4\pi^2 r_{a,b}^3} \mathbf{v}_{ab}^T \mathbf{S} \mathbf{v}_{ab} \tag{1}$$

where γ_a and γ_b are the gyromagnetic ratios of spin $1/2$ nuclei a and b, $\hbar$ is Plank's constant, $r_{a,b}$ is the internuclear distance between a and b, μ_0 is free-space magnetic permeability, $\mathbf{v}_{ab}$ is the unit column vector between a and b relative to an *arbitrary* coordinate frame, and $\mathbf{S}$ is the *alignment tensor* [26]. In essence, $\mathbf{S}$ describes the overall orientation of the molecule. When resonance assignments and a three dimensional model of the protein are available, it is possible to extract $\mathbf{S}$ using Equation 1. Different aligning agents can induce different average orientations. Thus, for a given protein, there are different alignment tensors associated with each aligning agent.

In addition to their use in establishing the orientation of bond vectors, RDCs have also been used to study the dynamics of proteins (e.g., [6, 21, 32]). RDCs are sensitive to motions up to the millisecond timescale [19, 21, 33, 32], and are thus very useful. More recently, important contributions made by both Griessinger [19] and Tolman [31] have introduced methods for extracting complete motion tensors for each atom, from which *both* the magnitude and orientation of the motion can be measured (Fig. 2). This is a significant departure from traditional methods for extracting dynamics using NMR, which were limited to measuring magnitude alone. Following the notation of [31], the motion tensors for each atom can be computed as

$$\mathbf{D} = -\frac{\mu_0 \gamma_a \gamma_b \hbar}{8\pi^3 r_{a,b}^3} \mathbf{BA} \tag{2}$$

where **D** is an $n \times m$ matrix where n is the number of experimentally measured RDCs, and m is the number of independent aligning media. Thus, **D** is constructed from the experimental data. In order to extract the motion tensors it is necessary that $m \geq 5$. Matrices **A** and **B** are $5 \times M$ and $N \times 5$, respectively. The columns of **A** encode the alignment tensors (which have 5 degrees of freedom) for each of the m media. The rows of **B** contain the motion tensors for each atom. The motion tensors also have five degrees of freedom. From each motion tensor, 5 parameters of interest can be computed. The variables S_i^2, η_i, α_i, β_i, and γ_i are used to denote these 5 parameters for atom i. S_i^2 is the magnitude of atom i's motion; η_i is a measure of the anisotropy of atom i's motion; α_i and β_i are related to the polar coordinates of the bond vector expressed in the initial arbitrary reference frame (i.e., the PDB frame). If the motion of the atom is anisotropic (i.e., $\eta \neq 0$), the final parameter, γ_i measures the principal orientation of the motion.

Note that the RDC-derived motion parameters are *local* measurements. That is, it is not possible, mathematically, to measure any coordination of motion between bond vectors. In practice, molecular dynamics simulations are run to establish the global nature of the motion. However, these simulations are limited to very fast (nanosecond) timescales. In contrast, our method infers the global motion using statistical methods which are not, in principle, limited to any timescale. Thus, our method can make full use of the range of motions measurable via RDCs

Our justification for using statistical methods is based on the observation that the number of distinct protein folds is small, relative to the number of proteins. If one assumes that a protein's range of motion is determined by its fold, then one can argue that the number of motions is also limited. Our method makes the assumption that there are distinct classes of motion. Our method does *not*, however, assume that a given fold can only experience a single kind of motion.

4 Simulations

Our experiments were run on simulated data. The reasons for using simulated data are two-fold. First, our method currently requires 5 sets of independent, assigned RDCs (i.e., five different aligning media) and at this time, there is only one protein for which such data have been published. Our primary aim was to demonstrate that such a method might work on a large number of different proteins representing different fold families and different kinds of motion. We are presently exploring extensions to our work that would relax either the requirement for RDCs from 5 independent media, resonance assignments, or both. The second reason our experiments are limited to simulated data is that the one protein for which the required data are published [21], ubiquitin, is known to be almost entirely rigid [9, 28, 29], and thus does not exhibit motions of the sort explored in this paper. In this section, we describe our source for models of protein motions, and the means by which we simulated RDCs in 5 media.

