JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Numbers 2–3, 2004

Pp. 277–298

# A Polynomial-Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments

CHRISTOPHER JAMES LANGMEAD,[1] ANTHONY YAN,[1] RYAN LILIEN,[1]
LINCONG WANG,[1] and BRUCE RANDALL DONALD[1,2,3]

## ABSTRACT

**High-throughput NMR structural biology can play an important role in structural genomics. We report an automated procedure for high-throughput NMR resonance assignment for a protein of known structure, or of a homologous structure. These assignments are a prerequisite for probing protein–protein interactions, protein–ligand binding, and dynamics by NMR. Assignments are also the starting point for structure determination and refinement. A new algorithm, called *Nuclear Vector Replacement (NVR)* is introduced to compute assignments that optimally correlate experimentally measured NH residual dipolar couplings (RDCs) to a given a priori whole-protein 3D structural model. The algorithm requires only uniform $^{15}$N-labeling of the protein and processes unassigned $H^N$-$^{15}$N HSQC spectra, $H^N$-$^{15}$N RDCs, and sparse $H^N$-$H^N$ NOE's ($d_{NN}$s), all of which can be acquired in a fraction of the time needed to record the traditional suite of experiments used to perform resonance assignments. NVR runs in minutes and efficiently assigns the ($H^N$,$^{15}$N) backbone resonances as well as the $d_{NN}$s of the 3D $^{15}$N-NOESY spectrum, in $O(n^3)$ time. The algorithm is demonstrated on NMR data from a 76-residue protein, human ubiquitin, matched to four structures, including one mutant (homolog), determined either by x-ray crystallography or by different NMR experiments (without RDCs). NVR achieves an assignment accuracy of 92–100%. We further demonstrate the feasibility of our algorithm for different and larger proteins, using NMR data for hen lysozyme (129 residues, 97–100% accuracy) and streptococcal protein G (56 residues, 100% accuracy), matched to a variety of 3D structural models. Finally, we extend NVR to a second application, 3D structural homology detection, and demonstrate that NVR is able to identify structural homologies between proteins with remote amino acid sequences using a database of structural models.**

**Key words:** automated NMR resonance assignments, structural homology detection, protein fold determination, nuclear vector replacement, residual dipolar couplings, structural genomics, molecular replacement.

[1]Computer Science Department, Dartmouth, Hanover, NH 03755.
[2]Chemistry Department, Dartmouth, Hanover, NH 03755.
[3]Department of Biological Sciences, Dartmouth, Hanover, NH 03755.

*Abbreviations used*: NMR, nuclear magnetic resonance; NVR, nuclear vector replacement; RDC, residual dipolar coupling; 3D, three-dimensional; HSQC, heteronuclear single-quantum coherence; $H^N$, amide proton; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; $d_{NN}$, nuclear Overhauser effect between two amide protons; MR, molecular replacement; SAR, structure activity relation; DOF, degrees of freedom; nt., nucleotides; SPG, Streptococcal protein G; $SO(3)$, special orthogonal (rotation) group in 3D.

## 1. INTRODUCTION

Current efforts in structural genomics are expected to determine experimentally many more protein structures, thereby populating the "space of protein structures" more densely. This large number of new structures should make techniques such as x-ray crystallography, molecular replacement (MR), and computational homology modeling more widely applicable for the determination of future structures. High-throughput NMR structural biology can play an equally important role in structural genomics. NMR techniques can determine solution-state structures (which are biochemically closer to physiological conditions than the crystalline state) and can be initiated immediately after protein purification, without resort to a lengthy search for high-quality crystals. NMR is ideally suited to probing and analyzing changes to the local electronic environments, yielding rapid, detailed studies of protein–protein and protein–ligand interactions and dynamics. A large fraction of the proteins of unknown function are NMR-accessible in terms of size and solubility. For these reasons, the NIH Protein Structure Initiative (NIH, 2002) has concentrated on both NMR and x-ray techniques as the paths to determine experimentally 10,000 new structures by 2010.

A key bottleneck in NMR structural biology is the resonance assignment problem. We seek to accelerate protein NMR resonance assignment and structure determination by exploiting a priori structural information. NMR assignments are valuable, even when the structure has already been determined by x-ray crystallography or computational homology modeling, because NMR can be used to probe protein–protein interactions (Fiaux *et al.*, 2002) (via chemical shift mapping [Chen *et al.*, 1993]), protein–ligand binding (via SAR by NMR [Shuker *et al.*, 1996] or line-broadening analysis [Fejzo *et al.*, 1999]), and dynamics (via, e.g., nuclear spin relaxation analysis [Palmer, 1997]). By analogy, in x-ray crystallography, the molecular replacement (MR) technique (Rossman and Blow, 1962) allows solution of the crystallographic phase problem when a "close" or homologous structural model is known a priori. It seems reasonable that knowing a structural model ahead of time could expedite resonance assignments. In the same way that MR attacks a critical informational bottleneck (phasing) in x-ray crystallography, an analogous technique for "MR by NMR" should address the NMR resonance assignment bottleneck. We propose a new RDC-based algorithm, called *Nuclear Vector Replacement (NVR)*, which computes assignments that correlate experimentally measured RDCs to a given a priori whole-protein 3D structural model. We believe this algorithm could form the basis for "MR by NMR."

NVR performs resonance assignment and structure refinement from a sparse set of NMR data. Performing resonance assignments given a structural model may be viewed as a combinatorial optimization problem—each assignment must match the experimental data, subject to the geometric and topological constraints of the known structure. Previous algorithms for solving the assignment problem using RDCs and a structural model (Al-Hashimi *et al.*, 2002; Hus *et al.*, 2002) require $^{13}C$-labeling and RDCs from many different internuclear vectors (for example, $^{13}C'$-$^{15}N$, $^{13}C'$-$H^N$, $^{13}C^{\alpha}$-$H^{\alpha}$, etc.) and more spectrometer time, and they are less efficient algorithms. In contrast, NVR requires only amide bond vector RDCs. Furthermore, NVR requires no triple-resonance experiments and uses only $^{15}N$-labeling, which is an order of magnitude less expensive than $^{13}C$-labeling. In NVR, the experimentally-measured internuclear bond vectors are conceptually "replaced" by model internuclear bond vectors to find the correct assignment. The NVR algorithm searches for the assignments that best correlate the experimental RDCs, $d_{NN}$s, and amide exchange rates with a whole-protein 3D structural model. NVR processes unassigned HSQC, $H^N$-$^{15}N$ RDCs (in two media), amide exchange data, and 3D $^{15}N$-NOESY spectra, all of which can be acquired in about one day using a cryoprobe.

NVR is demonstrated on NMR data from a 76-residue protein, human ubiquitin, matched to four structures determined either by X-ray crystallography or by *different* NMR experiments (without RDCs, and using a different NOESY spectrum than that processed by NVR), achieving an assignment accuracy of 92–100%. In other words, we did not fit the data to a model determined or refined by that same data.

TABLE 1. NVR EXPERIMENT SUITE: THE FIVE *Unassigned* NMR SPECTRA USED BY NVR TO PERFORM RESONANCE ASSIGNMENT AND STRUCTURE REFINEMENT[a]

| *Experiment/data* | *Information content* | *Role in NVR* | *Acquisition time* |
|---|---|---|---|
| $H^N$-$^{15}N$ HSQC | $H^N$,$^{15}N$ Chemical shifts | Backbone resonances, cross-referencing NOESY | 1/2 hr. |
| $H^N$-$^{15}N$ RDC (in 2 media) | Restraints on amide internuclear vector orientation | Tensor estimation, resonance assignment, structure refinement | 1/2 hr. + 1/2 hr. |
| H-D exchange HSQC | Identifies solvent exposed amide protons | Resonance assignment | 1/2 hr. |
| $H^N$-$^{15}N$ HSQC-NOESY | Distance restraints between spin systems | Resonance assignment | 12 hrs. |
| Structural model of backbone | Tertiary structure | Tensor estimation, resonance assignment, structure refinement | Assumed given |

[a]The HSQC provides the backbone resonances to be assigned. The two $H^N$-$^{15}N$ RDC spectra (which are modified HSQCs) provide independent, global restraints on the orientation of each backbone amide bond vector. The H-D exchange HSQC identifies fast exchanging amide protons. These amide protons are likely to be solvent-exposed and nonhydrogen bonded and can be correlated to the structural model. A sparse number ($< 1$ per residue) of $d_{\mathrm{NN}}$s can be obtained from the NOESY. These $d_{\mathrm{NN}}$s provide distance constraints between spin systems which can be correlated to the structural model. The data acquisition times are estimated assuming the spectrometer is equipped with a cryoprobe. Additional set-up time may be needed for each experiment.

Instead, we tested NVR using structural models that were derived using either (a) different techniques (x-ray crystallography) or (b) different NMR data. We further demonstrate the feasibility of our algorithm for different and larger proteins, using NMR data for hen lysozyme (129 residues) and streptococcal protein G (56 residues), matched to 16 different 3D structural models. Finally, when an homologous structure is employed as the model, it is straightforward to perform structure refinement after NVR. For this purpose one uses the assigned RDCs to facilitate rapid structure determination.

This paper reports an earlier version of the method presented by Langmead and Donald (2004a). The current paper emphasizes the computer science aspects of NVR, uses a different algorithm from Langmead and Donald (2004a), and develops a rotation search that is used in structural homology detection from unassigned NMR data (further developed by Langmead and Donald, 2003).

### *1.1. Organization of paper*

We begin, in Section 2, with a review of the specific NMR experiments used in our method, highlighting their information content. Section 3 describes existing techniques for resonance assignment from RDC data, including a discussion of their limitations and computational complexity. In Section 4, we detail our method and analyze its computational complexity. Section 5 presents the results of applying our method on real biological NMR data. Section 5.1 extends some of the key techniques in NVR to a new application, 3D structural homology detection. Finally, Section 6 discusses these results.

## 2. BACKGROUND

The experimental inputs to NVR are detailed in Table 1. Residual dipolar couplings (RDCs) (Tjandra and Bax, 1997) provide *global* orientational restraints on internuclear vectors[1] (these global restraints are often termed *"long-range"* in the literature). For good introductions to RDCs see Saupe (1968), Losonczi *et al.* (1999), and Tjandra and Bax (1997). For each RDC $D$, we have

$$D = D_{\max}\mathbf{v}^T\mathbf{S}\mathbf{v}, \tag{1}$$

[1]Often, these internuclear vectors are bond vectors (e.g., $^{15}N$-$^{1}H$).

where $D_{\max}$ is the dipolar interaction constant, $\mathbf{v}$ is the internuclear vector orientation relative to an arbitrary substructure frame, and $\mathbf{S}$ is the $3 \times 3$ *Saupe* order matrix, or alignment tensor specifying the orientation of the molecule in the laboratory frame (Saupe, 1968). Tensor $\mathbf{S}$ is a symmetric, traceless, rank-two tensor with five degrees of freedom, which describes the average substructure alignment in the dilute liquid crystalline phase (Losonczi *et al.*, 1999). The measurement of five or more RDCs in substructures of known geometry allows determination of $\mathbf{S}$. Furthermore, using Equation (1), substructures of the protein may be oriented relative to a common coordinate system, the *principal order frame*.

