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

Christopher James Langmead[*] **Anthony Yan**[*] **Ryan Lilien**[*]

Lincong Wang[*] **Bruce Randall Donald**[*,†,‡,§,¶]

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 an 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-labelling 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 average assignment accuracy of over 90%. We further demonstrate the feasibility of our algorithm for different and larger proteins, using NMR data for hen lysozyme (129 residues, 98% accuracy) and streptococcal protein G (56 residues, 95% 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.

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.

[*]Dartmouth Computer Science Department, Hanover, NH 03755, USA.

[†]Dartmouth Chemistry Department, Hanover, NH 03755, USA.

[‡]Dartmouth Department of Biological Sciences, Hanover, NH 03755, USA.

[§]Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: `brd@cs.dartmouth.edu`

[¶]This work is supported by the following grants to B.R.D.: National Institutes of Health (R01 GM-65982), National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, and EIA-9802068), and the John Simon Guggenheim Foundation.

RECOMB'03, April 10–13, 2003, Berlin, Germany.

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 modelling 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 [44] 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 modelling, because NMR can be used to probe protein-protein interactions [21] (via chemical shift mapping [12]), protein-ligand binding (via SAR by NMR [57] or line-broadening analysis [19]), and dynamics (via, e.g., nuclear spin relaxation analysis [47]). By analogy, in X-ray crystallography, the molecular replacement (MR) technique [52] 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

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

Table 1: NVR Experiment Suite: The 5 *unassigned* NMR spectra used by NVR to perform resonance assignment and structure refinement. 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 non-hydrogen bonded and can be correlated to the structural model. A sparse number (< 1 per residue) of d_{NN}s can be obtained from the NOESY. These d_{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.

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 [1, 32] 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^α, etc.), many days of spectrometer time, and use less efficient algorithms. In contrast, NVR requires only amide bond vector RDCs. Furthermore, NVR requires no triple-resonance experiments, and uses only ^{15}N-labelling, which is an order of magnitude less expensive than ^{13}C-labelling. 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 average assignment accuracy of over 90%. In other words, we did not fit the data to a model determined or refined by that same data. 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.

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) [59] provide *global* orientational restraints on internuclear vectors[1] (these global restraints are often termed *"long-range"* in the literature). For a good introduction to RDCs see [54]. For each RDC D, we have

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

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 3×3 *Saupe* order matrix, or alignment tensor specifying the orientation of the molecule in the laboratory frame [54]. $\mathbf{S}$ is a symmetric, traceless, rank 2 tensor with 5 degrees of freedom, which describes the average substructure alignment in the dilute

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

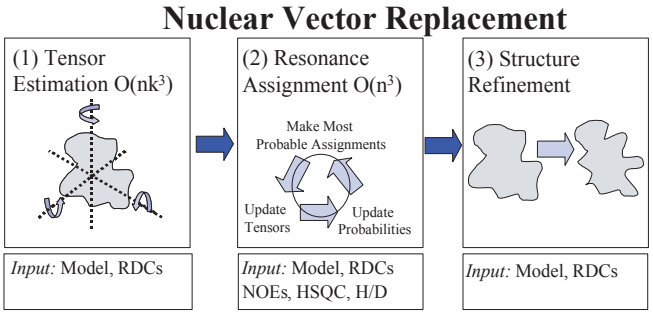


Figure 1: Nuclear Vector Replacement. **Schematic of the NVR algorithm for resonance assignment. The NVR algorithm takes as input a model of the target protein and several unassigned spectra, including the ^{15}N-HSQC, H^{N}-^{15}N RDC, ^{15}N-HSQC NOESY, and an H-D exchange-HSQC to measure amide exchange 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.**

liquid crystalline phase [41]. The measurement of five or more RDCs in substructures of known geometry allows determination of $\mathbf{S}$. Furthermore, using Eq. (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).

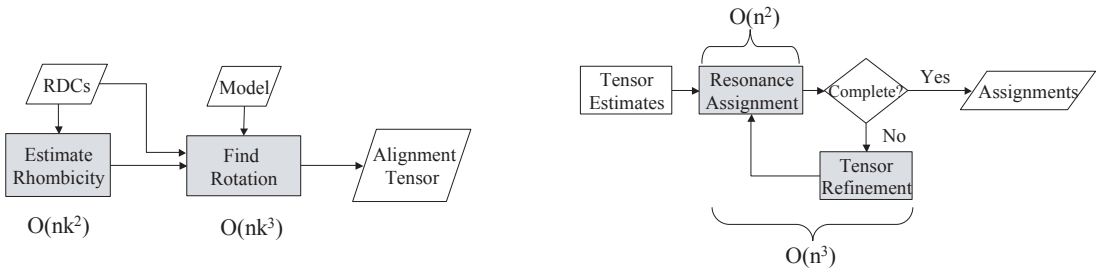
It is reasonable, in principle, to cast the problem of resonance assignment of a known structure using RDCs, into a combinatorial optimization framework [32]. Hence, initially, we attempted to treat the problem as an optimal bipartite matching problem. The use of multiple tensors for interpreting RDCs is a standard technique. Given estimates for the two alignment tensors, a bipartite graph was constructed between peaks and residues. Each edge weight was the difference between the observed RDC for a given peak, and the back-computed RDC for a given residue. A maximum bipartite matching algorithm [35] was implemented to compute the matching that minimized the sum of the edge weights in the bipartite graph. Interestingly, the matching that minimizes these weights, is typically not the correct matching. In experiments on experimental RDC data from human ubiquitin matched to 4 different structural models, maximum bipartite matchings contained, on average, only 25% correct assignments, and no higher than 40% (Table 4). Noise in the RDC help to explain these results. Experimentally recorded residual dipolar couplings 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. To overcome the uncertainty introduced by these deviations, NVR incorporates the additional, independent geometric constraints contained in amide exchange rates and NOEs.