The Gerstein lab has developed the Database of Macromolecular Movements [12]. This database contains dynamics trajectories for several hundred different proteins from many different fold families. The trajectories for each protein are often generated by interpolating between different ex-

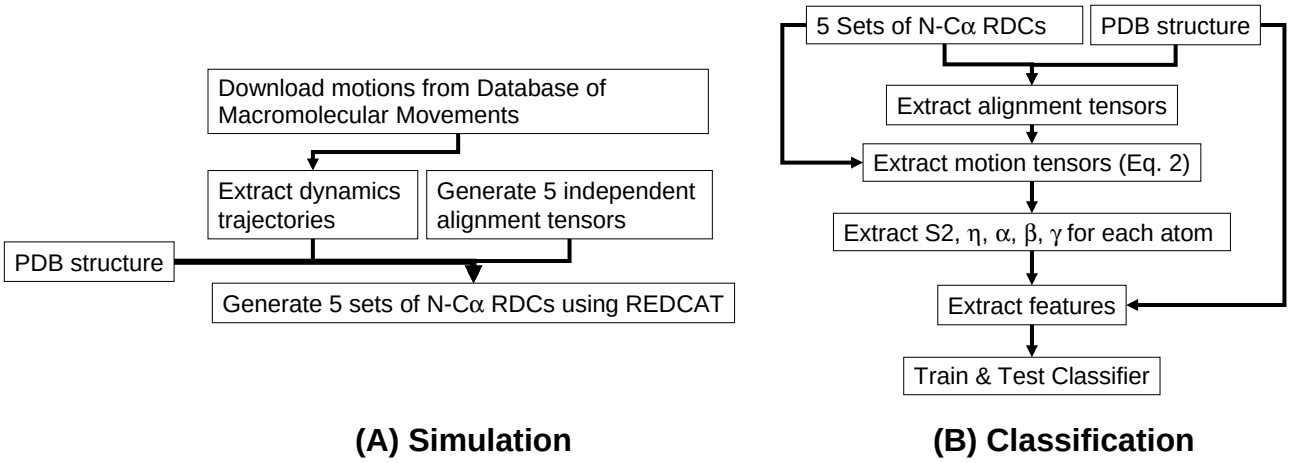

Figure 3: **Simulation and Classification.** (A) The flowchart for simulating the RDCs. (B) The flowchart for the classification step.

---

perimentally determined structural conformations of the same protein. Thus, these trajectories are not necessarily the result of actual molecular dynamics simulation, nor the result of experimental studies of dynamics. However, we feel that these trajectories are still representative of molecular motions, even if the details for any particular protein are not exact. In particular, the end-points of each trajectory are usually based on experimentally determined structures. Our method concerns the overall dynamics regime, and not the details of the motion.

In addition to serving as a repository of dynamics trajectories, the Database of Macromolecular Movements also defines a taxonomy of molecular motions. The different categories are largely defined by a) the size of the motion (e.g., fragment motions, domain motions), and b) a qualitative description (e.g., hinge motions, shear motions). For each protein in the database, there are often several different trajectories available. The trajectories are encoded as a set of ordered PDB files where each file reveals the configuration of the protein at a different point in "time". Our experiments were on 2,454 trajectories obtained from this database.

Five sets of RDCs were simulated for each of the 2,454 trajectories as follows (Fig 3 A). The program REDCAT [34] is capable of simulating the RDCs obtained under dynamics. REDCAT requires an alignment tensor, and a description of the dynamics of each atom, which we extract directly from the PDB files. For each trajectory, we randomly generated 5 independent alignment tensors. These alignment tensors, and the descriptions of the local dynamics were passed to REDCAT which generated the RDCs. We limited our experiments to RDCs for the backbone $N - C_\alpha$ bond. These simulated RDCs, and a single PDB file, representing the protein in a single configuration, are the inputs to our method, which is described in the next section.

5 Method

We treat the problem of characterizing a protein's global motion from RDCs as a classification problem (Fig 3 B). Using the taxonomy of motions defined by the Database of Macromolecular Movements, we separated the 2,454 trajectories into seven classes; 1) fragment motion, hinge; 2) fragment motion, shear; 3) domain motion, hinge; 4) domain motion, shear; 5) domain motion, partial refolding; 6) subunit motion, allosteric transition; 7) subunit motion; non-allosteric transition. Table 1 summarizes the number of trajectories falling into these seven classes. The terms "fragment", "domain", and "subunit" describe the size of the mobile element. Fragment motions can be as small as a single loop, while subunit motions are very large, potentially involving multiple domains. "Hinge", "shear", "partial refolding", and "allosteric" are qualitative descriptions of the motion, defined by the Database of Macromolecular Movements.

