Journal of Biomolecular NMR **29:** 111–138, 2004.

An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments

Christopher James Langmead[a] & Bruce Randall Donald[a,b,c,d,*,**]
[a]*Dartmouth Computer Science Department,* [b]*Dartmouth Chemistry Department,* [c]*Dartmouth Biological Sciences Department and* [d]*Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA*

Received 28 July 2003; Accepted 8 December 2003

Abstract

We report an automated procedure for high-throughput NMR resonance assignment for a protein of known structure, or of an homologous structure. Our algorithm performs *Nuclear Vector Replacement* (NVR) by *Expectation/Maximization* (EM) to compute assignments. NVR correlates experimentally-measured NH residual dipolar couplings (RDCs) and chemical shifts 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). NVR runs in minutes and efficiently assigns the (H^N,^{15}N) backbone resonances as well as the sparse d_{NN}s from 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 99%. We further demonstrate the feasibility of our algorithm for different and larger proteins, using different combinations of real and simulated NMR data for hen lysozyme (129 residues) and streptococcal protein G (56 residues), matched to a variety of 3D structural models.

Abbreviations: 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; EM, Expectation/Maximization; SVD, singular value decomposition.

Introduction

We seek to accelerate protein NMR resonance assignment and structure determination by exploiting *a priori* structural information. By analogy, in X-ray crystallography, the molecular replacement (MR) technique (Rossman and Blow, 1962) allows solution of the crystallographic phase problem when a 'close' or homologous structural model is known *a priori*, thereby facilitating rapid structure determination. In contrast, a key bottleneck in NMR structural biology is the resonance assignment problem. One would hope that knowing a structural model ahead of time could expedite assignments. An automated procedure for rapidly determining NMR resonance assignments given an homologous structure, would similarly accelerate structure determination. Moreover, even when the structure has already been determined by x-ray crystallography or computational homology modelling, NMR assignments are valuable because NMR can be used to probe protein-protein interactions (Fiaux et al., 2002) (via chemical shift mapping Chen

*To whom correspondence should be addressed. 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. E-mail: brd@cs.dartmouth.edu
**This work is supported by grants to B.R.D. from the National Institutes of Health (GM-65982) and the National Science Foundation (IIS-9906790, EIA-0305444, and EIA-9802068).

et al., 1993)), protein-ligand binding (via SAR by NMR (Shuker et al., 1996) or line-broadening analysis (Fejzo et al., 1999)), and dynamics (via, e.g., nuclear spin relaxation analysis (Palmer, 1997)).

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 MR and computational homology modelling more widely applicable for the determination of future structures. 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)* (Figure 1), which computes assignments that correlate experimentally-measured residual dipolar couplings, chemical shifts, H^N-H^N NOE's (d_{NN}s) and amide exchange rates to a given *a priori* whole-protein 3D structural model. We believe this algorithm could form the basis for 'MR by NMR'.

Residual dipolar couplings (RDCs) (Tjandra and Bax, 1997; Tolman et al., 1995) 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

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

where $D_{\max}$ is a constant, and $\mathbf{v}$ is the internuclear vector orientation relative to an arbitrary substructure frame and $\mathbf{S}$ is the 3×3 *Saupe order matrix* (Saupe, 1968). $\mathbf{S}$ is a symmetric, traceless, rank 2 tensor with 5 degrees of freedom, which describes the average substructure alignment in the dilute liquid crystalline phase. The measurement of five or more RDCs in substructures of known geometry allows determination of $\mathbf{S}$ (Losonczi et al., 1999). Once $\mathbf{S}$ has been determined, 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).

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

The idea of correlating unassigned experimentally measured RDCs with bond vector orientations from a known structure was first proposed by Al-Hashimi and Patel (2002) and subsequently demonstrated in Al-Hashimi et al. (2002) who considered permutations of assignments for RNA, and (Hus et al., 2002) who assigned a protein from a known structure using bipartite matching. Our algorithm builds on these works and offers some improvements in terms of isotopic labelling, spectrometer time, accuracy and computational complexity. Like Hus et al. (2002), we call optimal bipartite matching as a subroutine, but within an Expectation/Maximization framework which offers some benefits, which we describe below. Previous methods 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.). Our method addresses the same problem, but uses a different algorithm and requires only amide bond vector RDCs, no triple-resonance experiments, and no ^{13}C-labelling. Moreover, our algorithm is more efficient. 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, if a rotation search is required, the resolution k^3 of a rotation-space grid over $SO(3)$. The time-complexity of the RNA-assignment method, named CAP, proposed in Al-Hashimi et al. (2002) grows exponentially with n. In particular, CAP performs an exhaustive search over all permutations, making it difficult to scale up to larger RNAs. The method presented in Hus et al. (2002) runs in time $O(In^3)$, where $O(n^3)$ is the complexity of bipartite matching (Kuhn, 1955)

and I is the number of times that the bipartite matching algorithm is called. I may be bounded by $O(k^3)$, the size of the discrete grid search for the principal order frame over $SO(3)$ (using 3 Euler angles). Here, k is the resolution of the grid. Thus, the full time-complexity of the algorithm presented in Hus et al. (2002) is $O(k^3n^3)$. The method presented in (Langmead et al., 2003; Langmead and Donald, 2003) also performs a discrete grid search for the principal order frame over $SO(3)$, but uses a more efficient algorithm with time-complexity $O(nk^3)$. Once the principle order frame has been computed, resonance assignments are made in time $O(n^3)$. Thus, the total running time of the method presented in (Langmead et al., 2003) is $O(nk^3 + n^3)$. (Zweckstetter, 2003) has recently reported a technique for estimating alignment tensors (but not assignments) using permutations of assignments on a subset of the residues identified using either selective labelling or C_α and C_β chemical shifts. If m residues can be identified *a priori* (using, e.g., selective labelling) as being a unique amino acid type, then (Zweckstetter, 2003) provides an $O(nm^6)$ tensor estimation algorithm that searches over the possible assignment permutations for the m RDCs.