NOE distance restraints are extracted from the d_{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_{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_{NN}s, from an unassigned 3D ^{15}N-NOESY spectrum [30]. This amounts to fewer than 0.5 d_{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_{NN}s are interpreted as geometric constraints, as follows: If a particular spin system i has a d_{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 non-hydrogen bonded surface amide protons in the model.

3 Prior Work

Assigned RDCs have previously been employed by a variety of structure refinement [14] and structure determination methods, [31, 3, 63] including: orientation and placement of secondary structure to determine protein folds [22], pruning an homologous structural database [4, 42], *de novo* structure determination [51], in combination with a sparse set of assigned NOE's to determine the global fold [43], and a method developed by Bax and co-workers for fold determination that selects heptapeptide fragments best fitting the assigned RDC data [17]. Bax and co-workers 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 [67, 17]. In another study, RDCs were used by Prestegard and coworkers [58] 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 [58]) require ^{13}C-labelling and RDCs from many different internuclear vectors (for example, ^{13}C$'$-^{15}N, ^{13}C$'$-H^{N},



Tensor Estimation Resonance Assignment

Figure 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.

$^{13}C^{\alpha}$-H^{α}, etc.). The CAP method for small RNA assignment [1], also requires ^{13}C-labelling and many RDCs in addition to many through-bond, triple resonance experiments. More recently, Brüschweiler and co-workers [32] 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 $\mathbf{S}$ 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-labelling (cf. Wüthrich: [65] *"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-labelling 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 labelling: Prestegard and co-workers [58] 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.

From a computational standpoint, NVR adopts a minimalist approach [7], demonstrating the large amount of information available in a few key spectra. By eliminating the need for triple resonance experiments, NVR saves several 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 [1] 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 time $O(In^3)$, where $O(n^3)$ is the complexity of bipartite matching [35] and I is the number of times that the Kuhn-Munkres matching algorithm is called. [32] does 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 ofthe discrete grid search for the principal order frame over $SO(3)$ (using Euler angles α, β and γ). 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[2] 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 $\mathbf{S}$ in Eq. (1). Macromolecules align differently in different liquid crystals, thus $\mathbf{S}_1$ and $\mathbf{S}_2$ are different 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 3×3 matrix with five degrees of freedom. The five degrees of freedom correspond to three Euler angles (α, β and γ), describing the average partial 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 [41]. If the resonance assignments are not known, as in our case, these parameters must be estimated. It has been shown [41] that D_a and D_r can be decoupled from the Euler angles by diagonalizing the alignment tensor:

[2]For the purpose of exposition, we will refer specifically to bicelle and phage aligning media, as per the data we processed [15, 38, 56]. NVR, however, would work on residual dipolar couplings recorded in other media as well (e.g., stretched polyacrylamide gels [13]).

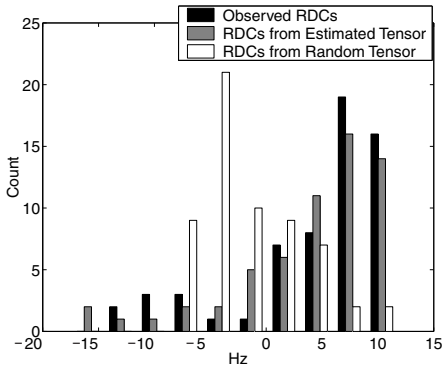


Figure 3: Distributions of Dipolar Couplings. A comparison of the distributions of dipolar couplings generated from 3 different alignment tensors. The black bars are the distribution of observed RDCs for human ubiquitin in the bicelle medium. The grey bars are the distribution of RDCs generated by the tensor estimated by NVR using 1UBI as a model. The black and grey 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 grey distributions.

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

Here, $\mathbf{V} \in SO(3)$ is a 3×3 rotation matrix[3] that defines a coordinate system called the *principal order frame.* Σ is a 3×3 diagonal and traceless matrix containing the eigenvalues of $\mathbf{S}$. The diagonal elements of Σ encode D_a and D_r: $D_a = \frac{S_{zz}}{2}, D_r = \frac{S_{xx}-S_{yy}}{3}$ where $S_{yy} < S_{xx} < S_{zz}$. S_{yy}, S_{xx} and S_{zz} are the diagonal elements of Σ 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 [63]. 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 Σ in Eq. (2) can be constructed using the relationship [41, 63] between the D_a and D_r and the diagonal elements of Σ. Next, the Euler angles α, β and γ of the principal order frame are estimated by considering rotations of the model. Given Σ (Eq. 2), for each rotation $V(\alpha, \beta, \gamma)$ of the model, a new Saupe matrix $\mathbf{S}$ is computed using Eq. (2). That matrix $\mathbf{S}$ is used to compute a set of back-computed RDCs using Eq. (1). The relative entropy, also known as the Kullback-Leibler distance [36], 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 [63] 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 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.

Although the initial tensor estimates are not perfect, they are accurate enough to bootstrap the second phase, *resonance assignment*, described below. For example, differences of up to 20° between the actual and estimated Euler angles were seen for one of our test proteins (Fig. 4). The magnitude of these deviations can be interpreted geometrically in terms of surface area on the unit sphere. The surface area of a region on the unit sphere enclosed by a latitudinal circle drawn η degrees from the North pole is $\int_0^\eta 2\pi \sin\theta d\theta$. Hence, the set of all deviations $\leq 20°$ represent only 3% of the total surface area of unit sphere (4π). Due to the symmetry of the dipolar operator, one must double that area. Still, relative to the distribution of possible errors, a 20° angular deviation falls into the 94 percentile of accuracy.