5.1 Features

Using Equation 2, it is possible to obtain 5 parameters describing the motion of each atom given a) a static model of the protein (e.g., a crystal structure), and b) the set of 5 assigned RDCs per atom. The five parameters, S^2, η, α, β, and γ, were discussed in Section 3. Our feature set consists of statistics of the distributions over these parameters. That is, the actual features are not atom-specific, but rather based on aggregates of atoms. This use of distributions is motivated by the need to handle "missing" data. Features based on distributions are more robust to these kinds of real-world concerns. Distribution-based features are also a natural way of dealing with proteins of different lengths, and therefore different numbers of atoms.

Our initial experiments were limited to passing in simple statistics (mean, variance, skewness, kurtosis) over distributions of the 5 parameters. These experiments were not successful. We then enriched our feature set substantially in three ways. First, we partitioned the backbone $N - C_\alpha$ bonds (i.e., the bonds for which we simulated RDCs) of each protein by amino acid category using the following types: non-polar, polar, acidic, basic. Thus, for each of these four residue types, we compute the mean, variance, skewness, and kurtosis for each of the 5 parameters. The motivation for this particular partitioning is that these categories are correlated with the expected location of the residue (surface, core, etc). Second, we computed a separate partitioning (independent of the first) based on which octant of the unit sphere each $N - C_\alpha$ bond vector is oriented[1]. That is, we bucketed the bonds by their α_i and β_i. The motivation for this partitioning is that it captures difference in relative motions in different orientations. Third, we considered the covariance of the RDCs among the different aligning media.

Our training set consisted of labeled examples. For each trajectory in the training set, the required features are extracted using the associated crystal structure and the assigned RDCs. The class label for each instance was one of the seven categories. We applied the discriminative learning technique of *Bagging* [5] to construct our classifier. Bagging (*Bootstrap Aggregating*) is an ensemble method of learning where the members of the ensemble are weak learners. Each of the weak learners is trained on a different bootstrap replicate (with replacement) of the training data. In

[1]The orientation of the bond vector is computed as going from the N to the C_α

Motion Scale	Motion Type	Number of trajectories	Precision (%)	Recall (%)
Fragment	Hinge	508	86.2	91.5
	Shear	146	90.5	78.3
subtotal		*654*	*88.1*	*89.6*
Domain	Hinge	646	95.7	92.1
	Shear	244	89.9	90.9
	Partial Refolding	241	81.6	84.9
subtotal		*1131*	*89.2*	*88.8*
Subunit	Allosteric Transition	656	97.4	97.6
	Non-Allosteric Transition	13	0	0
subtotal		*669*	*95.5*	*95.7*
Total		**2,454**	**90.6**	**90.9**

Table 1: **Results** A total of 2,454 molecular dynamics trajectories were used in our experiments. The table summarizes both the content of our training and test set and the accuracy of our method. For the purpose of comparison, the 2,454 trajectories are grouped by classes based on the size of the motion (column one) and the kind of motion (column two). Subtotals within each size subcategory are also given. Column three specifies the number of trajectories within a given class. Columns 4 and 5 report the precision and recall of our predictions, respectively. Precision and recall scores for subtotals are weighted averages. Note that our classifier makes predictions for all seven categories at once, and not just within a subcategory. The final row reports the overall scores over all seven categories and all 2,454 trajectories using 10-fold cross-validation.

the trained classifier, each of the weak learners is allowed to vote for the class for a given instance. The final prediction is the majority vote for classification tasks. In the language of statistical learning theory (e.g., [14]) and the bias-variance tradeoff, Bagging is a variance reduction technique. As such, it is a general method for resisting overfitting of training data. Intuitively, this can be understood by noticing that each of the weak learners is trained on a subset of the training set; thus, no learner can overfit the entire training set. Bagging is also very efficient and amenable to parallel implementations. We used a decision tree algorithm as our weak learner in our experiments.