The algorithm presented in the current paper requires neither a search over assignment permutations, nor a rotation search over $SO(3)$. Rather, the technique of Expectation/Maximization (EM) (Dempster et al., 1977) is used to correlate the chemical shifts of the H^N-^{15}N HSQC resonance peaks with the structural model. In practice, the application of EM on the chemical shift data is sufficient to uniquely assign a small number of resonance peaks. In particular, EM is able to assign a sufficient number of peaks for direct determination of the alignment tensor **S**. NVR eliminates the rotation grid-search over $SO(3)$, and hence any complexity dependency on a grid or its resolution k, running in $O(n^3)$ time, scaling easily to proteins in the middle NMR size range ($n = 56$ to 129 residues). Moreover, our algorithm elegantly handles missing data (both resonances and RDCs). We note that NVR both adopts a 'best-first' strategy and uses structural homology to make assignments; best-first and homology-based strategies for disambiguating assignments are well-established techniques (e.g., Hoch et al., 1990; Redfield et al., 1983).

From a computational standpoint, NVR adopts a minimalist approach (Bailey-Kellogg et al., 2000), demonstrating the large amount of information available in a few key spectra. By eliminating the need for triple resonance experiments, NVR saves spectrometer time. NVR processes unassigned H^N-^{15}N 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, when the NMR spectrometer is equipped with a cryoprobe. There are two classes of constraints used by our algorithm: geometric and probabilistic. The H^N-^{15}N RDC, H-D exchange and ^{15}N-NOESY each provide independent geometric constraints on assignment. A sparse number of d_{NN}s are extracted from the unassigned NOESY after the diagonal peaks of the NOESY are cross-referenced to the peaks in the HSQC. These d_{NN}s provide distance restraints for assignments. In general, there are only a small number of unambiguous d_{NN}s that can be obtained from an unassigned H^N-H^N NOESY. The amide exchange information probabilistically identifies the peaks in the HSQC corresponding to non hydrogen-bonded, solvent-accessible backbone amide protons. RDC's provide probabilistic constraints on each backbone amide-bond vector's orientation in the principle order frame. Finally, chemical shift prediction is employed to compute a probabilistic constraint on assignment. NVR exploits the geometric and probabilistic constraints by combining them within the Expectation/Maximization framework.

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 different NOESY data sets than that processed by NVR), achieving an average assignment accuracy of over 99%. 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. The feasibility of NVR for larger and different proteins is explored using different combinations of real and simulated NMR data for hen lysozyme (129 residues) and streptococcal protein G (56 residues).

Results and discussion

The experimental inputs to NVR are detailed in Table 1. The method is divided into two phases, *Tensor Determination* and *Resonance Assignment* (Figure 1). In the first phase, chemical shift predictions, d_{NN}s, and amide exchange rates are used to make a small number of assignments using Expectation/Maximization (EM). Specifically, this phase attempts to assign at

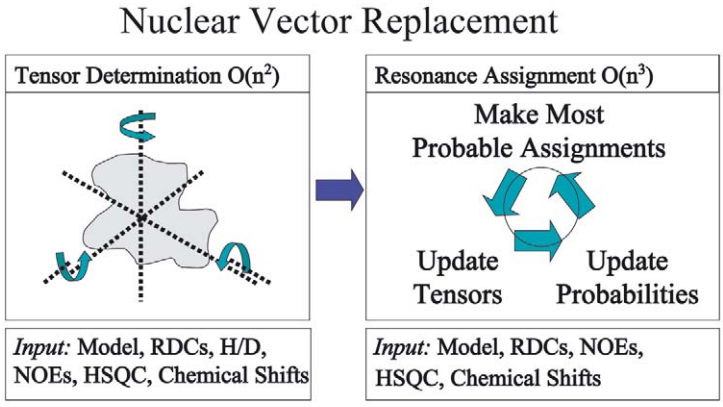

Figure 1. Nuclear Vector Replacement.

least 5 peaks for the purpose of determining the alignment tensors directly (Losonczi et al., 1999). The tensors are used to convert RDCs into probabilistic constraints. Algorithmically, the only difference between phases 1 and 2 is that phase 1 does not use RDCs (because the tensors have not yet been determined).

The molecular structure of human ubiquitin has been investigated extensively. A variety of data have been published including resonance assignments (Weber et al., 1987; Schneider et al., 1992), backbone amide residual dipolar couplings recorded in two separate liquid crystals (bicelle and phage) (Cornilescu et al., 1998), amide-exchange rates (Cornilescu et al., 1998), ^{15}N-HSQC and ^{15}N-HSQC NOESY spectra (Harris, 2002), and several independent high-resolution structures solved by both x-ray crystallography (Ramage et al., 1994; Vijay-Kumar et al., 1987) and NMR (Babu et al., 2001; Johnson et al., 1999). In 1998, the Bax lab published a new NMR structure for ubiquitin, (PDB Id: 1D3Z) (Cornilescu et al., 1998). Unlike previous ubiquitin structures, 1D3Z was refined using dipolar couplings. Table 2 summarizes the differences in tertiary structures between 1D3Z and four alternative high-resolution structures (PDB Ids: 1G6J, 1UBI, 1UBQ, 1UD7) of human ubiquitin, none of which have been refined using dipolar couplings. We report both all-atom and backbone RMSDs because while NVR processes atomic coordinates from backbone atoms, the two programs used for chemical shift prediction (SHIFTS and SHIFTX) both require all-atoms and are therefore affected by all-atom RMSD. 1G6J, 1UBI and 1UBQ have 100% sequence identity to 1D3Z. 1UD7 is mutant form 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. 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 (Cornilescu et al., 1998) for human ubiquitin were fit to the amide bond vectors of the selected model. ^{15}N-HSQC and ^{15}N-HSQC NOESY spectra (Harris, 2002) were processed to extract sparse, unassigned d_{NN}s.

