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

Christopher J. Langmead
Dartmouth College

Anthony K. Yan
Dartmouth College

Ryan Lilien
Dartmouth College

Lincong Wang
Dartmouth College

Bruce Randall Donald
Dartmouth College

Follow this and additional works at: http://repository.cmu.edu/compsci

Published In
Proceedings of The Seventh Annual International Conference on Computational Molecular Biology (RECOMB), 176-187.
A Polynomial-Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments

Christopher James Langmead*
Anthony K. 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 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 15N-labelling of the protein, and processes unassigned HN,15N HSQC spectra, HN,15N RDCs, and sparse HN-HN NOE’s (dNN8), 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 (HN,15N) backbone resonances as well as the dNN8 of the 3D 15N-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.

Abbreviations used: NMR, nuclear magnetic resonance; NVR, nuclear vector replacement; RDC, residual dipolar coupling; 3D, three-dimensional; HSQC, heteronuclear single-quantum coherence; HN, amide proton; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; dNN, 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 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 crystallography), 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 nuclear 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 [1] 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 [20] (via chemical shift mapping [12]), protein-ligand binding (via SAR by NMR [55] or line-broadening analysis [18]), and dynamics (via, e.g., nuclear spin relaxation analysis [45]). By analogy, in X-ray crystallography, the molecular replacement (MR) technique [50] 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 [2, 31] require 13C-labelling and RDCs from many different bonds (for example, 13C′-15N, 13C′-H,N, 13C′-H,O, etc.), many days of spectrometer time, and use inefficient algorithms. In contrast, NVR requires only amide bond vector RDCs. Furthermore, NVR requires no triple-resonance experiments, and uses only 15N-labelling, which is an order of magnitude less expensive than 13C-labelling. In NVR, the experimentally-measured bond vectors are conceptually “replaced” by model bond vectors to find the correct assignment. The NVR algorithm searches for the assignments that best correlate the experimental RDCs, \( d_{\text{NMR}} \) and amide exchange rates with a whole-protein 3D structural model. NVR processes unassigned HSQC, H,N-15N RDCs (in two media), amide exchange data, and 3D 15N-NOESY spectra, all of which can be acquired in about one day.

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 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 the application of our method on real biological NMR data. Finally, section 6 discusses these results. An optional appendix contains proofs, and tables of supporting information on NVR accuracy and performance.

2 Background

The experimental inputs to NVR are detailed in Table 1. Residual dipolar couplings (RDCs) [57] provide global orientational restraints on internuclear bond vectors (these global restraints are often termed “long-range” in the literature). For each RDC \( D \), we have

\[
D = D_{\text{max}} v^T S v,
\]

where \( D_{\text{max}} \) is a constant, and \( v \) is the internuclear vector orientation relative to an arbitrary substructure frame and \( S \) is the 3 x 3 Saupe order matrix [52]. \( S \) is a symmetric, traceless, rank 2 tensor with 5 degrees of freedom, which describes the average substructure alignment in the liquid crystalline phase [40]. The measurement of five or more RDCs in substructures of known geometry allows determination of \( S \). Furthermore, using Eq. (1), substructures of the protein may be oriented
relative to a common coordinate system, the principle order frame.

Once $S$ is estimated, RDCs may be simulated (back-calculated) given any other internuclear bond vector $v_j$. 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 $D_j$ be the measured RDC value corresponding to this peak. Then the RDC $D_i$ is assigned to amide bond vector $v_j$ of a known structure, and we should expect that $D_i \approx D_{\text{max}} v_j^T S v_j$ (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 [31]. Hence, initially, we attempted to treat the problem as an optimal bipartite matching problem. 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. However, the matching that minimizes these weights, is not the correct matching (see Tables 5, 6 and 7 in the appendix). One root of the problem is that experimentally recorded residual dipolar coupling 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 crystal. To overcome the uncertainty introduced by these deviations, NVR incorporates the additional, independent geometric constraints contained in amide exchange rates and NOEs.