Once $\mathbf{S}$ is estimated, RDCs may be simulated (back-calculated) given any other internuclear vector $\mathbf{v}_i$. In particular, suppose an ($H^N$,$^{15}N$) peak $i$ in an $H^N$-$^{15}N$ HSQC (subsequently termed simply "HSQC") spectrum is assigned to residue $j$ of a protein, whose crystal structure is known. Let $D_i$ be the measured RDC value corresponding to this peak. Then the RDC $D_i$ is assigned to amide bond vector $\mathbf{v}_j$ of a known structure, and we should expect that $D_i \approx D_{\max} \mathbf{v}_j^T \mathbf{S} \mathbf{v}_j$ (modulo noise, dynamics, crystal contacts in the structural model, etc).

Experimentally recorded RDCs often deviate from their predicted values. These deviations can be large or small and may be the result of dynamics, discrepancies between the idealized physics and the conditions in solution, and, when the model structure is derived from crystallography, crystal contacts and conformational differences between the protein in solution versus in the crystalline state. It is reasonable, in principle, to cast the problem of resonance assignment of a known structure using RDCs into a combinatorial optimization framework (Hus *et al.*, 2002); given an estimate for the alignment tensor, a weighted bipartite graph can be constructed between the resonance peaks in the spectrum and the amino acids in the primary sequence of the protein. Unfortunately, maximum bipartite matching is sensitive to outliers. In a set of preliminary experiments on bipartite graphs whose edge weights encode the difference between experimentally observed RDCs and back-computed RDCs, the matching that minimizes these weights is typically not the correct matching. Maximum bipartite matchings contained, on average, only 25% correct assignments, and no higher than 43% (as will be shown later in Tables 2, 3, 4). We conclude that $^1H$-$^{15}N$ RDCs alone do not have enough constraint to perform backbone resonance assignments. NVR incorporates the additional, independent geometric constraints contained in amide exchange rates and NOEs.

NOE distance restraints are extracted from the $d_{\mathrm{NN}}$ region of an unassigned $^{15}N$ HSQC-NOESY. NVR uses a *sparse* set of NOEs. By sparse, we mean a small number of unassigned NOEs. A sparse set of $d_{\mathrm{NN}}$s can be obtained from an unassigned NOESY spectrum, after it is referenced to the $^{15}N$-HSQC spectrum. In our trials on ubiquitin, for example, we obtained 34 $d_{\mathrm{NN}}$s, from an unassigned 3D $^{15}N$-NOESY spectrum (Harris, 2002). This amounts to fewer than 0.5 $d_{\mathrm{NN}}$s per residue on average. In contrast, when solving a protein structure using NMR, it is not uncommon to have 10–15 or more *assigned* NOEs per residue. In NVR, $d_{\mathrm{NN}}$s are interpreted as geometric constraints, as follows: If a particular spin system $i$ has a $d_{\mathrm{NN}}$ with spin system $j$ and $i$ is assigned to a particular residue $r$, then $j$'s possible assignments are constrained to the set of residues that are within 6 Å of $r$ in the model. Similarly, HSQC peaks that exchange rapidly with the solvent, as identified by amide exchange experiments, are constrained to be assigned to nonhydrogen bonded surface amide protons in the model.

## 3. PRIOR WORK

*Assigned* RDCs have previously been employed by a variety of structure refinement (Chou *et al.*, 2000) and structure determination methods (Hus *et al.*, 2000; Andrec *et al.*, 2001; Wedemeyer *et al.*, 2002), including orientation and placement of secondary structure to determine protein folds (Fowler *et al.*, 2000), pruning an homologous structural database (Annila *et al.*, 1999; Meiler *et al.*, 2000), *de novo* structure determination (Rohl and Baker, 2002), in combination with a sparse set of assigned NOE's to determine the global fold (Mueller *et al.*, 2000), and a method developed by Bax and coworkers for fold determination that selects heptapeptide fragments best fitting the assigned RDC data (Delaglio *et al.*, 2000). Bax and coworkers termed their technique "molecular fragment replacement," by analogy with x-ray crystallography MR techniques.

In contrast, our algorithm processes *unassigned* RDCs. Unassigned RDCs have been used to expedite resonance assignments. Chemical shift degeneracies (particularly $^{13}C$-resonance overlap) in triple-resonance through-bond correlation spectra can lead to ambiguity in determining the sequential neighbors of a residue.

RDC contributions have been shown to overcome these limitations (Zweckstetter and Bax, 2001; Delaglio *et al.*, 2000). In another study, RDCs were used by Prestegard and coworkers (Tian *et al.*, 2001) to prune the set of potential sequential neighbors indicated by a degenerate HNCA spectrum, yielding an algorithm for simultaneous resonance assignment and fold determination. These methods (except Tian *et al.*, 2001) require $^{13}$C-labelling and RDCs from many different internuclear vectors (for example, $^{13}C'$-$^{15}N$, $^{13}C'$-$H^N$, $^{13}C^{\alpha}$-$H^{\alpha}$, etc.). The CAP method for small RNA assignment (Al-Hashimi *et al.*, 2002) also requires $^{13}$C-labelling and many RDCs in addition to many through-bond, triple-resonance experiments. Brüschweiler and coworkers (Hus *et al.*, 2002) have reported a method for resonance assignment (which we eponymously term *HPB*) that uses RDCs to assign a protein of known structure. The HPB method iteratively solves for both the alignment tensor **S** and the resonance assignments. It requires several RDCs per residue and the recording of two $^{13}$C triple resonance experiments. Our method addresses the same problem as HPB, but uses a different algorithm and requires only amide bond vector RDCs, no triple-resonance experiments, and no $^{13}$C-labeling (cf. Wüthrich (2000): "A big asset with regard to future practical applications ... [is] ... straightforward, inexpensive experimentation. This applies to the isotope labelling scheme as well as to the NMR spectroscopy..."). In general, $^{13}$C-labeling is necessary both for triple resonance experiments and to measure two-bond $^{13}C'$-$^{1}$H and one-bond $^{13}C'$-$^{15}$N dipolar coupling constants. Of previous efforts in structure-based assignment, only one group has tried to minimize the cost of isotopic labeling: Prestegard and coworkers (Tian *et al.*, 2001) probed a rubredoxin protein that was small enough (54 residues) and soluble enough (4.5 mM) to explore using $^{15}$N enrichment, but with $^{13}$C at natural abundance. We note that NVR both adopts a "best-first" strategy and uses structural homology to make assignments; best-first and homology-based strategies for disambiguating assignments are well-established techniques (e.g., Hoch *et al.*, 1990; Redfield *et al.*, 1983).

From a computational standpoint, NVR adopts a minimalist approach (Bailey-Kellogg *et al.*, 2000), demonstrating the large amount of information available in a few key spectra. By eliminating the need for triple resonance experiments, NVR saves days of spectrometer time. The NVR protocol also confers advantages in terms of computational efficiency. The combinatorial complexity of the assignment problem is a function of the number $n$ of residues (or bases in a nucleic acid) to be assigned and the spectral complexity (degree of degeneracy and overlap in frequency space). For example, CAP (Al-Hashimi *et al.*, 2002) has been applied with $n = 27$ nt., and the time complexity of CAP grows exponentially with $n$. In particular, CAP performs an exhaustive search, making it difficult to scale up to larger RNAs. HPB runs in time $O(In^3)$, where $O(n^3)$ is the complexity of bipartite matching (Kuhn, 1955) and $I$ is the number of times that the Kuhn–Munkres matching algorithm is called. Hus *et al.* (2002) do not bound $I$ or prove convergence of HPB (i.e., how many times $I$ will the bipartite matching algorithm be called before HPB terminates). However, $I$ may be bounded by $O(k^3)$, the size of the discrete grid search for the principal order frame over $SO(3)$ (using Euler angles $\alpha$, $\beta$, and $\gamma$). Here, $k$ is the resolution of the grid. Thus, the full complexity of HPB is $O(k^3n^3)$. Our algorithm is combinatorially efficient, runs in minutes, and is guaranteed to converge in $O(nk^3 + n^3)$ time, scaling easily to proteins in the middle NMR size range ($n = 56$ to 129 residues).

## 4. NUCLEAR VECTOR REPLACEMENT

The NVR method has three stages: *tensor estimation*, *resonance assignment*, and *structure refinement* (Fig. 1). In the first stage, the alignment tensors for each aligning medium are estimated. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be the estimated tensors for the phage and bicelle media, respectively. These tensors correspond to the matrix **S** in Equation (1). Macromolecules align differently in different liquid crystals; thus, $\mathbf{S}_1$ and $\mathbf{S}_2$ are different matrices. Matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are used to bootstrap stage two. The output of stage two is the resonance assignments. These assignments, and the geometric constraints imposed from the RDCs, are used to refine the structural model in stage three.

### *4.1. Tensor estimation (Phase 1)*

An alignment tensor is a symmetric and traceless $3 \times 3$ matrix with five degrees of freedom. The five degrees of freedom correspond to three Euler angles ($\alpha$, $\beta$, and $\gamma$), describing the average partial

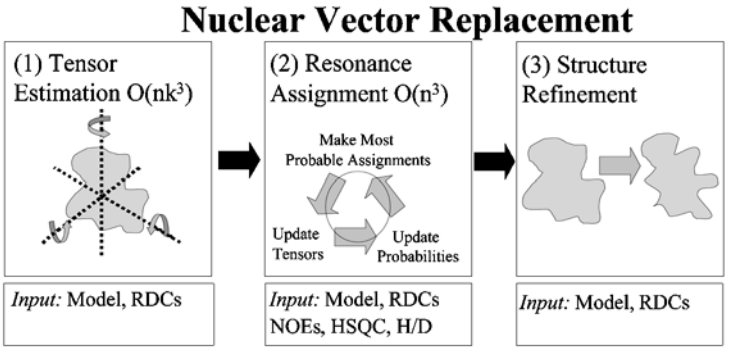


**FIG. 1.** Nuclear vector replacement. Schematic of the NVR algorithm for resonance assignment. The NVR algorithm takes as input a model of the target protein and several unassigned spectra, including the $^{15}$N-HSQC, $H^N$-$^{15}$N RDC, and $^{15}$N-HSQC NOESY and an H-D exchange-HSQC to measure amide exchange rates. In the first stage, NVR estimates the alignment tensors for both media. This step takes time $O(nk^3)$, where $n$ is the number of residues and $k$ is the resolution of the search grid. In the second phase, the estimated tensors are used to bootstrap an iterative process wherein the resonance assignments are computed using a Bayesian framework. This entire process runs in minutes and is guaranteed to converge in time $O(n^3)$. In the final phase, the model structure is refined using the residue-specific geometric constraints imposed by the RDCs (which were assigned in phase 2). When complete, NVR outputs both a refined structure and a set of resonance assignments.

alignment of the protein and the axial ($D_a$) and rhombic ($D_r$) components of an ellipsoid that scales the dipolar couplings. When resonance assignments and the structure of the macromolecule are known, all five parameters can be computed by solving a system of linear equations (Losonczi *et al.*, 1999). If the resonance assignments are not known, as in our case, these parameters must be estimated. It has been shown (Losonczi *et al.*, 1999) that $D_a$ and $D_r$ can be decoupled from the Euler angles by diagonalizing the alignment tensor:

$$\mathbf{S} = \mathbf{V}\Sigma\mathbf{V}^T. \tag{2}$$

here, $\mathbf{V} \in SO(3)$ is a $3 \times 3$ rotation matrix[2] that defines a coordinate system called the *principal order frame.* Matrix $\Sigma$ is a $3 \times 3$ diagonal and traceless matrix containing the eigenvalues of $\mathbf{S}$. The diagonal elements of $\Sigma$ encode $D_a$ and $D_r$: $D_a = \frac{S_{zz}}{2}$, $D_r = \frac{S_{xx}-S_{yy}}{3}$ where $S_{yy} < S_{xx} < S_{zz}$. Elements $S_{yy}$, $S_{xx}$, and $S_{zz}$ are the diagonal elements of $\Sigma$ and therefore the eigenvalues of $\mathbf{S}$. It has been shown that $D_a$ and $D_r$ can be estimated, using only unassigned experimentally recorded RDCs, by the powder pattern method (Wedemeyer *et al.*, 2002). The axial and rhombic components of the tensor can be computed in time $O(nk^2)$ (Fig. 2), where $n$ is the number of observed RDCs and $k$ is the resolution of the search-grid over $D_a$ and $D_r$.