In addition to the trials on ubiquitin, NVR was applied to two additional proteins, the 56-residue streptococcal protein G (SPG) and the 129-residue hen lysozyme. For both proteins, there exist published chemical shifts deposited into BMRB (Seavey et al., 1991), amide-bond RDC data (Kuszewski et al., 1999; Schwalbe et al., 2001) and RDC-refined structures (PDB Id: 3GB1 (Kuszewski et al., 1999)), (PDB Id: 1E8L (Schwalbe et al., 2001)). Several high-resolution alternative structures are also available (Tables 3 and 4). Using NVR, the experimentally recorded NH dipolar couplings (Kuszewski et al., 1999; Schwalbe et al., 2001) were fit to the amide bond vectors of the selected model. A set of sparse, unassigned d_{NN}s were simulated for SPG (using SPG's chemical shifts and the PDB restraint file (Kuszewski et al., 1999) of NOEs for 3GB1) and lysozyme (using Lysozyme's chemical shifts and the PDB restraint file (Schwalbe et al., 2001) of NOEs for 1E8L).

Table 1. NVR experiment suite

Experiment/data	Information content	Role in NVR
H^N-^{15}N HSQC	H^N,^{15}N Chemical shifts	Backbone resonances, cross-referencing NOESY
H^N-^{15}N RDC (in 2 media)	Restraints on amide bond vector orientation	Tensor determination, resonance assignment
H-D exchange HSQC	Identifies solvent exposed amide protons	Tensor determination
H^N-^{15}N HSQC-NOESY	Distance restraints between spin systems	Tensor determination, resonance assignment
Backbone structure	Tertiary structure	Tensor determination, resonance assignment, chemical shift prediction
Chemical shift predictions	Restraints on assignment	Tensor determination, Resonance assignment

Expectation/maximization

We outline in this section the EM algorithm, a variation of which is used in both the first and second phases of NVR. Details of the algorithm, and its implementation are presented in the Methods section. EM has been described previously (Dempster et al., 1977). EM is a statistical method for computing the maximum likelihood estimates of parameters for a generative model. EM has been a popular technique in a number of different fields, including machine learning and computer vision. It has been applied to bipartite matching problems in computer vision (Cross and Hancock, 1998). In the EM framework there are both observed and hidden (i.e., unobserved) random variables. In the context of resonance assignment, the observed variables are the chemical shifts, d_{NN}s, amide exchange rates, RDCs, and the 3D structure of the target protein. Let X be the set of observed variables.

The hidden variables $Y = Y_G \cup Y_S$ are the true (i.e., correct) resonance assignments Y_G, and Y_S, the correct, or 'true' alignment tensors. Of course, the values of the hidden variables are unknown. Specifically, Y_G is the set of edge weights of a bipartite graph, $G = \{ K, R, K \times R \}$, where K is the set of peaks in the HSQC and R is the set of residues in the protein. The weights Y_G represent *correct* assignments, therefore encode a perfect matching in G. Hence, for each peak $k \in K$ (respectively, residue $r \in R$), exactly one edge weight from k (respectively r) is 1 and the rest are 0. The probabilities on all variables in Y are parameterized by the 'model', which is the set Θ of all assignments made so far by the algorithm. Initially, Θ is empty. As EM makes more assignments, Θ grows, and both the probabilities on the edge weights Y_G and the probabilities on the alignment tensor values Y_S will change. The goal of the EM algorithm is to estimate Y accurately to discover the correct edge weights Y_G, thereby computing the correct assignments. The EM algorithm has two steps; the Expectation (E) step and the Maximization (M) step. The E step computes the expectation

$$E(\Theta \cup \Theta' | \Theta) = E(\log \mathbf{P}(X, Y | \Theta \cup \Theta')). \qquad (2)$$

Here, Θ' is a non-empty set of candidate new assignments that is disjoint from Θ. The M step computes the maximum likelihood new assignments Θ^*,

$$\Theta^* = \underset{\Theta'}{\operatorname{argmax}}\, E(\Theta \cup \Theta' | \Theta). \qquad (3)$$

Then the master list of assignments is updated, $\Theta \leftarrow \Theta \cup \Theta^*$. The alignment tensors are re-computed at the end of each iteration, using all the assignments in Θ. Thus, the tensor estimates are continually refined during the run of the algorithm. The algorithm terminates when each peak has been assigned. Care must be taken to implement the probabilistic EM framework efficiently. The details of how the E and M steps are implemented are presented in the Methods section.

Table 2. Human ubiquitin. The 4 structures of human ubiquitin used in the 4 separate trials of NVR

PDB ID	Exp. method	Comparison to 1D3Z		
		Sequence identity	All-atom RMSD	Backbone RMSD
1G6J (Babu et al., 2001)	NMR	100%	2.4 Å	2.0 Å
1UBI (Ramage et al., 1994)	X-ray (1.8 Å)	100%	1.3 Å	0.6 Å
1UBQ (Vijay-Kumar et al., 1987)	X-ray (1.8 Å)	100%	1.4 Å	0.6 Å
1UD7 (Johnson et al., 1999)	NMR	90%	2.5 Å	2.3 Å

Table 3. Streptococcal protein G (SPG). The 3 structures of SPG used in the 3 separate trials of NVR

PDB ID	Exp. method	Comparison to 3GB1	
		All-atom RMSD	Backbone RMSD
1GB1 (Gronenborn et al., 1991)	NMR	1.3 Å	1.0 Å
2GB1 (Gronenborn et al., 1991)	NMR	1.3 Å	1.0 Å
1PGB (Gallagher et al., 1994)	X-ray (1.92 Å)	1.2 Å	0.6 Å