Amide exchange rates, obtained by comparing the H-D exchange $^{15}$N-HSQC to the ordinary $^{15}$N-HSQC (in water), identify the peaks generated by non-hydrogen bonded amide protons of the surface residues. NOEs are extracted from the $d_{\text{NN}}$ region of an unassigned $^{15}$N HSQC-NOESY to provide distance restraints. NVR uses a sparse set of NOEs. By sparse, we mean a small number of unassigned NOEs. A sparse set of $d_{\text{NN}}$ can be obtained from an unassigned NOESY spectrum, once the peaks in the HSQC have been referenced to the $(H^N,^{15}N)$ coordinates of the diagonal peaks of the NOESY. In our trials on ubiquitin, for example, we obtained 34 $d_{\text{NN}}$, from an unassigned $3D^{15}$N-NOESY spectrum [29]. This amounts to fewer than 0.5 $d_{\text{NN}}$ 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, $d_{\text{NN}}$ are interpreted as geometric constraints, as follows: If a particular spin system $i$ has a $d_{\text{NN}}$ 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 5 Å of $r$ in the model. Similarly, amide spin systems that are found to be fast-exchanging using amide exchange studies 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, [30, 4, 61] including: orientation and placement of secondary structure to determine protein folds [21], pruning an homologous structural database [5, 41], de novo structure determination [49], in combination with a sparse set of assigned NOE’s to determine the global fold [42], and a method developed by Bax and

<table>
<thead>
<tr>
<th>Experiment/Data</th>
<th>Information Content</th>
<th>Role in NVR</th>
<th>Acquisition Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H^N,^{10}N$ HSQC</td>
<td>$H^N,^{10}N$ Chemical shifts</td>
<td>Backbone resonances, Cross-referencing NOESY</td>
<td>1/2 hr.</td>
</tr>
<tr>
<td>$H^N,^{15}N$ RDC (in 2 media)</td>
<td>Restraints on amide bond vector orientation</td>
<td>Tensor Estimation, Resonance Assignment, Structure Refinement</td>
<td>1/2 hr. + 1/2 hr.</td>
</tr>
<tr>
<td>H-D exchange HSQC</td>
<td>Identifies solvent exposed amide protons</td>
<td>Resonance Assignment</td>
<td>1/2 hr.</td>
</tr>
<tr>
<td>$H^N,^{15}N$ HSQC-NOESY</td>
<td>Distance restraints between spin systems</td>
<td>Resonance Assignment</td>
<td>12 hrs.</td>
</tr>
<tr>
<td>Structural model of backbone</td>
<td>Tertiary Structure</td>
<td>Tensor Estimation, Resonance Assignment, Structure Refinement</td>
<td>assumed given</td>
</tr>
</tbody>
</table>

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 $d_{\text{NN}}$ can be obtained from the NOESY. These $d_{\text{NN}}$ 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.
co-workers for fold determination that selects heptapeptide fragments best fitting the assigned RDC data [16]. 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 [66, 16]. In another study, RDCs were used by Prestegard and co-workers [56] 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 require $^{13}$C-labelling and RDCs from many different bonds (for example, $^{13}$C′-$^{15}$N, $^{13}$C′-H,N, $^{13}$Cα, H,N, etc.). The CAP method for small RNA assignment [2], also requires $^{13}$C-labelling and many RDCs in addition to many through-bond, triple resonance experiments. More recently, Brüschweiler and co-workers [31] 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 requires several RDCs per residue and the recording of several $^{13}$C triple resonance experiments. In contrast, NVR requires only amide bond vector RDCs, no triple-resonance experiments, and no $^{13}$C-labelling (cf. Wüthrich: “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′-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 [56] 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.