6 Results

Using the 2,454 trajectories previously outlined, we evaluated the performance of our classifier using 10-fold cross-validation. The results of these experiments are presented in Table 1. The overall accuracies, in terms of precision and recall, were 90.6% and 90.9, respectively. Precision is the percentage of predictions that are correct. Recall is the percentage of the total number of instances of a given class that are correctly classified. Table 1 also presents accuracy statistics for various combinations of classes. The statistics reported for groups of classes are weighted by the number of instances in each class.

Small motions (fragment) are divided into two classes, hinge and shear. Our data contained nearly 3.5 times as many instances of hinge as shear. While fairly similar, the precision statistic for hinge was lower than that for shear. The recall statistic for hinge, however, is significantly higher than that for shear. Recall is essentially a measure of the fraction false-negative predictions; it may be that for very small, localized motions, a shearing motion is not as well characterized by our feature set as a hinge motion. It is also possible that the category itself is not well-defined.

Medium size motions (domain) are divided into three classes, hinge, shear, and partial refolding. Hinge motions are again much more common than either of the other two classes, but in this category, hinge motions are uniformly more accurate in terms of both precision and recall. However, the recall statistic on the shearing motion is now comparable to the recall statistic of the hinge motion suggesting that larger scale motions are more easily detected and classified. Considering only shear and partial refoldings, the method does better at predicting shear motions than predicting partial refoldings. One possible explanation for this tendency may be that refoldings within different proteins may simply involve a variety of different motions.

Large scale motions are divided into two classes, allosteric, and non-allosteric motions. Notice that the number of instances of non-allosteric motions is very small — just 13 instances. The classifier does very well predicting allosteric motions and terribly predicting non-allosteric motions. This behavior is almost certainly due to the small number of training instances.

While the preceding discussion has been in terms of motions of different scales, we stress that the classifier made its predictions over all classes simultaneously, and not merely within a single scale of motion. Thus, it is appropriate to consider the accuracies among the scales, and overall. The precision and recall statistics between small and medium scale motions are very similar, within 1.1% in all cases. By comparison, the precision and recall for the largest motions is about 6% higher. This also suggests that it may be easier to identify large scale motions than it is smaller motions. This seems reasonable; our feature set is largely based on statistics on distributions of bonds, and a large movement will involve the motions of a larger number of bonds than a small one. Finally, the overall accuracy across all seven categories is quite high, suggesting that our approach works well both across different scales of motion, and across qualitatively different kinds of motion.

7 Discussion and Conclusion

The results presented in this paper suggest that it may be possible to classify a protein's global dynamics regime using RDCs. This is of interest because the RDCs report on the local dynamics of individual atoms; correlated motions among atoms, and thus global motions, cannot be extracted directly from the data. However, we note that the actual range of motions experienced by any protein is likely to be limited by the structure of the protein itself. It is known that there are far fewer distinct protein folds than there are proteins. Thus, the range of motions is also likely to be small. This would suggest that it is possible to construct statistical models of the relationships between distributions of local motions and the global regime. Our results support this hypothesis.

We note that the method is very general; the proteins used in this study come from well over 100 CATH [20] categories. Furthermore, while the technique does consider distributions of amino

acids types (polar, non-polar, etc), it does not use sequence identity. Thus, it is unlikely that our training set has significant biases in terms of fold types or sequence composition. Additionally, the fact that the technique makes use of distributions of motion tensors suggests that it should be robust to real-world issues such as missing, or corrupted data.

Our results are promising, but very preliminary. There are several important areas for future work. The first is that the Database of Macromolecular Movements, while very useful, does not represent the entire range of motions. Recently, the iGNM database [2] has made available the normal mode analyses for all PDB entries. These normal mode analyses are computed using the Gaussian Network Model [3] of motion. We intend to apply our technique to this larger database of motions. Secondly, the classification scheme defined by the Database of Macromolecular Motions, is not necessarily the only that can be used; alternative schemes have been considered by other investigators (e.g., [18]). Next, it would be interesting to establish the effective resolution of the classification method in terms of the minimum size of detectable motions, as well as the number of differentiable categories of motion. Fourth, our results are limited in that the RDCs used in this study were simulated. We believe that the performance of the method should not degrade significantly on real data. As previously mentioned, our feature set considers distributions of experimental values, so it should be robust to noise and missing data. Furthermore, RDCs are, in fact, very accurate measurements. That is, in practice, the difference between the experimentally recorded and the theoretically computed values is usually small (e.g., [24]). Thus, a classifier trained on simulated data may not degrade significantly when applied to real data. We are working with our experimental collaborators to obtain the required data. Finally, the requirement for 5 independent media is the most restrictive experimental consideration. We are exploring the behavior of the technique when fewer than five media are available and the null-space of the tensors needs to be considered.