Missing data

The EM algorithm affords an elegant and intuitive means for handling missing data. Table 5 summarizes the data processed in our 20 experiments on 3 proteins. In theory, the HSQC spectrum should contain one peak per residue in the protein (except prolines, and the N-terminus). In reality, some peaks may be 'missing' from the spectrum. For example, the ubiquitin HSQC data processed by NVR lacks peaks for Glu24 and Gly53. Furthermore, it is not always possible to record two RDCs for each backbone amide group. The ubiquitin RDC data processed by NVR lacks RDCs for residues Thr9, Glu24, Gly53, Leu73, Arg74, Gly75, and Gly76 in one medium, and for residues Thr9, Glu24, Gly53, Arg72, Leu73, Arg74, Gly75, and Gly76 in the other.[1] Our algorithm processed the data as-is and handles missing data directly. Missing data is handled in NVR with unbiased estimates. For example, in the ubiquitin data set, it is clear that two peaks are missing from the HSQC because we expect to see 72 peaks (76 residues − 3 prolines − N-terminus = 72), and only 70 peaks are present. In this case, the algorithm constructs and includes 2 'dummy' peaks that are interpreted as follows. Each dummy peak is assigned a uniform probability ($\mathbf{P} = 1/72$) to match all 72 expected residues when computing assignment probabilities using chemical shift data. That is, an unbiased (uniform) probability distribution is used. Similarly, if an RDC is missing in one or both media an unbiased probability distribution is used when computing assignment probabilities using RDCs.

Tensor determination (phase 1)

The experimentally determined RDC's cannot be interpreted as probabilistic constraints on assignment prior to the determination of the alignment tensor $\mathbf{S}$. $\mathbf{S}$ has five degrees of freedom. Therefore, at least five resonance assignments are needed to compute $\mathbf{S}$ directly. Consequently, the purpose of the first phase is to assign at least five peaks using the geometric and probabilistic constraints contained in the chemical shifts, d_{NN}s, and amide exchange rates. The details of how the chemical shifts, d_{NN}s and amide exchange rates are converted into constraints on assignment is described in the Methods section. The EM algorithm is used to assign a small number of peaks. In all of our 20 experiments, no more than 2 iterations of the EM algorithm on the chemical shift data, d_{NN}s and amide exchange rates were needed to obtain five assignments. Alignment tensors for both media are then computed directly by SVD (Losonczi et al., 1999). The computational complexity of each iteration of the EM algorithm is $O(n^2)$ time (see Methods). During

Table 4. Hen lysozyme. The 13 structures of hen lysozyme used in the 13 separate trials of NVR

PDB ID	Exp. method	Comparison to 1E8L	
		All-atom RMSD	Backbone RMSD
193L (Vaney et al., 1996)	X-ray (1.3 Å)	2.1 Å	1.5 Å
1AKI (Artymiuk et al., 1982)	X-ray (1.5 Å)	2.1 Å	1.5 Å
1AZF (Lim et al., 1998)	X-ray (1.8 Å)	2.1 Å	1.5 Å
1BGI (Oki et al., 1999)	X-ray (1.7 Å)	2.1 Å	1.5 Å
1H87 (Girard et al., 2001)	X-ray (1.7 Å)	2.1 Å	1.5 Å
1LSC (Kurinov and Harrison, 1995)	X-ray (1.7 Å)	2.2 Å	1.6 Å
1LSE (Kurinov and Harrison, 1995)	X-ray (1.7 Å)	2.2 Å	1.5 Å
1LYZ (Diamond, 1974)	X-ray (2.0 Å)	2.1 Å	1.5 Å
2LYZ (Diamond, 1974)	X-ray (2.0 Å)	2.1 Å	1.5 Å
3LYZ (Diamond, 1974)	X-ray (2.0 Å)	2.1 Å	1.5 Å
4LYZ (Diamond, 1974)	X-ray (2.0 Å)	2.1 Å	1.5 Å
5LYZ (Diamond, 1974)	X-ray (2.0 Å)	2.1 Å	1.5 Å
6LYZ (Diamond, 1974)	X-ray (2.0 Å)	2.1 Å	1.5 Å

Table 5. Missing data. The data processed on our experiments contained both missing peaks and missing RDCs. Column 2 indicates the number of HSQC peaks contained in our test data. Column 3 indicates the number of missing HSQC peaks (number of expected peaks − number of observed peaks). Columns 4–5 indicates the number of RDCs obtained in media 1 and 2. Columns 6–7 indicates the number of missing RDCs in media 1 and 2. The NVR algorithm processed all data as-is, and handles missing data

Protein	HSQC peaks		RDCs			
	Observed	'Missing' #, (%)	Observed		'Missing' #, (%)	
			Medium 1	Medium 2	Medium 1	Medium 2
Ubiquitin	70	2, (3%)	65	64	7 (10%)	8, (11%)
SPG	55	0, (0%)	48	46	7 (13%)	9, (16%)
Lysozyme	126	0, (0%)	107	102	19 (15%)	24, (19%)

the first phase, the EM algorithm is called a constant number of times. Thus, the cost of the first phase is $O(n^2)$.

Resonance assignment (phase 2)

Having determined the alignment tensors in the first phase, the RDCs are converted into constraints on assignment. Briefly, the amide bond vectors from the model are used to back-compute a set of RDCs using Equation 1. The difference between each back-computed RDC and each experimentally recorded RDC is converted into a probability on assignment. This conversion of RDCs into probabilities is described in detail in the Methods section. The EM algorithm is used to iteratively assign the remaining peaks. At least one assignment is made per iteration. Therefore, the second phase terminates in $O(n)$ steps. Each step takes $O(n^2)$ time, thus the computational complexity of the second phase, and therefore the entire algorithm, is $O(n^3)$. In practice, multiple peaks are assigned on each iteration and the algorithm quickly converges to a solution. In our experiments, no more than 10 iterations were ever needed to assign the HSQC spectrum, and the majority required 5 or fewer. Run-times ranged from 8 seconds for the smallest protein (SPG) to 4 min for the largest protein (lysozyme, 129 residues) on a Pentium 4-class workstation.