Once the axial and rhombic components have been estimated, matrix $\Sigma$ in Equation (2) can be constructed using the relationship (Losonczi *et al.*, 1999; Wedemeyer *et al.*, 2002) between the $D_a$ and $D_r$ and the diagonal elements of $\Sigma$. Next, the Euler angles $\alpha$, $\beta$, and $\gamma$ of the principal order frame are estimated by considering rotations of the model. Given $\Sigma$ (Equation 2) for each rotation $V(\alpha, \beta, \gamma)$ of the model, a new Saupe matrix $\mathbf{S}$ is computed using Equation (2). That matrix $\mathbf{S}$ is used to compute a set of back-computed RDCs using Equation (1). The relative entropy, also known as the Kullback–Leibler distance (1951), is computed between the histogram of the observed RDCs and the histogram of the back-computed RDCs. The rotation of the model that minimizes the relative entropy is chosen as the initial estimate for the Euler angles. The comparison of distributions to evaluate Euler angles is conceptually related to the premise used by the powder pattern method (Wedemeyer *et al.*, 2002) to estimate the axial and rhombic components of the tensor. In the powder pattern method, the observed RDCs are implicitly compared to a distribution of RDCs generated by a uniform distribution of internuclear vectors. When estimating the Euler angles, NVR explicitly compares the distributions using a relative entropy measure. Intuitively, the correct rotation of the

[2]While any representation of rotations may be employed, we use Euler angles $(\alpha, \beta, \gamma)$.

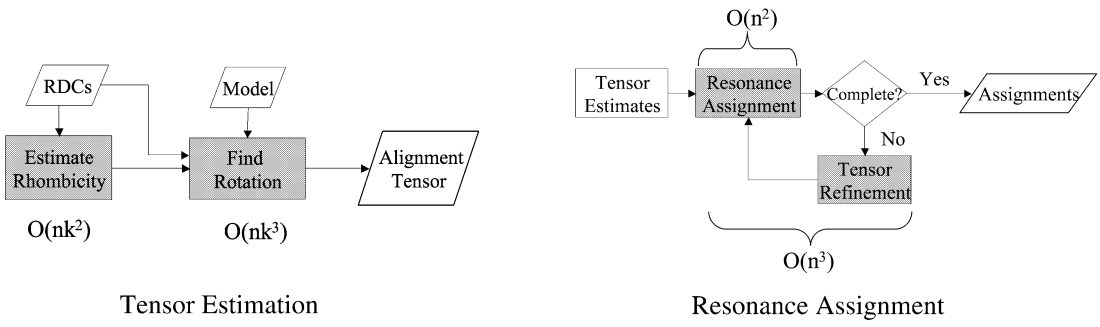


**FIG. 2.** Tensor estimation and resonance assignment. (**Left**) Tensor estimation: The NVR method estimates the alignment tensor for a given aligning medium in two steps. First, $D_a$ and $D_r$ are computed using the powder pattern method. Next, the best rotation of the model is computed using the estimated $D_a$ and $D_r$. This can be computed in $O(nk^3)$ time (see text). (**Right**) Resonance assignment: NVR computes resonance assignments using an iterative algorithm. Before the iteration begins, geometric constraints are extracted from the $^{15}$N-HSQC NOESY and H-D exchange HSQC and correlated to the model structure and the peaks in the HSQC. The initial tensor estimates bootstrap the iterative process. During each iteration, the probability of each remaining (resonance $\mapsto$ residue) assignment is (re)computed using the model, the tensors, and the RDCs. The most probable assignments are made, and the tensor estimates are refined at the end of each iteration (see Fig. 1). This process takes $O(n^2)$ time, where $n$ is the number of resonances. At least one residue is assigned each iteration. Thus, the entire protein is assigned in $O(n^3)$ time.

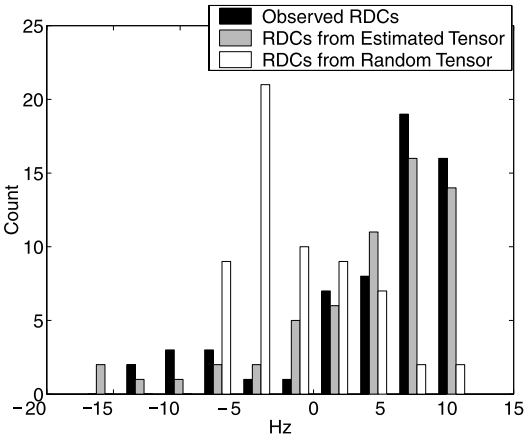


**FIG. 3.** Distributions of dipolar couplings. A comparison of the distributions of dipolar couplings generated from three different alignment tensors. The black bars are the distribution of observed RDCs for human ubiquitin in the bicelle medium. The grey bars are the distribution of RDCs generated by the tensor estimated by NVR using 1UBI as a model. The black and gray distributions are quite similar. The white bars are the distribution of RDCs from a random tensor. The white distribution is quite different from the black and gray distributions.

model will generate a distribution of RDCs that is similar to the unassigned distribution of experimentally measured RDCs (Fig. 3). The rotation minimizing the Kullback–Leibler distance can be computed exactly in polynomial time using the first-order theory of real-closed fields (see Appendix A); in practice, we implemented a discrete grid search. This rotation search (Fig. 2) takes $O(nk^3)$ time for $n$ residues on a $k \times k \times k$ grid. Thus, we can estimate alignment tensors in $O(nk^3)$ time. In practice, it takes NVR a few minutes to estimate the alignment tensors.

### *4.2. Resonance assignment (Phase 2)*

The input to phase 2 (Fig. 2) includes the two order matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ computed in phase 1. Each order matrix is used to compute a set of expected RDCs from the model using Equation (1). Let $Q$ be the set of HSQC peaks, $R$ be the set of residues in the protein, $D_m$ be the set of observed RDCs in medium $m$,

Table 2. Ubiquitin: Comparison of Assignment Algorithms[a]

| | *Accuracy* | | |
|---|---|---|---|
| *PDB ID* | *Maximum bipartite matching* | *NVR with RDC and amide exchange* | *NVR with RDC and NOE* |
| 1G6J[b] | 7% | 37% | 72% |
| 1UBI[c] | 25% | 65% | 73% |
| 1UBQ[d] | 40% | 42% | 85% |
| 1UD7[e] | 28% | 18% | 65% |

[a]The first column reports the accuracy of a maximum bipartite matching of a graph whose edge weights are the total distance between observed and back-calculated RDCs under both media. The maximum bipartite matching algorithm returns the matching that minimizes the total distance. Columns 2 and 3 are the results of running NVR with the alignment tensors it estimates using RDCs with amide exchange constraints and NOE constraints individually. The accuracies are far lower than those reported in Table 8.

[b]Babu *et al.*, 2001.

[c]Ramage *et al.*, 1994.

[d]Vijay-Kumar *et al.*, 1987.

[e]Johnson *et al.*, 1999.

and $B_m$ be the set of back-computed RDCs using the model and $\mathbf{S}_m$. For each medium $m$, an $n$-peak $\times$ $n$-residue probability matrix $\mathbf{M}_m$ is constructed. The rows of $\mathbf{M}_m$ correspond to some fixed ordering of the peaks in the HSQC. Similarly, the columns of $\mathbf{M}_m$ correspond to some fixed ordering of the residues in the protein. The assignment probabilities are computed as follows:

$$\mathbf{M}_m(q, r) = \mathbf{P}(q \mapsto r | S_m) = N(d_m(q) - b_m(r, S_m), \mu_m, \sigma_m) \tag{3}$$

where $q \in Q$ and $r \in R$, $d_m(q) \in D_m$, $b_m(r, S_m) \in B_m$. The function $N(d_m(q) - b(r, S_m), \mu_m, \sigma_m)$ is the probability of observing the difference $d_m(q) - b(r, S_m)$ in a normal distribution with mean $\mu_m$ and standard deviation $\sigma_m$. We used $\mu_m = 0$ Hz and $\sigma_m = 1$ Hz in all our trials. Intuitively, $\mathbf{M}_m(q, r)$ is the probability that peak $q$ is assigned to residue $r$ in medium $m$. An individual entry of $\mathbf{M}_m$ may be set to zero if the assignment $q \mapsto r$ violates a geometric constraint imposed by a $d_{\mathrm{NN}}$ or amide exchange.

On each iteration, the probabilities of assignment are (re)computed using Equation (3). For each row in $\mathbf{M}_1$ and $\mathbf{M}_2$, the most likely assignment is considered. Let $r_1(q) \in R$ and $r_2(q) \in R$ be the most likely resonance assignment for peak $q$ in media 1 and 2, respectively. The assignment $q \mapsto r$ is added to the master list of assignments if $r_1(q) = r_2(q)$ and the following condition is met:

$$r_m(q) \neq r_m(k), \qquad m = 1, 2; \ \forall k \in Q, \ k \neq q. \tag{4}$$

When an assignment is made, peak $q$ and residue $r$ are removed from consideration in subsequent iterations. Thus, the size of matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ decreases with each iteration. At the end of each iteration, alignment tensors $\mathbf{S}_1$ and $\mathbf{S}_2$ are refined by using the master list of assignments and the model, by means of the SVD method (Losonczi *et al.*, 1999). The tensors, which were coarsely estimated in phase 1 of NVR, begin to converge to their true values with each iteration.[3] At the end of phase 2, the principal axes of the final tensor estimates are typically within 3°, and the axial and rhombic components are within 1% of their correct values, respectively (Tables 2, 3, 4).

Intuitively, NVR only makes assignments that are a) unambiguous and b) consistent across both media. Figure 4 shows an example of the first few iterations of NVR on NMR data for human ubiquitin using 1UBQ as a model structure. The probabilistic nature of NVR means that it is straightforward to generate

[3]For the purposes of comparison and to quantitate the accuracy of NVR, "true" values of the alignment tensors were determined by (a) published values in the literature (Cornilescu *et al.*, 1998; Kuszewski *et al.*, 1999; Schwalbe *et al.*, 2001) and/or (b) computing the optimal Saupe matrix using the correct assignments.

TABLE 3. SPG: COMPARISON OF ASSIGNMENT ALGORITHMS[a]

| | *Accuracy* | | |
|---|---|---|---|
| *PDB ID* | *Maximum bipartite matching* | *NVR with RDC and amide exchange* | *NVR with RDC and NOE* |
| 1GB1[b] | 18% | 45% | 95% |
| 2GB1[b] | 43% | 48% | 95% |
| 1PGB[c] | 9% | 48% | 63% |

[a]See Table 2 for a description of each column.
[b]Gronenborn *et al.*, 1991.
[c]Gallagher *et al.*, 1994.