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 Eq. (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, and B_m be the set of back-computed RDCs using the model and $\mathbf{S}_m$. For each medium m, a 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 $r \in R$, $d_m(q) \in D_m$, $b_m(r, B_m) \in B_m$. 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 μ_m and standard deviation σ_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 reside 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_{NN} or amide exchange.

[3]While any representation of rotations may be employed, we use Euler angles (α, β, γ).

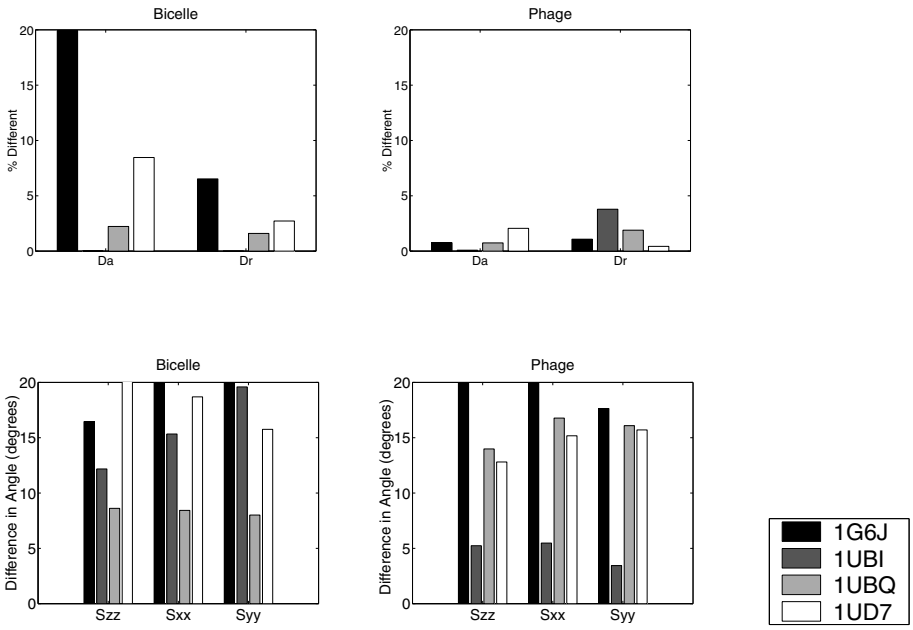


Figure 4: Ubiquitin Tensor Estimates. These panels demonstrate the accuracy of the first step of the NVR algorithm where two tensors are estimated, one for the bicelle medium, and one for the phage medium. (Upper Left Panel) Percentage difference for the axial and rhombic terms, D_a and D_r, for the four models, 1G6J, 1UBI, 1UBQ and 1UD7, vs. the actual axial and rhombic terms in the bicelle medium. (Lower Left Panel) Angular differences (in degrees) between the eigenvectors of the estimated tensors and the eigenvectors of the actual tensors in the bicelle medium. S_{zz} is the director of the tensor (i.e., the eigenvector associated with the largest eigenvalue of the tensor), S_{xx} and S_{yy} are eigenvectors associated with the second largest and smallest eigenvalue of the tensor, respectively. (Upper Right Panel, Lower Right Panel) Accuracy of the tensor estimates in the phage medium. Differences in the orientation of the eigenvectors are as large as 20°. However, angular deviations of 20° represent only 3% of the total surface area of the unit sphere (see text).

On each iteration, the probabilities of assignment are (re)computed using Eq. (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:

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

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 [41]. The tensors, which were coarsely estimated in phase 1 of NVR, begin to converge to their true values with each iteration.[4] At the end of phase 2, the principal axes of the final tensor estimates are typically within one degree, and the axial and rhombic components are within 1-2% of their correct values, respectively.

Intuitively, NVR only makes assignments that are a) unambiguous and b) consistent across both media. Figure 5 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 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 6).

The computational complexity of the second phase is as follows. $\mathbf{M}_1$ and $\mathbf{M}_2$ are each of size $O(n \times n)$, where n is the number of residues in the protein. Re-computing the tensors, using the Moore-Penrose pseudo-inverse of the $O(n) \times 5$ matrix takes time $O(n^2)$ [25]. 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 complete 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 Eq. (4) cannot be satisfied. This only occurs on the last few iterations when, for example, the remaining 2 peaks each vote for the same residue. NVR handles this case by performing a maximum bipartite matching [35] for those peaks, and the second phase terminates. This does not increase the time-complexity. As previously mentioned, bipartite matching did not perform well when run on all n residues and $O(n)$ peaks: we only use it in the endgame to resolve the very small number of remaining assignments that Eq. (4) cannot disambiguate.

4.3 Structure Refinements (Phase 3)

Once the final set of assignments has been computed, the (now)

[4]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[15, 38, 56] and/or (b) computing the optimal Saupe matrix using the correct assignments.

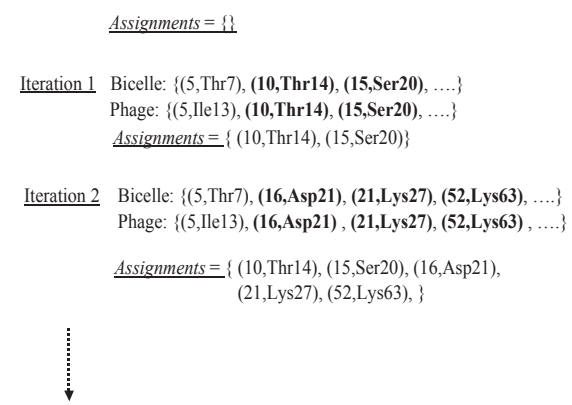


Figure 5: Iterative Assignments. The first two iterations of NVR with model 1UBQ. The assignment list is initially empty. At the end of the first iteration, both the phage and bicelle media "agree" that peaks 10 and 15 are residues Thr14 and Ser20, respectively. Consequently, those two assignments are added to the master assignment list. Note, there are only 2 assignments so there are not enough variables to update the tensors, S_1 and S_2, using Eq. (1). At the beginning of the 2nd 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 the phage and bicelle 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 5 assignments so S_1 and S_2 can be updated using Eq. (1). This procedure continues until the entire protein is assigned.