Accuracy

NVR achieves 100% accuracy in assigning the backbone peaks of the H^N-^{15}N HSQC spectrum using each of the four ubiquitin models. In the ubiquitin data,

Table 6. Backbone amide resonance assignment accuracy. Accuracies report the percentage of correctly-assigned backbone HSQC peaks. The 96% accuracy for 2GB1 reflects a single incorrect assignment

PDB ID	Accuracy
(A) Ubiquitin	
1G6J (Babu et al., 2001)	100%
1UBI (Ramage et al., 1994)	100%
1UBQ (Vijay-Kumar et al., 1987)	100%
1UD7 (Johnson et al., 1999)	100%
(B) SPG	
1GB1 (Gronenborn et al., 1991)	100%
2GB1 (Groenenborn et al., 1991)	96%
1PGB (Gallagher et al., 1994)	100%

Table 7. Backbone amide resonance assignment accuracy: Lysozyme

PDB ID	Accuracy
193L (Vaney et al., 1996)	100%
1AKI (Artymiuk et al., 1982)	100%
1AZF (Lim et al., 1998)	100%
1BGI (Oki et al., 1999)	100%
1H87 (Girard et al., 2001)	100%
1LSC (Kurinov and Harrison, 1995)	100%
1LSE (Kurinov and Harrison, 1995)	100%
1LYZ (Diamond, 1974)	100%
2LYZ (Diamond, 1974)	100%
3LYZ (Diamond, 1974)	100%
4LYZ (Diamond, 1974)	100%
5LYZ (Diamond, 1974)	100%
6LYZ (Diamond, 1974)	100%

Glu24 and Gly53 had missing HSQC peaks as well as missing RDCs in both media. In this case, NVR discovers and reports that residues Glu24 and Gly53 have missing peaks. NVR performed perfectly on 1UD7, a mutant of ubiquitin. This suggests that NVR might be extended to use homologous structures. The average accuracies on lysozyme and streptococcal protein G were 100% and 99%, respectively. NVR achieves consistently high accuracies, suggesting NVR is robust with respect to choice of model. Even residues that had fewer than 2 RDCs, but had chemical shifts were always assigned correctly. That is, the chemical shifts alone, or the chemical shifts and 1 RDC gave enough constraint to assign the peak.

The single assignment error made on the streptococcal protein G model 2GB1 is easily explainable. Residues Val29 and Phe30 were incorrectly swapped in the final iteration of the algorithm, due to differences between the observed and back-calculated RDCs. The observed dipolar couplings for these two residues were an average of 5.2 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 3.8 Hz.

Once the final set of assignments has been computed, the (now) assigned RDCs can be used to refine the structure of the model. This refined model can then be compared to the published RDC-refined structures of the 3 test proteins (1D3Z, 3GB1, 1E8L). If A is the model used by NVR, A' is the model obtained by refining A using the RDCs assigned by NVR, and B is the published RDC-refined structure, then ideally, $\texttt{RMSD}(A', B) < \texttt{RMSD}(A, B)$, where $\texttt{RMSD}(X, Y)$ returns the root-mean-squared distance between two models. Using an earlier version of the NVR algorithm (Langmead et al., 2003) and a Monte-Carlo algorithm for finding a (new) conformation of the model's ϕ and ψ backbone angles that best matches the observed RDCs, we have previously reported up to an 11% reduction in (backbone) RMSD. This illustrates the potential application to structural genomics, in which NVR could be used to assign and compute new structures based on homologous models. We are presently incorporating into NVR, a more advanced method for refining structures using RDCs (Wang and Donald, 2004).

Progression of assignments

Table 8 reports, for each model, the assignments made after the end of phase 1. It is these assignments that are used to construct the initial tensor estimates for phase 2. The most common amino acid type to be assigned during phase 1 is Glycine. This is perhaps expected due to Glycine's characteristic ^{15}N shift. Among models for a given protein, the residues assigned during phase 1 are fairly consistent, although there is variety. Note that at the end of phase 1, more than 5 peaks have been assigned for most models. This is consistent with the nature of our algorithm (see Methods); on a given iteration, the algorithm makes *all* unambiguous assignments. Furthermore, if a given peak is assigned, but no RDC has been recorded for that peak, the algorithm stays in phase 1 until 5 peaks with RDCs in both media have been assigned. Therefore,

Table 8. Phase 1 assignments. This table lists the assignments made by the end of phase 1 for each model. These assignments are used to construct the initial alignment tensors for phase 2. The residue numbers in italics had either 0 or 1 RDC

Protein	Model	First assignments (residue #)
Ubiquitin	1G6J	V5,I13,S20,I36,A46
	1UBI	S20,D21,T22,K33,G35,I36,A46
	1UBQ	I13,S20,D21,T22,,K33,G35,I36,A46,G47
	1UD7	S20,T22,K33,G35,I36,A46,Y59
SPG	1GB1	K4,N8,E25,K31,Q32,Y33,D36,D46,*D47*,A48,T49,K50,T51,F52
	1PGB	K4,I6,T11,G14,V21,E25,V29,*Y33*,G38,*G41*,D46,T49,K50,T51,F52
	2GB1	N8,E25,*Y33*,A34,D36,G38,D46,*D47*,A48,T49,K50,T51,F52
Lysozyme	193L	G4,H15,G16,G22,G26,S36,F38,N39,T40,N44
	1AKI	G4,G16,G26,S36,F38,N39,T40,N44,*R45*,D66
	1AZF	G4,G16,G26,S36,F38,N39,T40,*R45*,D66,I78
	1BGI	G4,H15,G16,G22,G26,S36,F38,T40,A42,D66
	1H8C	G4,M12,H15,G16,G22,G26,S36,F38,N39,T40,Q41,A42,N44,*Y53*,G54,D66
	1LSC	G4,H15,G16,G22,G26,S36,F38,N39,T40,Q41,N44,D66,*D119*
	1LSE	G4,G16,G26,S36,F38,N39,T40,N44,D66
	1LYZ	G4,H15,G26,T40,N44,D66
	2LYZ	G4,G26,V29,T40,D66
	3LYZ	G16,G26,V29,C30,F38
	4LYZ	G4,H15,G16,G22,G26,S36,F38,N39,T40,D66,I78
	5LYZ	G4,H15,G16,G22,G26,S36,F38,N39,T40,D66,I78
	6LYZ	G4,G16,G26,T40,D66