TABLE 4. LYSOZYME COMPARISON OF ASSIGNMENT ALGORITHMS

| | *Accuracy* | | |
|---|---|---|---|
| *PDB ID* | *Maximum bipartite matching* | *NVR with RDC and amide exchange* | *NVR with RDC and NOE* |
| 193L[b] | 24% | 23% | 93% |
| 1AKI[c] | 9% | 56% | 83% |
| 1AZF[d] | 15% | 19% | 84% |
| 1BGI[e] | 18% | 51% | 98% |
| 1H87[f] | 16% | 13% | 95% |
| 1LSC[g] | 23% | 17% | 94% |
| 1LSE[g] | 9% | 29% | 92% |
| 1LYZ[h] | 2% | 15% | 54% |
| 2LYZ[h] | 13% | 9% | 77% |
| 3LYZ[h] | 8% | 28% | 97% |
| 4LYZ[h] | 13% | 33% | 86% |
| 5LYZ[h] | 12% | 33% | 95% |
| 6LYZ[h] | 10% | 45% | 93% |

[a]See Table 2 for a description of each column.
[b]Vaney *et al.*, 1996.
[c]Artymiuk *et al.*, 1982.
[d]Lim *et al.*, 1998.
[e]Oki *et al.*, 1999.
[f]Girard *et al.*, 2001.
[g]Kurinov and Harrison, 1995.
[h]Diamond, 1974.

confidence scores for each assignment. These confidence scores are reported to the user. The highest-confidence assignments tend to be in regions of regular secondary structure (Fig. 5).

The computational complexity of the second phase is as follows. Matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are each of size $O(n \times n)$, where $n$ is the number of residues in the protein. Recomputing the tensors, using the Moore–Penrose pseudo-inverse of the $O(n) \times 5$ matrix takes time $O(n^2)$ (Golub and Van Loan, 1996). At least one residue is assigned per iteration; thus, the running time is $\sum_{i=1}^{n}(i^2 + i^2) = O(n^3)$, and the resonance assignment phase is guaranteed to be completed in $O(n^3)$ time. In practice, the resonance assignments can be computed in a couple of minutes on a Pentium-class workstation.

Occasionally, at the end of phase 2, it happens that Equation (4) cannot be satisfied. This occurs only on the last few iterations when, for example, the remaining two peaks each vote for the same residue. NVR handles this case by performing a maximum bipartite matching (Kuhn, 1955) for those peaks, and the second phase terminates. This does not increase the time complexity. As previously mentioned (Section 2), bipartite matching did not perform well (see Tables 1, 2 and 3) when run on all $n$ residues and $O(n)$

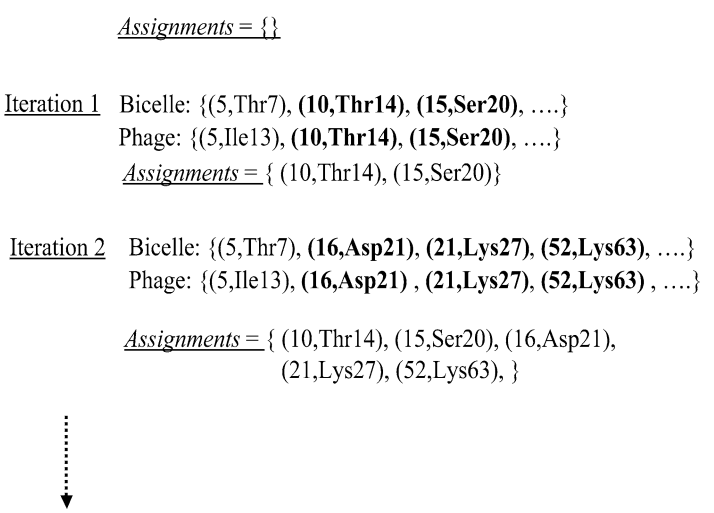


**FIG. 4.** Iterative assignments. The first two iterations of NVR with model 1UBQ. The assignment list is initially empty. At the end of the first iteration, both media "agree" that peaks 10 and 15 are residues Thr14 and Ser20, respectively. Consequently, those two assignments are added to the master assignment list. Note that there are only two assignments so there are not enough variables to update the tensors, $S_1$ and $S_2$, using Equation (1). At the beginning of the second iteration, the probability matrices, $M_1$ and $M_2$, are updated to reflect the fact that peaks Thr14 and Ser20 are already assigned. At the end of the second iteration, both media agree that peaks 16, 21, and 52 are Asp21, Lys27, and Lys63, respectively. These three assignments are added to the master assignment list. Now there are five assignments, so $S_1$ and $S_2$ can be updated using Equation (1). This procedure continues until the entire protein is assigned.

peaks: we only use it in the endgame to resolve the very small number of remaining assignments that Equation (4) cannot disambiguate.

### *4.3. Structure refinement (Phase 3)*

Once the final set of assignments has been computed, the (now) assigned RDCs are used to refine the structure of the model. Let $T \subset R$ be the set of residues whose back-computed RDCs values (one for each medium) are within 3 Hz of the experimentally observed RDCs. Set $T$ is used to refine the structure. A Monte Carlo algorithm was implemented to find a (new) conformation of the model's $\phi$ and $\psi$ backbone angles that best matches the observed RDCs. The program stops when either a) the RMSD between the RDCs associated with the set $T$ and those back-calculated from the modified structure is less than 0.3 Hz or b) one million structures have been considered, in which case the structure that best fits the data is output. The structure generated by the Monte Carlo method is then energy minimized using the Sander module of the program AMBER (Pearlman *et al.*, 1995). This minimization is done *in vacuo*. Figure 6 shows the results of the structure refinement of ubiquitin model 1G6J. An 11% reduction in RMSD was observed. This illustrates the potential application to structural genomics in which NVR could be used to assign and compute new structures based on homologous models.

## 5. RESULTS

### *5.1. Accuracy of tensor estimation algorithm*

Saupe matrices are completely specified by their eigenvalues and eigenvectors. Following standard notation (Wedemeyer *et al.*, 2002), we sort the eigenvectors by eigenvalue. We then compare eigenvalues and eigenvectors of the same rank. We compute the relative error between the estimated and actual eigenvalues (Tables 5–10). We then compare the directions of the estimated and actual eigenvectors. We show the angular error for each of the three eigenvectors (Tables 2–7).

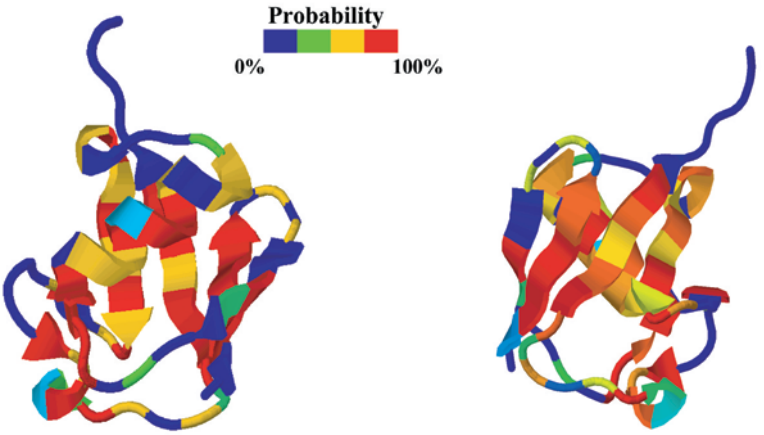


**FIG. 5.** Assignment confidences. NVR returns the *confidence* of each assignment. Here the structure of ubiquitin model 1UBQ (shown in two different orientations) is annotated with the confidence of each assignment. The color depicts the confidence with which the backbone amide group was assigned. Blue indicates low confidence, or missing data (e.g., prolines, which have no backbone amide group). Red indicates high confidence. The highest-confidence assignments tend to be in regions of regular secondary structure.

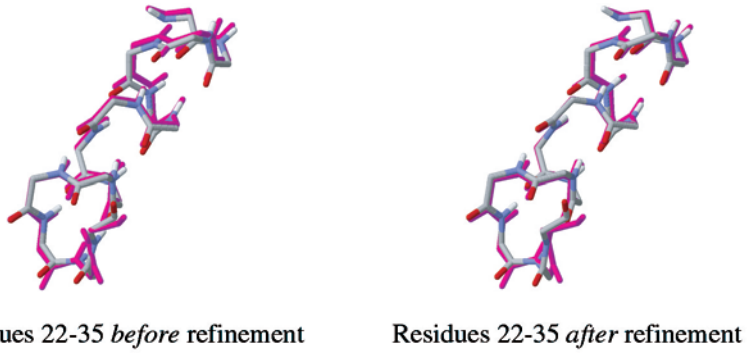


**FIG. 6.** 1G6J structure refinement. In magenta, the backbone of residues 22–35 from the structure 1D3Z. These residues form the first $\alpha$-helix in ubiquitin. Model 1D3Z is an RDC-refined model. In CPK-coloring are shown the backbone of residues 22–35 of model 1G6J (**left**) and a new structure (**right**) generated after structure refinement of 1G6J (using the RDC assignments from NVR). The RMSD between the two backbones on the right is 11% smaller than the RMSD of the backbones on the left.

To quantify the accuracy of the rotation search in the tensor estimation algorithm, we use the method of Yan *et al.* (2003) to compute a percentile that measures the fraction of all tensor orientations which fall within the angular deviations of the computed Saupe matrix from the actual Saupe matrix. Suppose we randomly and isotropically rotate the estimated Saupe matrix. We compute the probability $P_S$ that the eigenvectors of the randomly oriented Saupe matrix are simultaneously within the three angular errors measured. By integrating isotropically over $SO(3)$, we compute an upper bound on $P_S$, which includes a four-fold symmetry factor due to symmetry of the dipolar operator.[4]

Finally, we convert the probability $P_S$ to a percentile which measures the fraction of all orientations in $SO(3)$ which fall outside the angular eigenvector errors. Our results show upper bounds on $P_S$ of less than 19%, which translates into percentiles of at least 81% accuracy. The mean tensor accuracy in percentile is 97% (Ubiquitin), 97% (SPG), and 98% (Lysozyme) (see Tables 8, 9, 10).

[4]There is an inversion symmetry for each of the three eigenvectors. Therefore, there are eight isometries which leave the Saupe matrix unchanged. However, only four of those isometries are pure rotations ($SO(3)$). The other four are perversions in $O(3) - SO(3)$ (rotations composed with a reflection), and hence are not used to integrate over $SO(3)$.

TABLE 5. UBIQUITIN TENSOR IMPROVEMENTS[a]

| | *Bicelle 292°K* | | | | | | *Bicelle 298°K* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Percent difference* | | *Angular difference* | | | | *Percent difference* | | *Angular difference* | | | |
| *Model* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* |
| 1G6J | 0 | 0.6 | 0.2 | 0.3 | 0.2 | 100 | 0 | 0 | 0.2 | 0.2 | 0.1 | 100 |
| 1UBI | 0.1 | 0.2 | 2.3 | 2.4 | 0.6 | 100 | 0 | 0 | 0.2 | 0.2 | 0.1 | 100 |
| 1UBQ | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 1UD7 | 0 | 0.1 | 0.5 | 0.2 | 0.5 | 100 | 0 | 0 | 0.7 | 0.9 | 0.6 | 100 |

[a]The accuracies of the *final* tensor estimates, after NVR has completed the resonance assignment phase. The accuracy is improved from the initial tensor estimates (see Table 8).

TABLE 6. SPG TENSOR IMPROVEMENTS[a]

| | *Phage* | | | | | | *Bicelle* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Percent difference* | | *Angular difference* | | | | *Percent difference* | | *Angular difference* | | | |
| *Model* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* |
| 1GB1 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 2GB1 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 1PGB | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |

[a]The accuracies of the *final* tensor estimates, after NVR has completed the resonance assignment phase. The accuracy is improved from the initial tensor estimates (see Table 9).