Our method addresses the same problem as HPB, but uses a different algorithm. HPB requires RDCs for several different inter-molecular bond vectors in a residue, and records several triple-resonance spectra (HNCOCO, 3D CBCA(CO)HN, and 3D HNCACB) to group the RDCs into spin systems. The HPB method iteratively solves for both the alignment tensor S and the resonance assignments. Interestingly, the 3D CBCA(CO)HN and 3D HNCACB spectra alone (i.e., without a structural model or RDCs) have enough information to make most or all of the sequential assignments [63]. One might imagine using these assignments to solve for the tensor S immediately and then use the RDCs to assign the remaining resonances.

From a computational standpoint, NVR adopts a minimalist approach [8], 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 (nucleotides or bases) to be assigned, and the spectral complexity (degree of degeneracy and overlap in frequency space). For example, CAP [2] has been applied with $n = 27$ nt., and the time complexity of CAP grows exponentially with n. In particular, CAP performs exhaustive search, making it difficult to scale up to larger RNAs. HPB runs time $O(In^3)$, where I is the number of iterations of the main loop required to converge to a solution. While the convergence of HPB was not proven, in their algorithm, I is bounded by $O(k^3)$, reflecting the discrete grid search for the principal order frame over 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 $O(nk^3+n^2)$ 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 $S_1$ and $S_2$ be the estimated tensors for the phage and bicelle media, respectively. These tensors correspond to the matrix S in Eq. (1). Macromolecules align differently in different liquid crystals, thus $S_1$ and $S_2$ are different matrices. $S_1$ and $S_2$ 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)
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$^3$-H$^2$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 $O(n^7)$. 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.

}\hspace{1cm}\begin{align*}
\text{assignments and the structure of the macromolecule are known, all five parameters can be computed by solving a system of linear equations [40]. If the resonance assignments are not known, as in our case, these parameters must be estimated. It has been shown [40] that $D_a$ and $D_r$ can be decoupled from the Euler angles by diagonalizing the alignment tensor:} \hspace{1cm}
\begin{align*}
S &= \mathbf{V} \Sigma \mathbf{V}^T \tag{2}
\end{align*}
\end{align*}

Here, $\mathbf{V} \in SO(3)$ is a $3 \times 3$ rotation matrix$^2$ that defines a coordinate system called the *principal order frame*. $\Sigma$ is a $3 \times 3$ diagonal and traceless matrix containing the eigenvalues of $S$. The diagonal elements of $\Sigma$ encode $D_a$ and $D_r$: $D_a = \frac{S_{xx} + S_{yy} + S_{zz}}{3}$, $D_r = \frac{S_{xx} + S_{yy} + S_{zz}}{3}$, where $S_{xx} < S_{yy} < S_{zz}$. $S_{yy}$, $S_{xx}$ and $S_{zz}$ are the diagonal elements of $\Sigma$ and therefore the eigenvalues of $S$. It has been shown that $D_a$ and $D_r$ can be estimated, using only unassigned experimentally recorded RDCs, by the powder pattern method [61]. 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 Eq. (2) can be constructed using the relationship [40, 61] 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 $\Sigma$ (Eq. 2), for each rotation $V(\alpha, \beta, \gamma)$ of the model, a new Saupe matrix $S$ is computed using Eq. (2). That matrix $S$ 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 relative entropy measure is conceptually related to the premise used by the powder pattern method [61] 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 bond 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 unassigned RDCs that is similar to the unassigned distribution of experimentally measured RDCs. The interested reader is directed to Fig. 6 in the appendix for more information.

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.2); 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 $20^\circ$ between the actual and estimated Euler angles were seen for one of our test proteins (see Fig. 10 in the appendix). 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 lat-
Estimates bootstrap the iterative process. During each iteration, the probability of each (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.