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. T is used to refine the structure. A Monte-Carlo algorithm was implemented to find a (new) conformation of the model's ϕ and ψ 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) 1 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 [48]. This minimization is done *in vacuo*. Figure 7 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

The molecular structure of human ubiquitin has been investigated extensively. A variety of data have been published including resonance assignments [62, 55], backbone amide residual dipolar couplings recorded in two separate liquid crystals (bicelle and phage) [15], amide-exchange rates [15], ^{15}N-HSQC and ^{15}N-HSQC NOESY spectra [30], and several independent high-resolution structures solved by both X-ray crystallography [49, 61] and NMR [6, 33]. In 1998, the Bax lab published a new NMR structure for ubiquitin, (PDB Id: 1D3Z) [15]. 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. 1G6J, 1UBI and 1UBQ have 100% sequence identity to 1D3Z. 1UD7 is mutant of ubiquitin where 7 hydrophobic core residues have been altered (I3V, V5L, I13V, L15V, I23F, V26F, L67I). 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 [15] for human ubiquitin were fit to the amide bond vectors of the selected model. ^{15}N-HSQC and ^{15}N-HSQC NOESY spectra [30] were processed to extract sparse, unassigned d_{NN}s.

NVR achieves an average of over 90% accuracy for the four ubiquitin models (Table 2 A). The accuracies on NMR data for two additional proteins, the B1 domain of streptococcal protein G and lysozyme, were over 95% (See Table 2 B-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. Of all these errors, all but 1% involved dipolar couplings that were very different from their expected values. For example, in the trial on ubiquitin model 1G6J, Ser20 was mistaken for Gln49 and vice-versa. The observed dipolar couplings for these two residues were an average of 7.9 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 2.4 Hz.

There were only two cases, from our 20 separate trials, where a small chain of misassignments was seen. Both were from the trial on the lysozyme model 1LYZ. The following two chains were observed: Gly49 → Ser50 → Gly102 → Cys127 → Gly49 and Ser72 →Trp123 → Arg73 → Ser72. These cyclic errors are probably due to the relatively poor initial estimates for the alignment tensors (data not shown). We are presently extending the NVR method to include ^{1}H and ^{15}N chemical shift prediction [46, 64] to determine whether accurate chemical shift prediction will prevent these kinds of errors. Brüschweiler and co-workers describe a similar chain (cyclic permutation) of errors [32] for the one protein (1UBI) on which HPB was tested (Thr9 → Arg74 → Tyr59 → Gly53). NVR found no cyclic permutation of length longer than 2, for any ubiq-

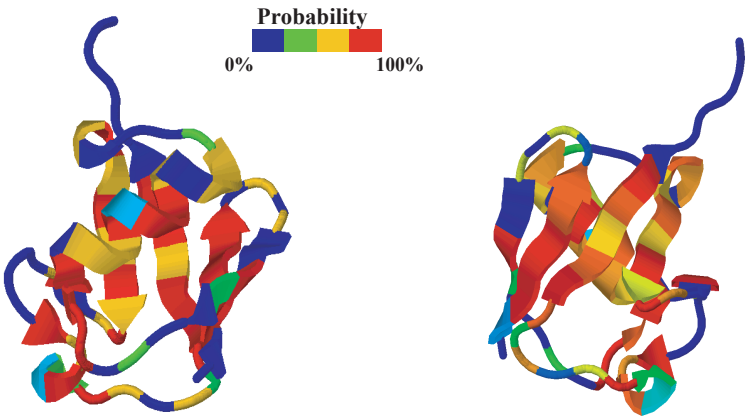
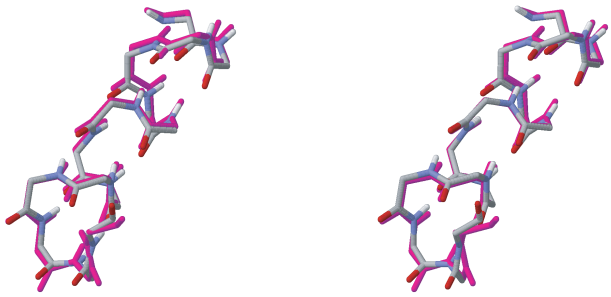


Figure 6: Assignment Confidences. **NVR returns the *confidence* of each assignment. Here the structure of ubiquitin model 1UBQ (shown in two different orientations) is annotated with the confidence of each assignment. The color depicts the confidence with which the backbone amide group was assigned. 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.**



Residues 22-35 *before* refinement Residues 22-35 *after* refinement

Figure 7: 1G6J Structure Refinement. **In magenta, the backbone of residues 22-35 from the structure 1D3Z. These residues form the first α-helix in ubiquitin. 1D3Z is an RDC-refined model. In CPK-coloring, the backbone of residues 22-35 of model 1G6J (on the left) and a new structure (on the right) generated after structure refinement of 1G6J (using the RDC assignments from NVR). The RMSD between the 2 backbones on the right is 11% smaller than the RMSD of the backbones on the left.**

uitin model, including 1UBI.