TABLE 7. LYSOZYME TENSOR IMPROVEMENTS[a]

| | *5% Bicelle* | | | | | | *7.5% Bicelle* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Percent difference* | | *Angular difference* | | | | *Percent difference* | | *Angular difference* | | | |
| *Model* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* |
| 193L | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 1AKI | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 1AZF | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 1BGI | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 1H87 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 1LSC | 0.1 | 0.1 | 0 | 0.1 | 0.1 | 100 | 0 | 0.1 | 0 | 0 | 0 | 100 |
| 1LSE | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 1LYZ | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 2LYZ | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 3LYZ | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 4LYZ | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 5LYZ | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 100 |
| 6LYZ | 1.5 | 3.3 | 0.7 | 1.2 | 1.0 | 100 | 1.9 | 5.8 | 0.8 | 5.3 | 5.2 | 100 |

[a]The accuracies of the *final* tensor estimates, after NVR has completed the resonance assignment phase. The accuracy is improved from the initial tensor estimates (see Table 10).

TABLE 8. UBIQUITIN TENSOR ESTIMATES[a]

| | *Bicelle 292°K* | | | | | | *Bicelle 298°K* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Percent difference* | | *Angular difference* | | | | *Percent difference* | | *Angular difference* | | | |
| *Model* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* |
| 1G6J | 2.3 | 0.2 | 20.8 | 25.1 | 21.8 | 98 | 12.0 | 5.0 | 28.1 | 30.3 | 16.1 | 96 |
| 1UBI | 1.1 | 3.7 | 27.3 | 28.2 | 7.1 | 96 | 15.2 | 8.3 | 28.4 | 17.8 | 27.7 | 96 |
| 1UBQ | 0.8 | 2.6 | 17.5 | 11.7 | 20.8 | 99 | 15.3 | 7.9 | 16.4 | 27.3 | 32.0 | 95 |
| 1UD7 | 0.2 | 2.2 | 21.2 | 16.5 | 25.8 | 98 | 14.7 | 6.9 | 16.9 | 16.3 | 7.4 | 99 |

[a]This table demonstrates the accuracy of the first step of the NVR algorithm–tensors estimation. (Columns 2 and 3) Percentage difference for the axial and rhombic terms, $D_a$ and $D_r$, for the four models, 1G6J, 1UBI, 1UBQ, and 1UD7, versus the actual axial and rhombic terms in the bicelle medium recorded at 292° K. The $D_a$ and $D_r$ differences are normalized by the range of the experimentally measured dipolar coupling values. (Columns 4–6) Angular differences (in degrees) between the eigenvectors of the estimated tensors and the eigenvectors of the actual tensors in the bicelle medium at 292° K: $S_{zz}$ is the director of the tensor (i.e., the eigenvector associated with the largest eigenvalue of the tensor), and $S_{xx}$ and $S_{yy}$ are eigenvectors associated with the second largest and smallest eigenvalue of the tensor, respectively. (Columns 8 and 9, columns 10–12) Accuracy of the tensor estimates in the bicelle medium recorded at 298° K. Columns 7 and 13 report the accuracy of the tensor estimate as a percentile (see Section 5.1).

TABLE 9. SPG TENSOR ESTIMATES

| | *Phage* | | | | | | *Bicelle* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Percent difference* | | *Angular difference* | | | | *Percent difference* | | *Angular difference* | | | |
| *Model* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* |
| 1GB1 | 0.6 | 6.0 | 26.8 | 23.3 | 21.4 | 97 | 2.4 | 6.6 | 17.9 | 20.5 | 22.3 | 98 |
| 2GB1 | 0.2 | 0.5 | 26.8 | 23.3 | 21.4 | 97 | 1.7 | 10.3 | 17.9 | 20.5 | 22.3 | 98 |
| 1PGB | 0.6 | 6.0 | 23.8 | 24.5 | 28.8 | 97 | 2.4 | 6.6 | 15.2 | 29.3 | 25.8 | 96 |

TABLE 10. LYSOZYME TENSOR ESTIMATES

| | *5% Bicelle* | | | | | | *7.5% Bicelle* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Percent difference* | | *Angular difference* | | | | *Percent difference* | | *Angular difference* | | | |
| *Model* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* | $D_a$ | $D_r$ | $S_{zz}$ | $S_{xx}$ | $S_{yy}$ | *Percentile* |
| 193L | 1.5 | 0.1 | 16.7 | 6.7 | 16.7 | 99 | 8.8 | 8.7 | 38.6 | 49.0 | 33.2 | 85 |
| 1AKI | 2.3 | 0.5 | 13.2 | 10.6 | 8.5 | 99 | 10.0 | 9.3 | 23.2 | 51.0 | 45.2 | 81 |
| 1AZF | 1.7 | 0.5 | 7.6 | 7.3 | 5.6 | 99 | 9.5 | 8.5 | 31.2 | 29.6 | 11.0 | 95 |
| 1BGI | 1.2 | 0.7 | 30.0 | 8.5 | 29.8 | 96 | 8.9 | 9.4 | 24.6 | 43.8 | 35.7 | 89 |
| 1H87 | 2.1 | 0.2 | 26.2 | 29.9 | 34.2 | 94 | 9.9 | 8.6 | 23.8 | 15.3 | 25.8 | 97 |
| 1LSC | 1.7 | 0.4 | 16.1 | 20.8 | 22.8 | 98 | 8.9 | 8.5 | 12.2 | 12.0 | 11.6 | 99 |
| 1LSE | 1.7 | 0.4 | 12.6 | 49.2 | 44.5 | 83 | 9.5 | 8.3 | 29.2 | 48.2 | 42.1 | 84 |
| 1LYZ | 9.8 | 5.0 | 10.7 | 21.4 | 18.5 | 99 | 18.9 | 8.5 | 21.3 | 21.0 | 24.1 | 98 |
| 2LYZ | 3.5 | 1.8 | 20.8 | 16.2 | 16.2 | 99 | 11.56 | 8.3 | 23.8 | 25.0 | 7.5 | 98 |
| 3LYZ | 4.3 | 2.4 | 20.0 | 31.4 | 25.2 | 96 | 12.7 | 8.0 | 27.8 | 38.1 | 4.4 | 96 |
| 4LYZ | 3.1 | 2.3 | 24.0 | 9.3 | 24.0 | 98 | 12.6 | 8.6 | 12.7 | 14.5 | 17.7 | 99 |
| 5LYZ | 3.1 | 2.3 | 23.9 | 9.3 | 24.0 | 98 | 12.6 | 8.6 | 12.7 | 14.5 | 17.7 | 99 |
| 6LYZ | 3.0 | 0.7 | 15.7 | 16.8 | 16.8 | 99 | 11.0 | 8.6 | 26.6 | 37.3 | 46.0 | 87 |

### 5.2. Accuracy of resonance assignment algorithm

The molecular structure of human ubiquitin has been investigated extensively. A variety of data have been published including resonance assignments (Weber *et al.*, 1987; Schneider *et al.*, 1992), backbone amide residual dipolar couplings recorded in two separate liquid crystals (Cornilescu *et al.*, 1998), amide-exchange rates (Cornilescu *et al.*, 1998), $^{15}$N-HSQC and $^{15}$N-HSQC NOESY spectra (Harris, 2002), and several independent high-resolution structures solved by both x-ray crystallography (Ramage *et al.*, 1994; Vijay-Kumar *et al.*, 1987) and NMR (Babu *et al.*, 2001; Johnson *et al.*, 1999). In 1998, the Bax lab published a new NMR structure for ubiquitin, (PDB Id: 1D3Z) (Cornilescu *et al.*, 1998). Unlike previous ubiquitin structures, 1D3Z was refined using dipolar couplings. NVR was tested on four alternative high-resolution structures (PDB Ids: 1G6J, 1UBI, 1UBQ, 1UD7) of human ubiquitin, none of which have been refined using dipolar couplings. Structures 1G6J, 1UBI, and 1UBQ have 100% sequence identity to 1D3Z. Structure 1UD7 is mutant of ubiquitin where seven hydrophobic core residues have been altered (I3V, V5L, I13V, L15V, I23F, V26F, L67I). Structure 1UD7 was chosen to test the effectiveness of NVR when the model is a close homolog of the target protein. We ran four independent trials, one for each of 1G6J, 1UBI, 1UBQ, and 1UD7. In each test, both sets of experimentally recorded backbone amide dipolar couplings (Cornilescu *et al.*, 1998) for human ubiquitin were fit to the amide bond vectors of the selected model. The $^{15}$N-HSQC and $^{15}$N-HSQC NOESY spectra (Harris, 2002) were processed to extract sparse, unassigned $d_{\mathrm{NN}}$s.

Best results were obtained when we used the final tensors generated after a complete run of NVR to bootstrap the resonance assignment phase. In this bootstrapping scheme, the assignment phase of NVR is run twice. At the end of the first run, the computed assignments are used to construct a "final" tensor estimate for each medium. This is done by using the assignments and the model, by means of the SVD method (Losonczi *et al.*, 1999). These final tensor estimates from the first run are used as initial tensor estimates for the second run of the assignment phase. The assignments computed in the first run, which have a median accuracy of 92%, are not used during the second run. In essence, the first run is used to refine the tensor estimates in preparation for the second run. This bootstrapping strategy was used uniformly to generate all the experimental results shown in Table 11A–D. On the second run, NVR achieves an assignment accuracy of 92–100% for the four ubiquitin models (Table 11A). The assignment accuracies on NMR data for the B1 domain of streptococcal protein G and lysozyme were 100% and 97–100%, respectively (See Table 11B–D). NVR performed well on 1UD7, a mutant of ubiquitin. This suggests that NVR might be extended to homologous structures. NVR achieves consistently high accuracies, suggesting NVR is robust with respect to choice of model.

We have found that the errors that our algorithm makes are, in general, easily explained. Almost all errors are symmetric. That is, if residue A was mistaken for residue B, then B was mistaken for A. All these errors involved dipolar couplings that were very different from their expected values. For example, in the trial on Lysozyme model 1LSC, Thr118 was mistaken for Leu129 and vice versa. The observed dipolar couplings for these two residues were an average of 2.9 Hz different from their expected values in both media. By making the incorrect assignment, the NVR method reduced the apparent discrepancy to an average of 1.3 Hz.

There were only two cases, from our 20 separate trials, where a small chain of misassignments was seen. The following chain was observed in Ubiquitin model 1UBI: Lys6 → Glu16 → Thr12 → Lys6. The following chain was observed in Lysozyme model 6LYZ: Met105 → Ala107 → Thr118 → Met105. These cyclic errors also reduce the apparent discrepancy between expected and observed dipolar couplings. We have recently extended the NVR method (Langmead and Donald, 2004a) to include $^{1}$H and $^{15}$N chemical shift prediction (Osapay and Case, 1991; Wishart *et al.*, 1997) and showed that accurate chemical shift prediction will prevent these kinds of errors. Brüschweiler and coworkers describe a similar chain (cyclic permutation) of errors (Hus *et al.*, 2002) for the one protein (1UBI) on which HPB was tested (Thr9 → Arg74 → Tyr59 → Gly53 → Thr9). NVR found no cyclic permutation of length longer than 3, for any model, including 1UBI.

As described in Section 4.2, NVR reports a confidence associated with each assignment. These confidences are expressed as percentages (see Fig. 5). We note that in all our experiments, no incorrect assignment ever yielded higher than a 44% confidence. Hence, the NMR structural biologist can use the confidence values to ensure 100% accuracy on a subset of the peaks, by selecting this threshold.