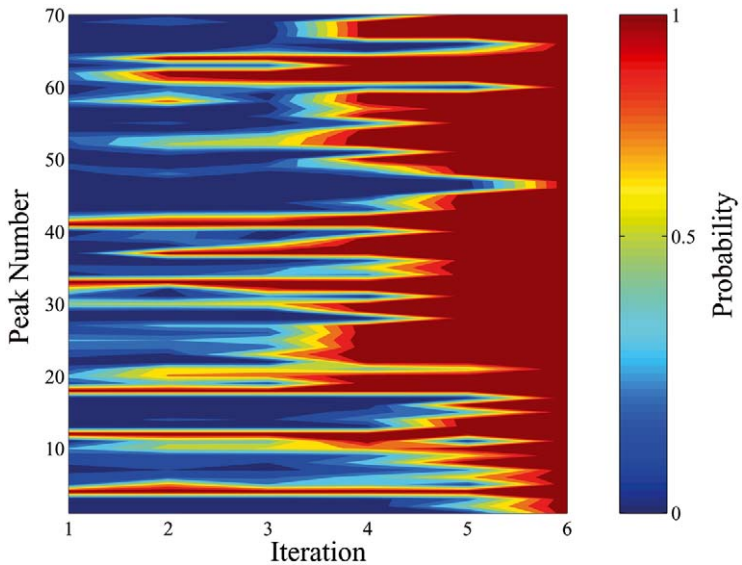

Figure 2. Evolution of probabilities: 1G6J: The evolution of each peak's assignment probability to the residue it is ultimately assigned to. Iteration number refers to the number of iterations in Phase 2 of the algorithm.

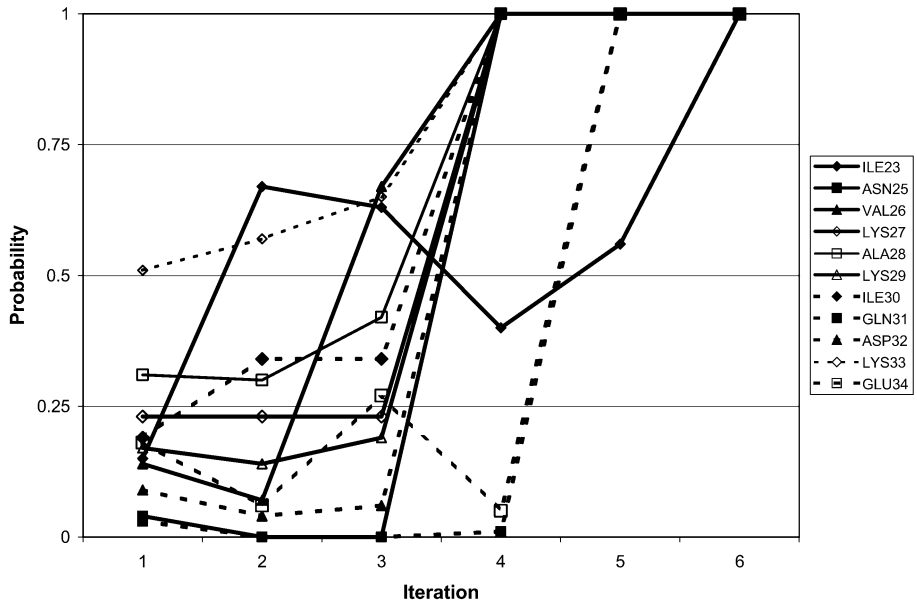

Figure 3. Evolution of residues 23-34: 1G6J: The evolution of residues 23–43 (ubiquitin's main α-helix) assignment probabilities.

it is possible for more than 5 assignments to be made during phase 1. The tensors constructed using these assignments are, of course, approximations to the true alignment tensors. However, during phase 2, the alignment tensors are refined on each iteration as more and more assignments are made.

Figure 2 depicts the evolution of the probability associated with the residue ultimately assigned a given peak for the Ubiquitin model 1G6J. The figure begins with the first iteration of phase 2 and ends at iteration 6, when all peaks have been assigned. Note that small groups of sequential residues tend to co-evolve in similar fashion. Figure 3 plots the individual probabilities of each residue in the main α-helix in Ubiquitin (residues 23-34). Figure 4 plots the evolution of the average probability, over all residues.

Table 9 lists, for each peak in Ubiquitin's HSQC, the iteration at which that peak was assigned, and the probability of that assignment at the time the assignment was made. The minimum and maximum number of assignments on a given iteration are 1 (iteration 2) and 24 (iteration 3). Thus, a single assignment can dramatically reduce the overall uncertainty.

Stability analysis

We have demonstrated that NVR performs well on real NMR data from a variety of different proteins. It is useful, however, to establish how an algorithm's performance degrades as the quality and quantity of

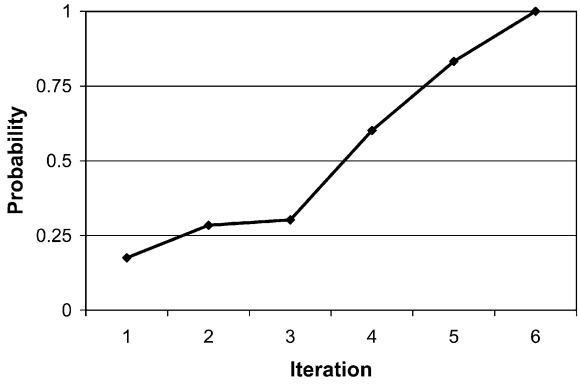

Figure 4. Evolution of average probability: 1G6J

experimental data decreases; this is especially important for sparse-data algorithms, like NVR. In this section, we discuss a series of controlled experiments to probe the stability of our algorithm under various perturbations of the data from our three test proteins.