There was one case where a mistake was made involving a "degenerate" pair of NH vectors (residues). In the trial on 1UBI, Ile23 was mistaken for His68. The angle between amide bond vectors from those residues is only 3.4°. Consequently, there was only a 0.35 Hz difference in the expected dipolar couplings under both media. The resolution of RDC is at best 0.2 Hz [50], and can be worse, making these internuclear vector orientations hard to distinguish.

In a separate set of trials, we used the final tensors generated in our first trials to bootstrap the resonance assignment phase of NVR. Overall, an increase in accuracy of 1% was seen. Additional iterations yielded no substantial improvement in accuracy. This suggests that the resonance assignment phase is stable with respect to the particular tensor estimate.

5.1 3D Structural Homology Detection

We have also extended NVR to a second application area —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 only considers the backbone geometry of each amino acid in the model, not the identity 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 individual 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 [10] 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 5 structural homologs for each of these three proteins (Table 3). These homologs have between 10-61% sequence ho-

PDB ID	Exp. Method	Accuracy
1G6J [6]	NMR	90
1UBI [49]	X-ray (1.8 Å)	90
1UBQ [61]	X-ray (1.8 Å)	93
1UD7 [33]	NMR	93

(A: Ubiquitin)

PDB ID	Exp. Method	**Accuracy**
193L [60]	X-ray (1.3 Å)	100%
1AKI [5]	X-ray (1.5 Å)	100%
1AZF [40]	X-ray (1.8 Å)	100%
1BGI [45]	X-ray (1.7 Å)	100%
1H87 [24]	X-ray (1.7 Å)	98%
1LSC [37]	X-ray (1.7 Å)	100%

(C: Lysozyme)

PDB ID	Exp. Method	**Accuracy**
1GB1 [29]	NMR	95%
2GB1 [29]	NMR	95%
1PGB [23]	X-ray (1.92 Å)	95%

(B: SPG)

PDB ID	Exp. Method	**Accuracy**
1LYZ [18]	X-ray (2.0 Å)	91%
2LYZ [18]	X-ray (2.0 Å)	98%
3LYZ [18]	X-ray (2.0 Å)	100%
4LYZ [18]	X-ray (2.0 Å)	96%
5LYZ [18]	X-ray (2.0 Å)	98%
6LYZ [18]	X-ray (2.0 Å)	98%

(D: Lysozyme (cont))

Table 2: Accuracy. (A) NVR achieves an average accuracy of over 90% on the four ubiquitin models. The structure 1D3Z [15] is the only published structure of ubiquitin to have been refined against RDCs. The RDCs used in [15] have been published and were used in each of the 4 NVR trials. 1G6J, 1UBI and 1UBQ have 100% sequence identity to 1D3Z. 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 [38] and hen lysozyme [56] were obtained from the PDB. NOEs and amide exchange data were extracted from their associated restraints files. NVR achieves an average of 95% (Table B) and 98% (Tables C and D).

mology to 1D3Z, 1E8L and 3GB1. The database contains only the backbone geometry, the length of the primary sequence, and the percentage of α and β 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 [16]. 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 (Sec. 4.1) against each model in the database. Note that the tensor estimation phase does not require NOEs nor 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 α_t and β_t are the predicted percentages of α and β structure for the target protein, t, and α_m and β_m are the actual percentages of α and β structure taken from the model, m. Let Δ_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] (Sec. 4.1). A model's score is computed as follows:

$$I_m = \Delta_\alpha + \Delta_\beta + \Delta_l + KL_1 + KL_2. \quad (5)$$

Each model is then ranked according to its score. As seen in Table 3, the highest ranking structure is the native structure. The five homologous structures are also highly ranked. Figure 8 is a scatterplot of the scores computed by NVR vs. the backbone RMSD of 1D3Z to all the models in the database. The native and homologous structures form a cluster. Thus, NVR is able to identify structural homologies between proteins with remote amino acid sequences.

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-labelling 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 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 efficiently assigns the ^{15}N-HSQC spectrum as well as the d_{NN}s of the 3D ^{15}N-NOESY spectrum, in $O(n^3)$ time. We tested NVR on NMR data from 3 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 average assignment accuracy of over 90%. Similarly good results were obtained on NMR data for streptococcal protein G (95%) and hen lysozyme (98%) 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 are presently extending it to handle the case when either the set of resonances or RDCs are incomplete. We are also exploring incorporating ^{1}H and ^{15}N chemical shift prediction [46, 64] for NVR.

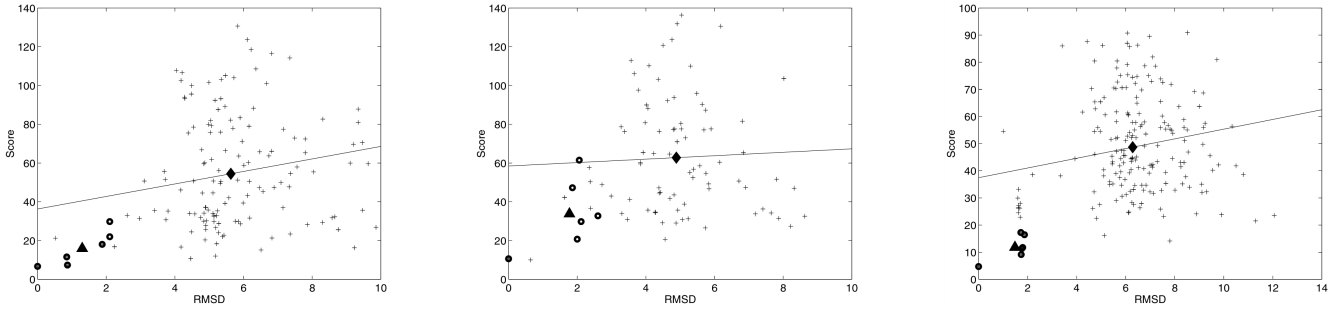
Finally, we have demonstrated that NVR can be used to identify 3D structural homologies between remote amino acid sequences. Furthermore, our success in assigning 1UD7, which is a mutant of ubiquitin, suggests that NVR could be applied more broadly to assign spectra based on homologous structures. Using the results