TABLE 11. ACCURACY[a]

| (A: Ubiquitin) | | | (B: SPG) | | |
|---|---|---|---|---|---|
| *PDB ID* | *Exp. method* | *Accuracy* | *PDB ID* | *Exp. method* | *Accuracy* |
| 1G6J[b] | NMR | 97 | 1GB1[l] | NMR | 100% |
| 1UBI[c] | X-ray (1.8 Å) | 92 | 2GB1[l] | NMR | 100% |
| 1UBQ[d] | X-ray (1.8 Å) | 100 | 1PGB[m] | X-ray (1.92 Å) | 100% |
| 1UD7[e] | NMR | 93 | | | |
| (C: Lysozyme) | | | (D: Lysozyme (*continued*)) | | |
| *PDB ID* | *Exp. method* | *Accuracy* | *PDB ID* | *Exp. method* | *Accuracy* |
| 193L[f] | X-ray (1.3 Å) | 100% | 1LYZ[n] | X-ray (2.0 Å) | 100% |
| 1AKI[g] | X-ray (1.5 Å) | 100% | 2LYZ[n] | X-ray (2.0 Å) | 100% |
| 1AZF[h] | X-ray (1.8 Å) | 100% | 3LYZ[n] | X-ray (2.0 Å) | 100% |
| 1BGI[i] | X-ray (1.7 Å) | 100% | 4LYZ[n] | X-ray (2.0 Å) | 100% |
| 1H87[j] | X-ray (1.7 Å) | 100% | 5LYZ[n] | X-ray (2.0 Å) | 100% |
| 1LSC[k] | X-ray (1.7 Å) | 98% | 6LYZ[n] | X-ray (2.0 Å) | 97% |
| 1LSE[k] | X-ray (1.7 Å) | 100% | | | |

[a](A) NVR achieves an accuracy of 92–100% on the four ubiquitin models. The structure 1D3Z (Cornilescu *et al.*, 1998) is the only published structure of ubiquitin to have been refined against RDCs. The RDCs used by Cornilescu *et al.* (1998) have been published and were used in each of the 4 NVR trials; 1G6J, 1UBI, and 1UBQ have 100% sequence identity to 1D3Z, and 1UD7 is a mutant form of human ubiquitin. As such, it demonstrates the effectiveness of NVR when the model is a close homolog of the target protein. (B–D) The RDCs for the B1 domain of streptococcal protein G (Kuszewski *et al.*, 1999) and hen lysozyme (Schwalbe *et al.*, 2001) were obtained from the PDB. NOEs and amide exchange data were extracted from their associated restraints files. NVR achieves an accuracy of 100% (Table B) and 97–100% (Tables C and D), respectively.

[b]Babu *et al.*, 2001.
[c]Ramage *et al.*, 1994.
[d]Vijay-Kumar *et al.*, 1987.
[e]Johnson *et al.*, 1999.
[f]Vaney *et al.*, 1996.
[g]Artymiuk *et al.*, 1982.
[h]Lim *et al.*, 1998.
[i]Oki *et al.*, 1999.
[j]Girard *et al.*, 2001.
[k]Kurinov and Harrison, 1995.
[l]Gronenborn *et al.*, 1991.
[m]Gallagher *et al.*, 1994.
[n]Diamond, 1974.

### *5.3. Discussion*

The key difference between NVR and traditional maximum bipartite matching (MBM) algorithms is that NVR incorporates the conditional geometric constraints encoded in NOEs. They are conditional in the sense that they only become a constraint given the geometric relationship between two different residues. Computing MBM, given such constraints, is known to be NP-hard (Xu *et al.*, 2002). The approximation scheme employed by Xu and coworkers (2002), achieves 45–56% accuracy. We note that it is well known that MBM is sensitive to noise and to outliers in particular. NVR is, by nature, greedy, and not an approximation algorithm. NVR takes a conservative approach, making only a few, likely assignments. However, by making these assignments, NVR is then able to leverage the conditional probabilities encoded in NOEs. Intuitively, a peak whose assignment may be ambiguous in iteration $i$ may become unambiguous in iteration $i + 1$.

There are a number of limitations to our algorithm worth noting. The first is that we have only tested NVR only on models with both high sequence and structural homology. Consequently, the present form of the algorithm may be best applied to scenarios where a crystal structure of the same protein is available, as may

be the case in a SAR by NMR study. Models with significantly less homology will likely have somewhat different networks of hydrogen bonds and NOEs, as well as different amide bond-vector orientations. The probabilistic framework in which RDCs are interpreted will likely be robust to reasonable amounts of variation. In contrast, the hard constraints employed by NVR in the interpretation of amide-exchange rates and $d_{\mathrm{NN}}$'s may force assignment errors in these cases (Langmead and Donald, 2004a, pg. 123). A more comprehensive analysis of the performance of the algorithm under varying amounts of homology (both structural and sequential) remains an important goal. Computational modeling could be used to construct a variety of alternative models having strictly controlled amounts of homology. These models may inform, for example, the minimum amount of homology required by NVR for a given set of experimental data. Similarly, a thorough analysis of the performance of the tensor estimation algorithm (phase 1) is warranted. In particular, it will be useful to establish both the sensitivity of the algorithm to statistical outliers, as well as a comparison to alternatives to relative entropy as a similarity measurement. We note, however, that recent enhancements to the NVR algorithm (Langmead and Donald, 2004a) obviate the need for an explicit search over $SO(3)$ for the tensors. Finally, phase 3 of the NVR algorithm (structure refinement) presently involves unconstrained energy minimization, which is known to be sensitive to initial conditions. We are presently exploring alternative means for constructing and refining structural models using assigned RDCs, including the use of exact algorithms (Wang and Donald, 2004).

### *5.4. 3D structural homology detection*

We have also extended NVR to a second application—3D structural homology detection. While many sequence-based homology prediction methods exist, an important challenge remains: two highly dissimilar sequences can have similar folds. For example, the backbone RMSD between the human ubiquitin structure (PDB Id: 1D3Z) and the structure of the Ubx domain from human Faf1 (PDB Id: 1H8C) is quite small (1.9 Å), yet they have only 16% sequence identity. NVR is well suited for identifying these remote homologies because it considers only the backbone geometry of each amino acid in the model, not the geometry of side chains. In particular, given a 3D model of the backbone of *any* protein, NVR can compute how well the experimental RDC data fits that model. One would expect that a structural homolog would fit the data quite well, while an unrelated structure would not. NVR can also be used to confirm or refute structural predictions made by other techniques, such as protein threading or sequence homology.

We have assembled a database of 2,456 backbone structural models from the Protein Data Bank (Berman *et al.*, 2000) representing a variety of different fold-families. The database includes the structures of ubiquitin (PDB Id: 1D3Z), lysozyme (PDB Id: 1E8L), and SPG (PDB Id: 3GB1) and five structural homologs for each of these three proteins (Table 12). These homologs have between 10–61% sequence homology to 1D3Z, 1E8L, and 3GB1. The database contains only the backbone geometry, the length of the primary sequence, and the percentage of $\alpha$ and $\beta$ secondary structure for each protein. The protein's primary sequence is not used.

Using the primary sequences of our three test proteins (1D3Z, 1E8L, and 3GB1), we estimated their secondary structure using the program JPRED (Cuff *et al.*, 1998). The native fold was not used to estimate secondary structure. Next, using the experimental RDCs for the three test proteins, we ran NVR's tensor estimation (Section 4.1) against each model in the database. Note that the tensor estimation phase does not require NOEs or amide-exchange data. Therefore, it is not necessary to record these experiments in order to perform homology detection. Alternatively, homology detection could proceed in parallel while these experiments are being recorded. The tensor estimation phase takes $O(nk^3)$ time. Thus, a database consisting of $p$ structural models can be searched in $O(pnk^3)$ time.

Each model in the database is assigned a score. Let $\Delta_\alpha = \mid \alpha_t - \alpha_m \mid$ and $\Delta_\beta = \mid \beta_t - \beta_m \mid$, where $\alpha_t$ and $\beta_t$ are the predicted percentages of $\alpha$ and $\beta$ structure for the target protein, $t$, and $\alpha_m$ and $\beta_m$ are the actual percentages of $\alpha$ and $\beta$ structure taken from the model, $m$. Let $\Delta_l$ be the difference in length between $t$ and $m$. Finally, let $KL_1$ and $KL_2$ be the Kullback–Leibler distances of the two tensor estimates[5] (Section 4.1). A model's score is computed as follows:

$$I_m = \Delta_\alpha + \Delta_\beta + \Delta_l + KL_1 + KL_2. \tag{5}$$

[5]Both $\Delta_\alpha$ and $\Delta_\beta$ are multiplied by 100 so that they have the same order of magnitude as $\Delta_l$, $KL_1$, and $KL_2$.

TABLE 12. STRUCTURAL HOMOLOGY DETECTION RESULTS[a]

| *PDB ID* | *Homolog* | *Sequence identity* | *RMSD* | *Rank* |
|---|---|---|---|---|
| 1D3Z | | 100% | 0 Å | 1 |
| | 1NDD | 55.6% | 0.6 Å | 2 |
| | 1BT0 | 61.0% | 0.7 Å | 3 |
| | 1H8C | 15.7% | 1.9 Å | 11 |
| | 1GUA | 11.6% | 2.1 Å | 19 |
| | 1C1Y | 11.6% | 2.1 Å | 38 |
| 1E8L | | 100% | 0 Å | 1 |
| | 2EQL | 49.2% | 1.8 Å | 2 |
| | 1ALC | 35.8% | 1.8 Å | 3 |
| | 1HFZ | 38.3% | 1.8 Å | 4 |
| | 1A4V | 38.2% | 1.8 Å | 5 |
| | 1F6S | 38.7% | 1.7 Å | 6 |
| 3GB1 | | 100% | 0 Å | 1 |
| | 1HZ5 | 14.5% | 2.2 Å | 2 |
| | 1JML | 12.8% | 1.8 Å | 5 |
| | 1HEZ | 12.7% | 2.0 Å | 12 |
| | 2GCC | 10.0% | 2.6 Å | 24 |
| | 1HZ6 | 14.5% | 2.2 Å | 55 |

[a]The sequence identity and RMSD of the three test proteins and their respective five homologs. The final column is the rank of that model (out of 2,546 structures) based on the score computed by NVR.

Each model is then ranked according to its score. As seen in Table 12, the highest ranking structure is the native structure. The five homologous structures are also highly ranked, relative to the 2,500 structures in the database. The scores associated with the native fold and the five homologs are statistically significantly lower than the scores of unrelated proteins ($p$-values of $2.6 \times 10^{-5}$, $2.3 \times 10^{-5}$, and $2.9 \times 10^{-5}$ for 1D3Z, 1E8L, and 3GB1, respectively). Thus, NVR is able to identify structural homologies between proteins with remote amino acid sequences, without employing or performing resonance assignments. Figure 7 is a scatter-plot of the scores computed by NVR versus the backbone RMSD of our three test proteins and the models in the database. The native and homologous structures tend to form a cluster. Unfortunately, NVR also tends to report some unrelated structures (i.e., false positives), as seen in Fig. 7. Furthermore, while the scores between homologous and nonhomologous structures are statistically significantly different, the correlation between score and RMSD is weak. This suggests that the particular scoring mechanism employed here may be most useful as a coarse filter. We note, however, that recent enhancements to NVR (Langmead and Donald, 2004a), coupled with a new algorithm for structural homology detection (Langmead and Donald, 2004c), have eliminated this weakness. In particular, our new algorithm for structural homology detection from sparse NMR data, called HD, reports no false positives or false negatives on a larger set of proteins against a larger database (Langmead and Donald, 2004c).