RDCs from a single aligning medium

Recording RDCs from two separate aligning media is a standard technique for addressing the degeneracies in RDCs. However, it is of interest to determine how well NVR performs when RDCs from a single medium alone are used. To test this, we ran NVR on each of the 20 models using only one set of RDCs. The tests were exhaustive; that is, for each of the 20 models,

Table 9. First assignments: 1G6J. For each residue, this table lists when that residue was assigned to a peak. 'Phase 1' means that the residue was assigned during the first phase of NVR. Otherwise, the number refers to the iteration number in phase 2. Note that while human ubiquitin has 76 residues, this table has 70 rows. The missing rows correspond to the N-terminus, 3 prolines (residues 19,37,38), and residues 24 and 53, for which no peaks appear in the ^{15}N-HSQC of ubiquitin

Residue	Iteration	Probability (%)	Residue	Iteration	Probability (%)
Q2	5	5	Q41	4	18
I3	5	49	R42	2	99
F4	5	22	L43	3	66
V5	Phase 1	32	I44	3	3
K6	5	14	F45	4	13
T7	5	68	A46	Phase 1	99
L8	5	28	G47	3	67
T9	4	36	K48	4	12
G10	5	63	Q49	4	40
K11	4	73	L50	4	2
T12	5	11	E51	5	3
I13	Phase 1	56	D52	5	5
T14	4	5	R54	4	15
L15	4	26	T55	4	52
E16	5	31	L56	3	1
V17	4	1	S57	5	99
E18	5	7	D58	3	56
S20	Phase 1	99	Y59	3	51
D21	3	18	N60	4	97
T22	3	77	I61	3	21
I23	4	56	Q62	4	12
N25	3	1	K63	3	73
V26	3	67	E64	3	11
K27	3	23	S65	5	2
A28	3	42	T66	5	4
K29	3	19	L67	3	91
I30	3	34	H68	1	26
Q31	4	1	L69	3	16
D32	3	6	V70	1	24
K33	4	6	L71	5	57
E34	4	65	R72	4	2
G35	3	87	L73	3	2
I36	Phase 1	99	R74	3	1
D39	4	27	G75	3	5
Q40	4	64	G76	4	2

we ran NVR using the first aligning medium alone, and then with the second aligning medium alone, for a total of 40 separate experiments. The mean accuracy over these 40 trials was 92%. The standard deviation was 13% and the median accuracy was 97%. Thus, the overall drop in average accuracy is modest. In twelve instances (Ubiquitin model: 1UBQ, medium 2; SPG models: 2GB1, medium 1; 1PGB, medium 1; 1PGB, medium 2; Lysozyme models: 193L, medium 1; 193L, medium 2; 1AKI, medium 1; 1H8C, medium 2; 1LSC, medium 1; 1LSC, medium 2; 2LYZ, medium 2; 6LYZ, medium 1) the algorithm achieves 100% accuracy on a single medium. It is interesting that the SPG model 2GB1 actually does slightly better using one medium,

and not two. However, in one case (Ubiquitin model 1G6J, first medium), the accuracy dropped to 31%. We conclude that two sets of RDCs are important for overall robustness.

Reducing the number of HSQC peaks

As shown in Table 5, an HSQC may lack peaks for one or more residues. In our next set of experiments, we varied the number of 'missing' HSQC peaks. We ran 100 trials on each of the 20 models. On each trial, we randomly threw away x% of the peaks in the HSQC. These experiments were run for $x \in \{5, 10, 20, 30, 40, 50\}$. Thus, we ran 12,000 trials (20 proteins $\times$ 100 random selections $\times$ 6 percentages). Of course, when an HSQC peak is discarded, the associated RDCs, amide exchange rates, and any relevant d_{NN}'s are discarded as well. As shown in Table 10, accuracies decrease sharply when HSQC peaks are discarded. When as few as 5% of the peaks are missing, the average accuracy drops to 81%. Due to the poor performance on as few as 5% missing HSQC peaks, we ran follow-up experiments wherein we removed x HSQC peaks, randomly. These experiments were run for $x \in \{1, 2, 3, 4, 5\}$. That is, we ran an additional 10,000 trials (20 proteins $\times$ 100 random selections $\times$ 5 cases). As shown in Table 11, the average accuracy drops below 90% when 3 HSQC peaks are missing.

Reducing the number of unambiguous d_{NN}'s

NVR operates on a sparse set of unambiguous unassigned d_{NN}'s. *Unambiguous* is defined as a NOESY cross-peak whose two ^{1}H and one ^{15}N chemical shifts can be cross-referenced to a unique pair of peaks in the HSQC. The notion of unambiguous d_{NN}'s is well established (e.g., Grishaev and Llinas, 2002), and we precisely define the term in the Methods section. Whether a given d_{NN} is unambiguous depends on the local density of peaks in the spectrum, which cannot be controlled. Note that, in general, it is unlikely that any given d_{NN} is mislabelled; that is, it is unlikely we will mistakenly attribute a given d_{NN} to an incorrect pair of peaks in the HSQC. This is simply because, by definition, we only use a d_{NN} when there are only two peaks in the HSQC that could explain the data. It is reasonable, however, to consider the effects of reducing the number of unambiguous d_{NN}'s available to NVR. To test this, we ran 100 trials on each of the 20 models. On each trial, we randomly threw away x% of the d_{NN}'s. These experiments were run for $x \in \{10, 20, 30, 40, 50\}$. Thus, we ran 10,000 trials (20 proteins $\times$ 100 random selections $\times$ 5 percentages). As shown in Table 12, average accuracies above 90% are still obtained when up to 20% of the d_{NN}'s are discarded. When 50% of the d_{NN}'s are discarded, NVR still achieves greater than 70% accuracy, on average.