[5] Δ_α and Δ_β are multiplied by 100 so that they have the same order of magnitude as Δ_l, KL_1, and KL_2

PDB ID	Homolog	Sequence Identity	RMSD	Rank
1D3Z		100%	0Å	1
	1NDD	55.6%	0.6Å	2
	1BT0	61.0%	0.7Å	3
	1H8C	15.7%	1.9Å	11
	1GUA	11.6%	2.1Å	19
	1C1Y	11.6%	2.1Å	38

PDB ID	Homolog	Sequence Identity	RMSD	Rank
1E8L		100%	0Å	1
	2EQL	49.2%	1.8Å	2
	1ALC	35.8%	1.8Å	3
	1HFZ	38.3%	1.8Å	4
	1A4V	38.2%	1.8Å	5
	1F6S	38.7%	1.7Å	6
3GB1		100%	0Å	1
	1HZ5	14.5%	2.2Å	2
	1JML	12.8%	1.8Å	5
	1HEZ	12.7%	2.0Å	12
	2GCC	10.0%	2.6Å	24
	1HZ6	14.5%	2.2Å	55

Table 3: Structural Homology Detection Results. The sequence identity and RMSD of the 3 test proteins and their respective 5 homologs. The final column is the rank of that model (out of 2546 structures), based on the score computed by NVR.



Ubiquitin (1D3Z) SPG (3GB1) Lysozyme (1E8L)

Figure 8: RMSD vs NVR Homology Score. 3 Scatter plots of the backbone RMSD between the native structures of Ubiquitin (left), SPG (center) and Lysozyme (right) and the models in the database vs 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 non-homologous 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. The scores associated with the native fold and the 5 homologs are statistically significantly lower than the scores of unrelated proteins (p-values of 2.6×10^{-5}, 2.3×10^{-5}, and 2.9×10^{-5} for 1D3Z, 1E8L, and 3GB1, respectively).

of a sequence alignment algorithm [2], protein threading [39, 66], or homology modelling [11, 20, 26, 34, 53], 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.

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.

8. REFERENCES

[1] AL-HASHIMI, H.M. AND GORIN, A. AND MAJUMDAR, A. AND GOSSER, Y. AND PATEL, D.J. Towards Structural Genomics of RNA: Rapid NMR Resonance Assignment and Simultaneous RNA Tertiary Structure Determination Using Residual Dipolar Couplings. *J. Mol. Biol. 318* (2002), 637–649.

[2] ALTSCHUL, S.F. AND GISH, W. AND MILLER, W. AND MYERS, E.W. AND LIPMAN, D.J. Basic local alignment search tool. *J. Mol. Biol. 215* (1990), 403–410.

[3] 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 (2001), 335–347.

[4] ANNILA, A. AND AITIO, H. AND THULIN, E. AND DRAKENBERG. T. Recognition of protein folds via dipolar couplings. *J. Biom. NMR 14* (1999), 223–230.

[5] ARTYMIUK, P. J. AND BLAKE, C. C. F. AND RICE, D. W. AND WILSON, K. S. The Structures of the Monoclinic and Orthorhombic Forms of Hen Egg-White Lysozyme at 6 Angstroms Resolution. *Acta Crystallogr B Biol Crystallogr 38* (1982), 778.

[6] BABU, C. R. AND FLYNN, P. F. AND WAND, A. J. Validation of Protein Structure from Preparations of Encapsulated Proteins Dissolved in Low Viscosity Fluids. *J.Am.Chem.Soc. 123* (2001), 2691.

[7] BAILEY-KELLOGG, C. AND WIDGE, A. AND KELLEY III, J.J. AND BERARDI, M.J. AND BUSHWELLER, J.H. AND DONALD, B.R. The NOESY Jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J Comput Biol 7*, 3-4 (2000), 537–58.

[8] BASU, S. An Improved Algorithm for Quantifier Elimination Over Real Closed Fields. *IEEE FOCS* (1997), 56–65.

[9] BASU, S. AND ROY, M.F. On the combinatorial and algebraic complexity of quantifier elimination. *Journal of the ACM (JACM) 43*, 6 (1996), 1002–1045.

[10] 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* (2000), 235–242.

[11] BLUNDELL, T.L. AND SIBANDA, B.L. AND STERNBERG, M.J. AND THORNTON, J.M. Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules. *Nature 326* (1987), 347–352.

PDB ID	Accuracy		
	Maximum Bipartite Matching	NVR with RDC and Amide Exchange	NVR with RDC and NOE
1G6J [6]	7%	37%	72%
1UBI [49]	25%	65%	73%
1UBQ [61]	40%	42%	85%
1UD7 [33]	28%	18%	65%

Table 4: Ubiquitin: Comparison of Assignment Algorithms. 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 2 (A).

[12] CHEN, Y. AND REIZER, J. AND SAIER JR., M. H. AND FAIRBROTHER, W. J. AND WRIGHT, P. E. Mapping of the binding interfaces of the proteins of the bacterial phosphotransferase system, HPr and IIAglc. *Biochemistry 32*, 1 (1993), 32–37.

[13] CHOU, J.J AND GAEMERS, S. AND HOWDER, B. AND LOUIS, J.M. AND BAX, A. A simple apparatus for generating stretched polyacrylamide gels, yielding uniform alignment of proteins and detergent micelles. *J. Biom. NMR 21*, 4 (2001), 377–82.