## 6. CONCLUSION

We have described a fast, automated procedure for high-throughput NMR resonance assignments for a protein of known structure, or of an homologous structure. NMR assignments are useful for probing protein–protein interactions, protein–ligand binding, and dynamics by NMR, and they are the starting point for structure refinement. A new algorithm, Nuclear Vector Replacement (NVR) was introduced to compute assignments that optimally correlate experimentally measured NH residual dipolar couplings (RDCs) to a given a priori whole-protein 3D structural model. NVR requires only uniform $^{15}$N-labeling of the protein and processes unassigned $^{15}$N-HSQC and H-D exchange-HSQC spectra, $H^N$-$^{15}$N RDCs, and sparse $H^N$-$H^N$

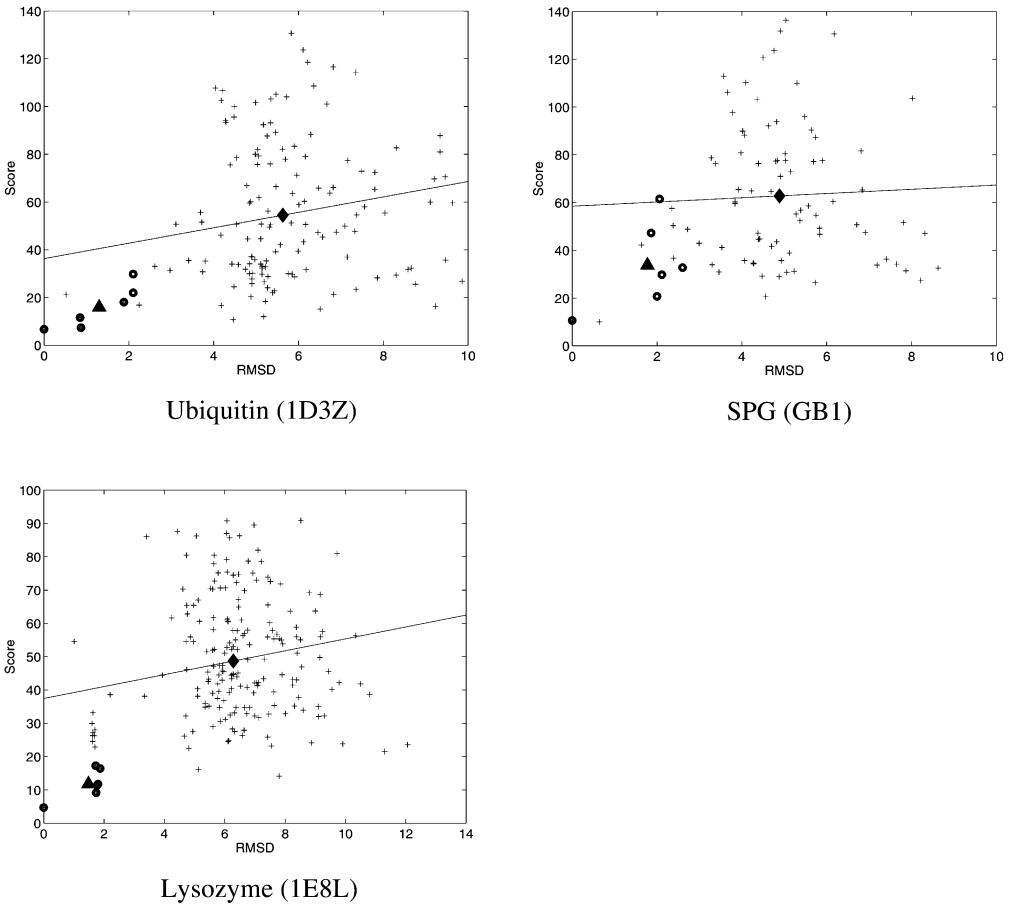


**FIG. 7.** RMSD versus NVR homology score. 3 Scatter plots of the backbone RMSD between the native structures of Ubiquitin (**top left**), SPG (**top right**), and Lysozyme (**bottom**) and the models in the database versus the score computed by NVR. Only those proteins whose length is within 10% of the native structure are shown. The open circles are the data points for the native structure and five homologous structures. The + signs are the data points associated with nonhomologous proteins. The diamond is the 2D mean of the +'s while the triangle is the 2D mean of the open circles. The trend line shows the correlation between the score computed by NVR and RMSD for all the data points.

NOE's ($d_{\mathrm{NN}}$s), all of which can be acquired in a fraction of the time needed to record the traditional suite of experiments used to perform resonance assignments. NVR efficiently assigns the $^{15}$N-HSQC spectrum, as well as the $d_{\mathrm{NN}}$s of the 3D $^{15}$N-NOESY spectrum, in $O(n^3)$ time. We tested NVR on NMR data from three proteins using 20 different alternative structures. When NVR was run on NMR data from the 76-residue protein, human ubiquitin (matched to four structures, including one mutant/homolog), we achieved an assignment accuracy of 92–100%. Similarly good results were obtained on NMR data for streptococcal protein G (100%) and hen lysozyme (97–100%) when they were matched by NVR to a variety of 3D structural models.

We have shown that NVR works well on proteins in the 56–129 residue range. It is to be expected that some modifications may be needed when scaling NVR to larger proteins. The accuracy of the powder pattern method is known to increase as the number of RDCs increases. Thus, our ability to estimate the axial and rhombic components of the alignment tensors should increase with protein size. Estimating the eigenvectors of the tensors, however, will become harder as the distribution of amide bond vectors becomes more uniform. The current version of the NVR algorithm assumes nearly complete data. We have recently extended it to handle the case when either the set of resonances or RDCs are incomplete (Langmead and Donald, 2004a). We are also incorporating $^{1}$H and $^{15}$N chemical shift prediction (Osapay and Case, 1991; Wishart *et al.*, 1997) into NVR (Langmead and Donald, 2004a).

Finally, we have demonstrated that NVR can be used to identify 3D structural homologies between proteins with remote amino acid sequences. Furthermore, our success in assigning 1UD7, which is a mutant

of ubiquitin, suggests that NVR could be applied more broadly to assign spectra based on homologous structures. Using the results of a sequence alignment algorithm (Altschul *et al.*, 1990), protein threading (Lathrop and Smith, 1996; Xu *et al.*, 2000), or homology modeling (Blundell *et al.*, 1987; Fetrow and Bryant, 1993; Greer, 1991; Johnson *et al.*, 1994; Sali *et al.*, 1990), one would modify NVR to perform assignments by matching RDCs to an homologous structure. It is likely that the structure refinement phase would be folded into the main iterative loop so that the homologous structure would be simultaneously assigned and refined. Thus, NVR could play a role in structural genomics.

See Langmead and Donald, 2004a; Langmead and Donald, 2003; Langmead and Donald, 2004b; and Langmead and Donald, 2004c for improvements in accuracy, complexity, and robustness.

## APPENDIX

### A. *Complexity of minimum Kullback–Leibler distance*

We implemented an $O(nk^3)$ discrete-grid rotation search for initial tensor estimation. We now show how the rotation minimizing the Kullback–Leibler distance can be computed in polynomial time (without a grid search) using the first-order theory of real-closed fields (Grigor'ev, 1988; Grigor'ev and Vorobjov, 1988; Basu and Roy, 1996; Basu, 1997). Hence the $O(nk^3)$ discrete-grid rotation search in Section 4.1 can be replaced by a combinatorially precise algorithm, eliminating all dependence of the rotation search upon the resolution $k$.

Suppose two variables of the same type are characterized by their probability distributions $f$ and $f'$. The relative entropy formula is given by $KL(f, f') = \sum_{i=1}^{m} f_i \ln(f_i/f_i')$, where $m$ is the number of levels of the variables. We will use a polynomial approximation to $\ln(\cdot)$. Let us represent rotations by unit quaternions, and use the substitution $u = \tan(\theta/2)$ to 'rationalize' the equations using rotations, thereby yielding purely algebraic (polynomial) equations. Let $V$ be such a rotation (quaternion), $D$ be the unassigned experimentally-measured RDCs, $E$ be the set of model NH vectors and $B(V)$ be the set of unassigned, back-computed RDCs (parameterized by $V$). Hence, from Eqs. (1,2), $B(V) = E^T \mathbf{S} E = (E^T (V^T \Sigma V) E) = \{ \mathbf{w}^T (V^T \Sigma V) \mathbf{w} \mid \mathbf{w} \in E \}$. (We have ignored $D_{\max}$ here for the simplicity of exposition). We wish to compute

$$\operatorname*{argmin}_{V \in S^3} KL(D, B(V)) \tag{6}$$

(We use the unit 3-sphere $S^3$ instead of $SO(3)$, since the quaternions are a double-covering of rotation space). Equation (6) can be tranformed into a sentence in the language of semi-algebraic sets (the first order theory of real closed fields):

$$\exists V_0 \in S^3,\ \forall V \in S^3 :\ KL(D, B(V_0)) \leq KL(D, B(V)). \tag{7}$$

$S^3$ and $SO(3)$ are semi-algebraic sets, and Equation (7) is a polynomial inequality with bounded quantifier alternation ($a = 1$). The number of DOF (the number of variables) is constant ($r = 3$ DOF for rotations), and the size of the equations is $O(n)$. Hence Equation (7) can be decided exactly, in polynomial time, using the theory of real-closed fields. We will use Grigor'ev's algorithm (Grigor'ev, 1988; Grigor'ev and Vorobjov, 1988) for deciding a Tarski sentence, which is singly-exponential in the number of variables, and doubly-exponential only in the number of quantifer alternations. The time complexity of Grigor'ev's algorithm is $n^{O(r)^{4a-2}}$, which in our case ($a = 1$, $r = 3$) reduces to $n^{O(1)}$ which is polynomial time.

## 7. ACKNOWLEDGEMENTS

Some of the key ideas in this paper arose in discussions with Dr. T. Lozano-Pérez, and we are grateful for his advice and support. We thank Drs. A. Anderson, C. Bailey-Kellogg, J. Hoch, and B. Hare, Ms. E. Werner-Reiss, and all members of Donald Lab for helpful discussions and comments on drafts.

This work was supported by the following grants to BRD: National Institutes of Health (R01 GM-65982), National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, EIA-0305444, and EIA-9802068), and the John Simon Guggenheim Foundation.

## REFERENCES


Al-Hashimi, H.M., Gorin, A., Majumdar, A., Gosser, Y., and Patel, D.J. 2002. Towards structural genomics of RNA: Rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings. *J. Mol. Biol.* 318, 637–649.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

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

Annila, A., Aitio, H., Thulin, E., and Drakenberg, T. 1999. Recognition of protein folds via dipolar couplings. *J. Biom. NMR* 14, 223–230.

Artymiuk, P.J., Blake, C.C.F., Rice, D.W., and Wilson, K.S. 1982. The structures of the monoclinic and orthorhombic forms of hen egg-white lysozyme at 6 angstroms resolution. *Acta Crystal. B Biol. Crystal.* 38, 778.

Babu, C.R., Flynn, P.F., and Wand, A.J. 2001. Validation of protein structure from preparations of encapsulated proteins dissolved in low viscosity fluids. *J. Am. Chem. Soc.* 123, 2691.

Bailey-Kellogg, C., Widge, A., Kelley III, J.J., Berardi, M.J., Bushweller, J.H., and Donald, B.R. 2000. The NOESY Jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comp. Biol.* 7(3–4), 537–558.