The input to phase 2 (Fig. 2) includes the two order matrices \(S_1\) and \(S_2\) 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 \(S_m\). For each medium \(m\), a \(n\)-peak \(\times\) \(n\)-residue probability matrix \(M_m\) is constructed. The rows of \(M_m\) correspond to some fixed ordering of the peaks in the HSQC. Similarly, the columns of \(M_m\) correspond to some fixed ordering of the residues in the protein. The assignment probabilities are computed as follows:

\[
M_m(q, r) = P(q \mapsto r | S_m) = N(d_m(q) - b_m(r), S_m, \mu_m, \sigma_m)
\]

where \(q \in Q\) and \(r \in R\), \(d_m(q) \in D_m\), \(b_m(r) \in B_m\). 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 \(\mu\) and standard deviation \(\sigma\). We used \(\mu = 0\) Hz and \(\sigma = 1\) Hz in all our trials. Intuitively, \(M_m(q, r)\) is the probability that peak \(q\) is assigned to reside \(r\) in medium \(m\). An individual entry of \(M_m\) may be set to zero if the assignment \(q \mapsto r\) violates a geometric constraint imposed by a \(d_{NN}\) or amide exchange.

On each iteration, the probabilities of assignment are (re)computed using Eq. (3). For each row in \(M_1\) and \(M_2\) 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:

\[
r_m(q) \neq r_m(k) \quad m = 1, 2; \forall k \in Q, k \neq q.
\]

When an assignment is made, peak \(q\) and residue \(r\) are removed from consideration in subsequent iterations. Thus, the size of matrices \(M_1\) and \(M_2\) decreases with each iteration. At the end of each iteration alignment tensors \(S_1\) and \(S_2\) are refined by using the master list of assignments and the model, by means of the SVD method [40]. 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 one degree, and the axial and rhombic components are within 1-2% of their correct values, respectively. The interested reader is directed to Fig. 10 in the appendix for more information.

The computational complexity of the second phase is as follows. The tensors, using the Moore-Penrose pseudoinverse of the \(O(n) \times 5\) matrix takes time \(O(n^2)\) [24].

\(^3\)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[15, 34, 54] and/or (b) computing the optimal Saupe matrix using the correct assignments.
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. NV R 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 (see Tables 5, 6 and 7 in the appendix) 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.

Intuitively, NVR only makes assignments that are a) unambiguous and b) consistent across both media. Figure 3 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 assignments. These confidence scores are reported to the user. The highest-confidence assignments tend to be in regions of regular secondary structure (Fig 4).

### 4.3 Structure Refinements (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. \(T\) 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) 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 [46]. This minimization is done in vacuo. Figure 5 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 [60, 53], backbone amide residual dipolar couplings recorded in two separate liquid crystals (bicelle and phage) [15], amide-exchange rates [15], \(^{15}\text{N}-\text{HSQC}\) and \(^{15}\text{N}-\text{HSQC}\) NOESY spectra [29], and several independent high-resolution structures solved by both x-ray crystallography [47, 59] and NMR [7, 32]. 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 form of ubiquitin where 7 hydrophobic core residues have been altered (13V, V5L, L15V, L23F, V26F, L67I). 1UD7 was chosen to test the effectiveness of NVR when the model is a close homolog of the target protein. Our algorithm performs resonance assignment by fitting experimentally recorded dipolar couplings to bond vectors from structural models. 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 to bond vectors from structural models.

We have found that the errors that our algorithm makes are, in general, easily explained. Almost all errors are symmetric. That 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 tri-
Figure 3: 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.

Figure 4: Assignment Confidences. NVR returns the confidence of each assignment. Here the structure of ubiquitin model 1UBQ (two different rotations) 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.

Figure 5: 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.
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 [48], and can be worse, making these bond 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.

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 15N-labeling of the protein, and processes unassigned 15N-HSQC and H-D exchange-HSQC spectra, HN, 15N RDCs, and sparse HN-HN NOE’s (dNNN), 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 15N-HSQC spectrum as well as the dNNN of the 3D 15N-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. We are currently exploring 1H and 15N chemical shift prediction [44, 62] for NVR. 1H and 15N chemical shift predictions will be incorporated into NVR as a probabilistic constraint on assignments. Potentially, chemical shift prediction might obviate the need for recording the NOESY, which is at this time, the most time-consuming experiment in the NVR method (12 hours using a cryoprobe).