[14] CHOU, J.J AND LI, S. AND BAX, A. Study of conformational rearrangement and refinement of structural homology models by the use of heteronuclear dipolar couplings. *J. Biom. NMR 18* (2000), 217–227.

[15] CORNILESCU, G. AND MARQUARDT, J. L. AND OTTIGER, M. AND BAX, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J.Am.Chem.Soc. 120* (1998), 6836–6837.

[16] CUFF, J.A. AND CLAMP, M.E. AND SIDDIQUI, A.S. AND FINLAY, M. AND BARTON, G.J. Jpred: A Consensus Secondary Structure Prediction Server. *Bioinformatics 14* (1998), 892–893.

[17] DELAGLIO, F. AND KONTAXIS, G. AND BAX, A. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc 122* (2000), 2142–2143.

[18] DIAMOND, R. Real-space refinement of the structure of hen egg-white lysozyme. *J Mol. Biol. 82* (1974), 371–391.

[19] FEJZO, J. AND LEPRE, C.A. AND PENG, J.W. AND BEMIS, G.W. AND AJAY AND MURCKO, M.A. AND MOORE, J.M. The SHAPES strategy: An NMR-based approach for lead generation in drug discovery. *Chem. and Biol. 6* (1999), 755–769.

[20] FETROW, J.S. AND BRYANT, S.H. New Programs for Protein Tertiary Structure Prediction. *Bio/Technology 11* (1993), 479–484.

[21] FIAUX, J. AND BERTELSEN, E. B. AND HORWICH, A. L. AND WÜTHRICH, K. NMR analysis of a 900K GroELGroES complex. *Nature 418* (2002), 207 – 211.

[22] FOWLER, C.A. AND TIAN, F. AND AL-HASHIMI, H. M. AND PRESTEGARD, J. H. Rapid Determination of Protein Folds Using Residual Dipolar Couplings. *J. Mol. Bio 304*, 3 (2000), 447–460.

[23] GALLAGHER, T. AND ALEXANDER, P. AND BRYAN, P. AND GILLILAND, G. L. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry 33* (1994), 4721–4729.

[24] GIRARD, E. AND CHANTALAT, L. AND VICAT, J. AND KAHN, R. 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 Crystallogr D Biol Crystallogr. 58* (2001), 1–9.

[25] GOLUB, G.H. AND VAN LOAN, C.F. *Matrix Computations*, 3rd ed. The Johns Hopkins Univeristy Press, 2715 North Charles Street, Baltimore, Maryland 21218-4319, 1996, ch. 5, pp. 253–254.

[26] GREER, J. Comparative Modeling of Homologous Proteins. . *Meth. Enzymol. 202* (1991), 239–252.

[27] GRIGOR'EV, D.Y. Complexity of deciding Tarski algebra. *Journal of Symbolic Computation 5*, 1-2 (February/April 1988), 65–108.

[28] GRIGOR'EV, D.Y. AND VOROBJOV,N.N. Solving systems of polynomial inequalities in subexponential time. *Journal of Symbolic Computation 5*, 1-2 (February/April 1988), 37–64.

[29] GRONENBORN, A. M. AND FILPULA, D. R. AND ESSIG, N. Z. AND ACHARI, A. AND WHITLOW, M. AND WINGFIELD, P. T. AND CLORE, G. M. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science 253* (1991), 657.

[30] HARRIS, R. The Ubiquitin NMR Resource Page. BBSRC Bloomsbury Center for Structural Biology http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/index.html, 2002.

[31] 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 (2000), 927–936.

[32] HUS, J.C. AND PROPMERS, J. AND BRÜSCHWEILER, R. Assignment strategy for proteins of known structure. *J. Mag. Res 157* (2002), 119–125.

[33] JOHNSON, E.C. AND LAZAR, G. A. AND DESJARLAIS, J. R. AND HANDEL, T. M. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. *Structure Fold Des. 7* (1999), 967–976.

[34] JOHNSON, M.S. AND SRINIVASAN, N. AND SOWDHAMINI, R. AND BLUNDELL, T.L. Knowledge-Based Protein Modeling. *Mol. Biochem. 29* (1994), 1–68.

[35] KUHN, H.W. Hungarian method for the assignment problem. *Nav. Res. Logist. Quarterly 2* (1955), 83–97.

[36] KULLBACK, S. AND LEIBLER, R. A. On Information and Sufficiency. *Annals of Math. Stats. 22* (1951), 79–86.

[37] KURINOV, I. V. AND HARRISON, R. W. The influence of temperature on lysozyme crystals - structure and dynamics of protein and water. *Acta Crystallogr D Biol Crystallogr 51* (1995), 98–109.

[38] KUSZEWSKI, J. AND GRONENBORN, A. M. AND CLORE, G. M. Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. *J. Am. Chem. Soc. 121* (1999), 2337–2338.

[39] LATHROP, R.H. AND SMITH, T.F. Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions. *J. Mol. Biol. 255* (1996), 641–665.

[40] LIM, K. AND NADARAJAH, A. AND FORSYTHE, E. L. AND PUSEY, M. L. Locations of bromide ions in tetragonal lysozyme crystals. *Acta Crystallogr D Biol Crystallogr. 54* (1998), 899–904.

[41] LOSONCZI, J.A. AND ANDREC, M. AND FISCHER, W.F. AND PRESTEGARD J.H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson 138*, 2 (1999), 334–42.

[42] MEILER, J. AND PETI, W. AND GRIESINGER, C. DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts. *J. Biom. NMR 17* (2000), 283–294.