Basu, S. 1997. An improved algorithm for quantifier elimination over real closed fields. *IEEE FOCS*, 56–65.

Basu, S., and Roy, M.F. 1996. On the combinatorial and algebraic complexity of quantifier elimination. *J. ACM* 43(6), 1002–1045.

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

Blundell, T.L., Sibanda, B.L., Sternberg, M.J., and Thornton, J.M. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326, 347–352.

Chen, Y., Reizer, J., Saier Jr., M.H., Fairbrother, W.J., and Wright, P.E. 1993. Mapping of the binding interfaces of the proteins of the bacterial phosphotransferase system, HPr and IIAglc. *Biochemistry* 32(1), 32–37.

Chou, J.J., Li, S., and Bax, A. 2000. Study of conformational rearrangement and refinement of structural homology models by the use of heteronuclear dipolar couplings. *J. Biom. NMR* 18, 217–227.

Cornilescu, G., Marquardt, J.L., Ottiger, M., and Bax, A. 1998. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.* 120, 6836–6837.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M., and Barton, G.J. 1998. Jpred: A consensus secondary structure prediction server. *Bioinformatics* 14, 892–893.

Delaglio, F., Kontaxis, G., and Bax, A. 2000. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc.* 122, 2142–2143.

Diamond, R. 1974. Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol.* 82, 371–391.

Fejzo, J., Lepre, C.A., Peng, J.W., Bemis, G.W., Ajay, Murcko, M.A., and Moore, J.M. 1999. The SHAPES strategy: An NMR-based approach for lead generation in drug discovery. *Chem. Biol.* 6, 755–769.

Fetrow, J.S., and Bryant, S.H. 1993. New programs for protein tertiary structure prediction. *Bio/Technology* 11, 479–484.

Fiaux, J., Bertelsen, E.B., Horwich, A.L., and Wüthrich, K. 2002. NMR analysis of a 900K GroELGroES complex. *Nature* 418, 207–211.

Fowler, C.A., Tian, F., Al-Hashimi, H.M., and Prestegard, J.H. 2000. Rapid determination of protein folds using residual dipolar couplings. *J. Mol. Biol.* 304(3), 447–460.

Gallagher, T., Alexander, P., Bryan, P., and Gilliland, G.L. 1994. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* 33, 4721–4729.

Girard, E., Chantalat, L., Vicat, J., and Kahn, R. 2001. Gd-HPDO3A, a complex to obtain high-phasing-power heavy atom derivatives for SAD and MAD experiments: Results with tetragonal hen egg-white lysozyme. *Acta Crystal. D Biol. Crystal.* 58, 1–9.

Golub, G.H., and Van Loan, C.F. 1996. *Matrix Computations*, 3rd ed., 253–254, Johns Hopkins Univeristy Press, Baltimore, MD.

Greer, J. 1991. Comparative modeling of homologous proteins. *Meth. Enzymol.* 202, 239–252.

Grigor'ev, D.Y. 1988. Complexity of deciding Tarski algebra. *J. Symbolic Computation* 5(1–2), 65–108.

Grigor'ev, D.Y., and Vorobjov, N.N. 1988. Solving systems of polynomial inequalities in subexponential time. *J. Symbolic Computation* 5(1–2), 37–64.

Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T., and Clore, G.M. 1991. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* 253, 657.

Harris, R. 2002. The Ubiquitin NMR Resource Page, BBSRC Bloomsbury Center for Structural Biology, *www.biochem.ucl.ac.uk/bsm/nmr/ubq/index.html*.

Hoch, J., Burns, M.M., and Redfield, C. 1990. *Frontiers of NMR in Molecular Biology*, 167–175, Alan R. Liss, NY.

Hus, J.C., Marion, D., and Blackledge, M. 2000. *De novo* determination of protein structure by NMR using orientational and long-range order restraints. *J. Mol. Biol.* 298(5), 927–936.

Hus, J.C., Prompers, J., and Brüschweiler, R. 2002. Assignment strategy for proteins of known structure. *J. Mag. Res.* 157, 119–125.

Johnson, E.C., Lazar, G.A., Desjarlais, J.R., and Handel, T.M. 1999. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. *Structure Fold Des.* 7, 967–976.

Johnson, M.S., Srinivasan, N., Sowdhamini, R., and Blundell, T.L. 1994. Knowledge-based protein modeling. *Mol. Biochem.* 29, 1–68.

Kuhn, H.W. 1955. Hungarian method for the assignment problem. *Nav. Res. Logist. Quarterly* 2, 83–97.

Kullback, S., and Leibler, R.A. 1951. On information and sufficiency. *Annals Math. Statist.* 22, 79–86.

Kurinov, I.V., and Harrison, R.W. 1995. The influence of temperature on lysozyme crystals—structure and dynamics of protein and water. *Acta Crystal. D Biol. Crystal.* 51, 98–109.

Kuszewski, J., Gronenborn, A.M., and Clore, G.M. 1999. Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. *J. Am. Chem. Soc.* 121, 2337–2338.

Langmead, C., and Donald, B.R. 2003. 3D Structural homology detection via unassigned residual dipolar couplings, *Proc. IEEE Computer Society Bioinformatics Conference (CSB)*, Stanford CA. (August 10, 2003), pp. 209–217. ISBN 0-7695-2000-6.

Langmead, C., and Donald, B.R. 2004a. An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J. Biomol. NMR* 29, 111–138.

Langmead, C., and Donald, B.R. 2004b. An improved nuclear vector replacement algorithm for nuclear magnetic resonance assignment. Dartmouth Computer Science Technical Report TR2004-494, Hanover, NH. *http://www.cs.dartmouth.edu/reports/abstracts/TR2004-494/*

Langmead, C., and Donald, B.R. 2004c. High-throughput 3D homology detection via NMR resonance assignment. Dartmouth Computer Science Technical Report TR2004-487, Hanover, NH. *http://www.cs.dartmouth.edu/reports/abstracts/TR2004-487/*

Lathrop, R.H., and Smith, T.F. 1996. Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.* 255, 641–665.

Lim, K., Nadarajah, A., Forsythe, E.L., and Pusey, M.L. 1998. Locations of bromide ions in tetragonal lysozyme crystals. *Acta Crystal. D Biol. Crystal.* 54, 899–904.

Losonczi, J.A., Andrec, M., Fischer, W.F., and Prestegard, J.H. 1999. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J. Magn. Reson.* 138(2), 334–342.

Meiler, J., Peti, W., and Griesinger, C. 2000. DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts. *J. Biom. NMR* 17, 283–294.

Mueller, G.A., Choy, W.Y., Yang, D., Forman-Kay, J.D., Venters, R.A., and Kay, L.E. 2000. Global folds of proteins with low densities of NOEs using residual dipolar couplings: Application to the 370-residue maltodextrin-binding protein. *J. Mol. Biol.* 300, 197–212.

National Institute of General Medical Sciences. 2002. The Protein Structure Initiative. URL: *www.nigms.nih.gov/psi/*.

Oki, H., Matsuura, Y., Komatsu, H., and Chernov, A.A. 1999. Refined structure of orthorhombic lysozyme crystallized at high temperature: Correlation between morphology and intermolecular contacts. *Acta Crystal. D Biol. Crystal.* 55, 114.

Osapay, K., and Case, D.A. 1991. A new analysis of proton chemical shifts in proteins. *J. Am. Chem. Soc.* 113, 9436–9444.

Palmer III, A.G. 1997. Probing molecular motion by NMR. *Curr. Opin. Struct. Biol.* 7, 732–737.

Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structures and energies of molecules. *Comp. Phy. Comm.* 91, 1–41.

Ramage, R., Green, J., Muir, T.W., Ogunjobi, O.M., Love, S., and Shaw, K. 1994. Synthetic, structural and biological studies of the ubiquitin system: The total chemical synthesis of ubiquitin. *J. Biochem.* 299, 151–158.

Redfield, C., Hoch, J., and Dobson, C. 1983. Chemical shifts of aromatic protons in protein NMR spectra. *FEBS Lett.* 159, 132–136.

Rohl, C.A., and Baker, D. 2002. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* 124(11), 2723–2729.

Rossman, M.G., and Blow, D.M. 1962. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystal.* 15, 24–31.

Sali, A., Overington, J.P., Johnson, M.S., and Blundell, T.L. 1990. From comparisons of protein sequences and structures to protein modelling and design. *Trends Biochem. Sci.* 15, 235–240.

Saupe, A. 1968. Recent results in the field of liquid crystals. *Angew. Chem.* 7, 97–112.

Schneider, D.M., Dellwo, M.J., and Wand, A.J. 1992. Fast internal main-chain dynamics of human ubiquitin. *Biochemistry* 31(14), 3645–3652.

Schwalbe, H., Grimshaw, S.B., Spencer, A., Buck, M., Boyd, J., Dobson, C.M., Redfield, C., and Smith, L.J. 2001. A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Sci.* 10, 677–688.

Shuker, S.B., Hajduk, P.J., Meadows, R.P., and Fesik, S.W. 1996. Discovering high affinity ligands for proteins: SAR by NMR. *Science* 274, 1531–1534.

Tian, F., Valafar, H., and Prestegard, J.H. 2001. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. *J. Am. Chem. Soc.* 123, 11791–11796.

Tjandra, N., and Bax, A. 1997. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 278, 1111–1114.

Vaney, M.C., Maignan, S., Ries-Kautt, M., and Ducruix, A. 1996. High-resolution structure (1.33 Angstrom) of a HEW lysozyme tetragonal crystal grown in the APCF apparatus. Data and structural comparison with a crystal grown under microgravity from SpaceHab-01 mission. *Acta Crystal. D Biol. Crystal.* 52, 505–517.

Vijay-Kumar, S., Bugg, C.E., and Cook, W.J. 1987. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* 194, 531–544.

Wang, L., and Donald, B.R. 2004. 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*. In press.

Weber, P.L., Brown, S.C., and Mueller, L. 1987. Sequential 1H NMR assignments and secondary structure identification of human ubiquitin. *Biochemistry* 26, 7282–7290.

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

Wishart, D.S., Watson, M.S., Boyko, R.F., and Sykes, B.D. 1997. Automated 1H and 13C chemical shift prediction using the BioMagResBank. *J. Biomol. NMR* 10, 329–336.

Wüthrich, K. 2000. Protein recognition by NMR. *Nat. Struct. Biol.* 7(3), 188–189.

Xu, Y., Xu, D., Crawford, O.H., Einstein, J.R., and Serpersu, E. 2000. Protein structure determination using protein threading and sparse NMR data. *RECOMB'00*.

Xu, Y., Xu, D., Kim, D., Olman, V., Razumovskaya, J., and Jiang, T. 2002. Automated assignment of backbone NMR peaks using constrained bipartite matching. *IEEE Comput. Sci. Eng.* 4(1), 50–62.

Yan, A., Langmead, C., and Donald, B.R. 2003. A probability-based similarity measure for saupe alignment tensors with applications to residual dipolar couplings in NMR structural biology. Dartmouth Computer Science Technical Report TR2003-474, Hanover, NH. *http://www.cs.dartmouth.edu/reports/abstracts/TR2003-474/*

Zweckstetter, M., and Bax, A. 2001. Single-step determination of protein substructures using dipolar couplings: Aid to structural genomics. *J. Am. Chem. Soc.* 123(38), 9490–9491.

Address correspondence to:
*Bruce Randall Donald*
*6211 Sudikoff Laboratory*
*Dartmouth Computer Science Department*
*Hanover, NH 03755*

*E-mail:* brd@cs.dartmouth.edu