We believe that a method like ours, could potentially be used in two ways. First, motion tensors obtained via RDCs capture motions up to the millisecond range. The classifications made by our method could potentially be used to generate external constraints for molecular dynamics simulations. This opens the possibility for performing simulations over long timescales. Second, we feel that the predictions may be useful to biologists in interpreting existing functional data, and more generally, as an aid when investigating the relationships between structure, dynamics, and function.

8 Acknowledgements

The authors would like to thank Joel Tolman, Michael Erdmann, and Gordon Rule for their helpful comments. We would also like to thank Homayoun Valafar for his help with the REDCAT program. This work is supported by a grant to C.J.L. from the Pittsburgh Life Sciences Greenhouse. R.V. was supported by the Merck Summer Scholars Program.

References

[1] Andrec, M. and Du, P. and Levy, R.M. Protein backbone structure determination using only residual dipolar couplings from one ordering medium. *J Biomol NMR*, 21(4):335–347, 2001.

[2] I. Bahar. The iGNM database. http://ignm.ccbb.pitt.edu/, 2003.

[3] Bahar, I. and Atilgan, A. R. and Erman, B. . Direct evaluation of thermal fluctuations in protein using a single parameter harmonic potential. *Folding & Design*, 2:173–181, 1997.

[4] Berman, H.M and Westbrook, J. and Feng, Z. and Gilliland, G. and Bhat, T.N. and Weissig, H. and Shindyalov, I.N. and Bourne, P.E. The Protein Data Bank. *Nucl. Acids Res.*, 28:235–242, 2000.

[5] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[6] Briggman, K.B. and Tolman, J.R. De Novo Determination of Bond Orienations and Order Parameters from Residual Dipolar Couplings with High Accuracy. *J. Am. Chem. Soc.*, 125:10164–10165, 2003.

[7] Callender, R. and Dyer, R.B. Probing Protein Dynamics Using Temperature Jump Relaxation Spectroscopy. *Curr. Op. Struc. Biol.*, 12:628–633, 2002.

[8] Case, D.A. and Karplus, M. Dynamics of ligand binding to heme proteins. *J. Mol. Biol*, 132:343–368, 1979.

[9] de Alba, E. and Barber, J.L. and Tjandra, N. The Use of Residual Dipolar Coupling in Concert with Backbone Relaxation Rates to Identify Conformational Exchange by NMR. *J. Am. Chem. Soc.*, 121:4282–4283, 1999.

[10] Deng, H. and Zhadin, N. and Callender, R. The Dynamics of Protein Ligand Binding on Multiple Time Scales: NADH Binding to Lactate Dehydrogenase. *Biochemistry*, 40:3767–3773, 2001.

[11] Desamero, R. and Rozovsky, S. and Zhadin, N. and McDermott, A. and Callender, R. Active Site Loop Motion in Triosephosphate Isomerase: T-jump relaxation spectroscopy of thermal activation. *Biochemistry*, 42:2941–2951, 2003.

[12] Echols, E. and Milburn, D. and Gerstein, M. MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res*, 31:478–82, 2003.

[13] Feher, V.A. and Cavanagh, J. Millisecond-timescale motions contribute to the function of the bacterial response regulator protein Spo0F. *Nature*, 400:289–293, 1999.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[15] Hus, J.C. and Marion, D. and Blackledge, M. *De novo* Determination of Protein Structure by NMR using Orientational and Long-range Order Restraints. *J. Mol. Bio*, 298(5):927–936, 2000.

[16] Jacrot, B. and Cusack, S. and Dianoux, A.J. and Engelman, D.M. Inelastic neutron scattering analysis of hexokinase dynamics and its modification on binding of glucose. *Nature*, 3002:84–86, 1982.

[17] Kay, L.E. Protein dynamics from NMR. *Nat. Struct. Biol*, NMR Supplement:513–516, 1998.

[18] Kim, M.E. and Chirikjian, G.S. and Jernigan, R.L. Elastic Models of Conformational Transitions in Macromolecules. *J. Mol Graphics and Modelling*, 21:151–160, 2002.