[43] MUELLER, G.A. AND CHOY, W.Y. AND YANG, D. AND FORMAN-KAY, J.D. AND VENTERS, R.A. AND KAY, L.E. 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* (2000), 197–212.

[44] NATIONAL INSTITUTE OF GENERAL MEDICAL SCIENCES. The Protein Structure Initiative. The National Institute of General Medical Sciences, 2002. URL: `http://www.nigms.nih.gov/funding/psi.html`.

[45] OKI, H. AND MATSUURA, Y. AND KOMATSU, H. AND CHERNOV, A. A. Refined structure of orthorhombic lysozyme crystallized at high temperature: correlation between morphology and intermolecular contacts. *Acta Crystallogr D Biol Crystallogr. 55* (1999), 114.

[46] OSAPAY, K. AND CASE,D.A. A new analysis of proton chemical shifts in proteins. *J. Am. Chem. Soc. 113* (1991), 9436–9444.

[47] PALMER III, A. G. Probing Molecular Motion By NMR. *Current Opinion in Structural Biology 7* (1997), 732–737.
[48] PEARLMAN, D.A. AND CASE, D.A. AND CALDWELL, J.W. AND ROSS, W.S. AND CHEATHAM, T.E. AND DEBOLT, S. AND FERGUSON, D. AND SEIBEL, G. AND KOLLMAN, P. 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* (1995), 1–41.
[49] RAMAGE, R. AND GREEN, J. AND MUIR, T. W. AND OGUNJOBI, O. M. AND LOVE, S. AND SHAW, K. Synthetic, structural and biological studies of the ubiquitin system: the total chemical synthesis of ubiquitin. *J. Biochem 299* (1994), 151–158.
[50] 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* (1998), 9106–9107.
[51] ROHL, C.A AND BAKER, D. De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta. *J. Am. Chem. Soc. 124*, 11 (2002), 2723–2729.
[52] ROSSMAN, M.G. AND BLOW, D.M. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr 15* (1962), 24–31.
[53] SALI, A. AND OVERINGTON, J.P. AND JOHNSON, M.S. AND BLUNDELL, T.L. From Comparisons of Protein Sequences and Structures to Protein Modelling and Design. *Trends Biochem. Sci. 15* (1990), 235–240.
[54] SAUPE, A. Recent Results in the field of liquid crystals. *Angew. Chem. 7* (1968), 97–112.
[55] SCHNEIDER, D.M. AND DELLWO, M.J. AND WAND, A. J. Fast Internal Main-Chain Dynamics of Human Ubiquitin. *Biochemistry 31*, 14 (1992), 3645–3652.
[56] SCHWALBE, H. AND GRIMSHAW, S. B. AND SPENCER, A. AND BUCK, M. AND BOYD, J. AND DOBSON, C. M. AND REDFIELD, C. AND SMITH, L. J. A Refined Solution Structure of Hen Lysozyme Determined Using Residual Dipolar Coupling Data. *Protein Sci. 10* (2001), 677–688.
[57] SHUKER, S. B. AND HAJDUK, P. J. AND MEADOWS, R. P. AND FESIK, S. W. Discovering high affinity ligands for proteins: SAR by NMR. *Science 274* (1996), 1531–1534.
[58] TIAN, F. AND VALAFAR, H. AND PRESTEGARD, J. H. A Dipolar Coupling Based Strategy for Simultaneous Resonance Assignment and Structure Determination of Protein Backbones. *J. Am. Chem. Soc. 123* (2001), 11791–11796.
[59] TJANDRA, N. AND BAX, A. Direct Measurement of Distances and Angles in Biomolecules by NMR in a Dilute Liquid Crystalline Medium. *Science 278* (1997), 1111–1114.
[60] VANEY, M. C. AND MAIGNAN, S. AND RIES-KAUTT, M. AND DUCRUIX, A. 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 Crystallogr D Biol Crystallogr 52* (1996), 505–517.
[61] VIJAY-KUMAR, S. AND BUGG, C. E. AND COOK, W. J. Structure of ubiquitin refined at 1.8 A resolution. *J Mol. Biol. 194* (1987), 531–544.
[62] WEBER, P. L. AND BROWN, S. C. AND MUELLER, L. Sequential 1H NMR Assignments and Secondary Structure Identification of Human Ubiquitin. *Biochemistry 26* (1987), 7282–7290.
[63] 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* (2002), 137–151.
[64] WISHART, D.S. AND WATSON, M.S. AND BOYKO, R.F. AND SYKES, B.D. Automated 1H and 13C Chemical Shift Prediction Using the BioMagResBank. *J.Biomol. NMR 10* (1997), 329–336.
[65] WÜTHRICH, K. Protein recognition by NMR. *Nat. Struct. Biol 7*, 3 (2000), 188–189.
[66] XU, Y. AND XU, D. AND CRAWFORD, O. H. AND EINSTEIN, J. R. AND SERPERSU, E. Protein Structure Determination Using Protein Threading and Sparse NMR Data. In *Proc. RECOMB* (2000), pp. 299–307.
[67] ZWECKSTETTER, M. AND BAX, A. Single-step determination of protein substructures using dipolar couplings: aid to structural genomics. *J Am Chem Soc 123*, 38 (2001), 9490–1.

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 [27, 28, 9, 8]. Hence the $O(nk^3)$ discrete-grid rotation search in Sec. 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), $B(V) = E^T \mathbf{S} E = (E^T (V^T \Sigma V) E) = \{ \mathbf{w}^T (V^T \Sigma V) \mathbf{w} \,|\, \mathbf{w} \in E \}$. (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). Eq. (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 Eq. (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 Eq. (7) can be decided exactly, in polynomial time, using the theory of real-closed fields. We will use Grigor'ev's algorithm [27, 28] 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.