Finally, 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 [3], protein threading [38, 65], or homology modelling [11, 19, 25, 33, 51], 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. Another application would be to use NVR to identify homologous structures.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Exp. Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1G6J</td>
<td>NMR</td>
<td>90</td>
</tr>
<tr>
<td>1UBI</td>
<td>X-ray (1.8 Å)</td>
<td>90</td>
</tr>
<tr>
<td>1UBQ</td>
<td>X-ray (1.8 Å)</td>
<td>93</td>
</tr>
<tr>
<td>1UD7</td>
<td>NMR</td>
<td>93</td>
</tr>
</tbody>
</table>
In principle, one could run NVR in parallel, using many structures from different fold-families. It is possible that the structure(s) that are in fact homologous to the target protein would have higher confidence scores than those that are unrelated.

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 and C. Bailey-Kellogg, Ms. E. Werner-Reiss, and all members of Donald Lab for helpful discussions and comments on drafts.

References


[52] SAUPE, A. Recent Results in the field of liquid crystals. Angew. Chem. 7 (1968), 97–112.


A Appendix

A.1 Supplementary Results and Supporting Material

For the interested reader, the following figures and tables are provided as supplementary results and supporting material. Appendix A.2 discusses the computational complexity of the Minimum Kullback-Leibler
Figure 6 illustrates the comparison of distributions used in the tensor estimation phase of NVR. Figures 7, 8 and 9 report the accuracies of the initial tensor estimated for ubiquitin, streptococcal protein G and lysozyme matched to 20 models. Figure 10 compares the initial tensor estimate (phase 1) for each of the 4 ubiquitin structure models (1G6J, 1UBQ, 1UBI, 1UD7) to the final tensor estimates returned after resonance assignment (phase 2). Tables 3 and 4 report the accuracy of NVR on streptococcal protein G and lysozyme matched to 16 structures. Finally, Tables 5, 6 and 7 compares the accuracies of several different assignment algorithms on ubiquitin, streptococcal protein G and lysozyme matched to 20 different structures.

A.2 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 [26, 27, 10, 9]. 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), $B(V) = E^T SE = (E^T V \Sigma V) E = \{ w^T (V^T \Sigma V) w \mid w \in E \}$. (We have ignored $D_{\text{max}}$ here for the simplicity of exposition). We wish to compute

$$\arg\min_{V \in S^3} KL(D, B(V))$$

(5)

(We use the unit 3-sphere $S^3$ instead of $SO(3)$, since the quaternions are a double-covering of rotation space). Eq. (5) can be tranformed into a sentence in the language of semi-algebraic sets (the first order theory of real closed fields):

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

(6)

$S^3$ and $SO(3)$ are semi-algebraic sets, and Eq. (6) is a polynomial inequality with bounded quantifier alternation ($\alpha = 1$). 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. (6) can be decided exactly, in polynomial time, using the theory of real-closed fields. We will use Grigor’ev’s algorithm [26, 27] for deciding a Tarski sentence, which is singly-exponential in the number of variables, and doubly-exponential only in the number of quantifier alternations. The time complexity of Grigor’ev’s algorithm is $n^{O(r)^{1\alpha-2}}$, which in our case ($\alpha = 1$, $r = 3$) reduces to $n^{O(1)}$ which is polynomial time.
Figure 6: 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.

Figure 7: 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$^\circ$. However, angular deviations of 20$^\circ$ represent only 3% of the total surface area of the unit sphere (see text). See Fig. 10 for the (improved) final tensor estimates.
Figure 8: Streptococcal Protein G 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, 1GB1, 2GB1 and 1PGB, 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 40°. However, angular deviations of 40° represent only 12% of the total surface area of the unit sphere (see text).