[19] Meiler, J. and Prompers, J.J. and Peti, W. and Griesinger, C. and Bruschweiler, R. Model-Free Approach to the Dynamic Interpretation of Residual Dipolar Couplings in Globular Proteins. *J. Am. Chem. Soc.*, 123:6098–6107, 2001.

[20] Orengo, C.A. and Michie, A.D. and Jones, S. and Jones, D.T. and Swindells, M.B. and Thornton, J.M. CATH- A Hierarchic Classification of Protein Domain Structures. *Structure*, 5(8):1093–1108, 1997.

[21] Peti, W. and Meiler, J. and Bruschweiler, R. and Griesinger, C. Model-Free Analysis of Protein Backbone Motion from Residual Dipolar Couplings. *J. Am. Chem. Soc.*, 124:5822–5833, 2002.

[22] Qiu, X. and Janson, C. A. and Blackburn, M. N. and Chohan, I. K. and Hibbs, M. and Abdel-Meguid, S. S. Cooperative Structural Dynamics and a Novel Fidelity Mechanism in Histidyl-tRNA Synthetases. *Biochemistry*, 38:12296, 1999.

[23] Qu, Y. and Guo, J. and Olman, V. and Xu, Y. Protein fold recognition using residual dipolar coupling data. *Nucl. Acids Res.*, 32(2):551–561, 2004.

[24] Ramirez, B.E. and Bax, A. Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. *J. Am. Chem. Soc.*, 120:9106–9107, 1998.

[25] Sassaman, C. Dynamics of a lactate dehydrogenase polymorphism in the wood louse Porcellio scaber latr.: evidence for partial assortative mating and heterosis in natural populations. *Genetics*, 88(3):591–609, 1978.

[26] Saupe, A. Recent Results in the field of liquid crystals. *Angew. Chem.*, 7:97–112, 1968.

[27] Stroud, R.M. and Finer-Moore, J.S. Conformational dynamics along an enzymatic reaction pathway: thymidylate synthase, "the movie". *Biochemistry*, 42(2):239–247, 2003.

[28] Tjandra, N. and Feller, S.E. and Pastor, R.W. and Bax, A. Rotational diffusion anisotropy of human ubiquitin from ^{15}N NMR relaxation. *J. Am. Chem. Soc.*, 117:12562–12566, 1995.

[29] Tjandra, N. and Szabo, A. and Bax, A. Protein backbone dynamics and N-15 chemical shift anisotropy from quantitative measurement of relaxation interference effects. *J. Am. Chem. Soc.*, 118:6986–6991, 1996.

[30] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. Nuclear magnetic dipole interactions in field-oriented proteins:information for structure determination in solution. *Proc. Natl. Acad. Sci. USA*, 92:9279–9283, 1995.

[31] Tolman, J.R. A Novel Approach to the Retreival of Structural and Dynamic Information from Residual Dipolar Couplings Using Several Oriented Media in Biomolecular NMR Spectroscopy. *J. Am. Chem. Soc.*, 124:12020–12030, 2002.

[32] Tolman, J.R. and Al-Hashimi, H.M. and Kay, L.E. and Prestegard, J.H. Structural and dynamic analysis of residual dipolar coupling data for proteins. *J. Am. Chem. Soc.*, 123:1416–1424, 2001.

[33] Tolman, J.R. and Flanagan, J.M. and Kennedy, M.A. and Prestegard, J.H. NMR evidence for slow collective motions in cyanometmyoglobin. *Nat Struct. Biol.*, 4:292–297, 1997.

[34] Valafar, H. and Prestegard, J. REDCAT; a residual dipolar coupling analysis tool. *J. Magnetic Res*, 167:228–241, 2004.

[35] Wang, L. and Donald, B.R. Exact Solutions for Internuclear Vectors and Backbone Dihedral Angles from NH Residual Dipolar Couplings in Two Media, and Their Application in a Systematic Search Algorithm for Determining Protein Backbone Structure. *J. Biomol NMR*, 29(3):223–242, 2003.

[36] Wedemeyer, W. J. and Rohl, C. A. and Scheraga, H. A. Exact solutions for chemical bond orientations from residual dipolar couplings. *J. Biom. NMR*, 22:137–151, 2002.