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GB1</td>
<td>95%</td>
</tr>
<tr>
<td>2GB1</td>
<td>95%</td>
</tr>
<tr>
<td>1PGB</td>
<td>95%</td>
</tr>
</tbody>
</table>

Table 3: Resonance Assignments. NVR achieves an average of 95% accuracy on the 3 SPG models.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>193L</td>
<td>100%</td>
</tr>
<tr>
<td>1AKI</td>
<td>100%</td>
</tr>
<tr>
<td>1AZF</td>
<td>100%</td>
</tr>
<tr>
<td>1BGI</td>
<td>98%</td>
</tr>
<tr>
<td>1H87</td>
<td>100%</td>
</tr>
<tr>
<td>1LS3</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1LYZ</td>
<td>91%</td>
</tr>
<tr>
<td>2LYZ</td>
<td>98%</td>
</tr>
<tr>
<td>3LYZ</td>
<td>100%</td>
</tr>
<tr>
<td>4LYZ</td>
<td>96%</td>
</tr>
<tr>
<td>5LYZ</td>
<td>98%</td>
</tr>
<tr>
<td>6LYZ</td>
<td>98%</td>
</tr>
</tbody>
</table>

Table 4: Lysozyme Resonance Assignments. NVR achieves an average of 98% accuracy on the 13 lysozyme models.
Figure 9: Lysozyme 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 thirteen models, 193L, 1AKI, 1AZF, 1BGI, 1H87, 1LSC, 1LSE, 1LYZ, 2LYZ, 3LYZ, 4LYZ, 5LYZ, and 6LYZ, 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. Here, $S_{zz}$ 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 $45^\circ$. However, angular deviations of $45^\circ$ represent only 15% of the total surface area of the unit sphere (see text).

Table 5: 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 using 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 3 (A).
Figure 10: Ubiquitin Tensor Improvements. Left-hand panels are the accuracies of the initial tensor estimates for ubiquitin in the bicelle medium. The right hand panels are the accuracies of the final tensor estimates, after NVR has completed the resonance assignment phase. The final axial and rhombic components (\(D_a, D_r\)) are within 1-2% of their true values, while the eigenvectors are within 1-2\(^\circ\) of their true values. Similar improvements are seen for ubiquitin in the phage medium and for streptococcal protein G and lysozyme (data not shown).

Table 6: SPG: 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 using 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 3.
Table 7: Lysozyme: 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 using 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 4.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Maximum Bipartite Matching</th>
<th>NVR with RDC and Amide Exchange</th>
<th>NVR with RDC and NOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>193L</td>
<td>21%</td>
<td>23%</td>
<td>93%</td>
</tr>
<tr>
<td>1AKI</td>
<td>9%</td>
<td>56%</td>
<td>83%</td>
</tr>
<tr>
<td>1AZF</td>
<td>15%</td>
<td>19%</td>
<td>84%</td>
</tr>
<tr>
<td>1BGI</td>
<td>18%</td>
<td>51%</td>
<td>98%</td>
</tr>
<tr>
<td>1HS7</td>
<td>16%</td>
<td>13%</td>
<td>95%</td>
</tr>
<tr>
<td>1LSC</td>
<td>23%</td>
<td>17%</td>
<td>94%</td>
</tr>
<tr>
<td>1LSE</td>
<td>9%</td>
<td>29%</td>
<td>92%</td>
</tr>
<tr>
<td>1LYZ</td>
<td>2%</td>
<td>15%</td>
<td>54%</td>
</tr>
<tr>
<td>2LYZ</td>
<td>13%</td>
<td>9%</td>
<td>77%</td>
</tr>
<tr>
<td>3LYZ</td>
<td>8%</td>
<td>28%</td>
<td>97%</td>
</tr>
<tr>
<td>4LYZ</td>
<td>13%</td>
<td>33%</td>
<td>86%</td>
</tr>
<tr>
<td>5LYZ</td>
<td>12%</td>
<td>33%</td>
<td>95%</td>
</tr>
<tr>
<td>6LYZ</td>
<td>10%</td>
<td>45%</td>
<td>93%</td>
</tr>
</tbody>
</table>