Corrupting d_{NN}'s

While it is very unlikely that an unambiguous d_{NN} might be mislabelled, we decided to test this scenario. To test this, we ran 100 trials on each of the 20 models. On each trial, we randomly mislabelled x of the d_{NN}'s. These experiments were run for

Table 10. Stability analysis: Reducing the number of HSQC peaks. Assignment accuracies, over 12,000 trials, when $x \in \{5, 10, 20, 30, 40, 50\}$ percent of randomly selected HSQC peaks are discarded

	Percentage of discarded HSQC peaks					
Accuracy	5	10	20	30	40	50
Mean	81%	71%	53%	38%	24%	13%
St. dev.	20%	21%	21%	20%	16%	12%
Median	88%	75%	53%	35%	21%	17%

Table 11. Stability analysis: Reducing the number of HSQC peaks, part 2. Assignment accuracies, over 10,000 trials, when $x \in \{1, 2, 3, 4, 5\}$ randomly selected HSQC peaks are discarded

	No. of discarded HSQC peaks				
Accuracy	1	2	3	4	5
Mean	96%	93%	86%	83%	82%
St. dev.	9%	12%	20%	21%	20%
Median	100%	98%	94%	91%	89%

Table 12. Stability analysis: Reducing the number of unambiguous d_{NN}'s. Assignment accuracies, over 10,000 trials, when $x \in \{10, 20, 30, 40, 50\}$ percent of randomly selected d_{NN}'s are discarded

	Percentage of discarded d_{NN}'s				
Accuracy	10	20	30	40	50
Mean	94%	91%	87%	81%	74%
St. dev.	14%	14%	16%	19%	20%
Median	98%	97%	93%	87%	79%

Table 13. Stability analysis: Corrupting d_{NN}'s. Assignment accuracies, over 10,000 trials, when $x \in \{1, 2, 3, 4, 5\}$ randomly selected unambiguous d_{NN}'s are mislabelled

	No. of mislabelled d_{NN}'s				
Accuracy	1	2	3	4	5
Mean	70%	51%	45%	38%	30%
St. dev.	32%	30%	27%	28%	23%
Median	82%	46%	39%	29%	23%

$x \in \{1, 2, 3, 3, 5\}$. Thus, we ran 10,000 trials (20 proteins $\times$ 100 random selections $\times$ 5 cases). As shown in Table 13, NVR is sensitive to mislabelled d_{NN}'s; mean accuracies drop to 70% when as few as one d_{NN} is mislabelled.

Reducing the number of slow-exchanging peaks

In our next set of experiments, we varied the number of slow-exchanging peaks input to NVR. We ran 100 trials on each of the 20 models. On each trial, we randomly mis-labelled x% of the slow-exchanging peaks in the HSQC as being fast-exchanging. These experiments were run for $x \in \{10, 20, 30, 40, 50\}$. Thus, we ran 10,000 trials (20 proteins $\times$ 100 random selections $\times$ 5 percentages). Note that in NVR, the edge-weights between slow-exchanging peaks and non-hydrogen bonded surface residues are set to zero (see Methods); by re-labelling a peak as being fast-exchanging, we remove a constraint, but do not eliminate any correct assignments from consideration. In contrast, in the next section, we consider the case where a peak may be incorrectly identified as a slow-exchanging peak, thereby disallowing a correct assignment from consideration. As shown in Table 14, NVR is quite robust to a reduced set of slow-exchanging peaks; accuracies above 90% are reported when 50% of the slow-exchanging peaks are discarded.

Corrupting the amide exchange data

Finally, we tested the effect of corrupted amide-exchange data. We ran 100 trials on each of the 20 models. Recall that each peak in the HSQC is labelled as either being a slow- or a fast-exchanging peak. On each trial, we flipped the designation of x randomly selected fast-exchanging HSQC peaks. These experiments were run for $x \in \{1, 2, 3, 4, 5\}$. Thus, we ran 10,000 trials (20 proteins $\times$ 100 random selections

Table 14. Stability analysis: Reducing the number of slow-exchanging peaks. Assignment accuracies, over 10,000 trials, when $x \in \{10, 20, 30, 40, 50\}$ percent of randomly selected slow-exchanging peaks are re-labelled as fast-exchanging

	Percentage of discarded slow-exchanging peaks				
Accuracy	10	20	30	40	50
Mean	95%	95%	94%	94%	93%
St. dev.	13%	12%	12%	11%	12%
Median	100%	100%	100%	100%	98%

Table 15. Stability analysis: Corrupted amide exchange data. Assignment accuracies, over 10,000 trials, when $x \in \{1, 2, 3, 4, 5\}$ randomly selected peak's amide exchange data is corrupted (see text)

	Number of corrupted amide exchange rates				
Accuracy	1	2	3	4	5
Mean	85%	79%	76%	67%	64%
St. dev.	22%	26%	30%	30%	30%
Median	96%	92%	92%	75%	70%

$\times$ 5 cases). As shown in Table 15, NVR is sensitive to corrupted amide exchange data. While median accuracies above 90% are observed for up to 3 corrupted peaks, the mean accuracy falls to 85% when as few as one fast-exchanging peak is (incorrectly) labelled as a slow-exchanging peak. In contrast, in the previous section we saw that when a slow-exchanging peak is (incorrectly) labelled as being fast-exchanging, accuracy remains high when up to 50% of the slow-exchanging peaks are labelled as fast-exchanging. This behavior is to be expected given the way NVR applies amide-exchange rates: The edge-weights emanating from those peaks designated as slow-exchanging and connected to solvent-accessible, labile protons are set to zero. We ran a follow-up experiment where these same edge weights were set to a small value $\epsilon > 0$, rather than zero. In this experiment, the mean accuracy was improved to 93% (st. dev 11%, median 98%) when up to 50% of the amide exchange data is corrupted.

The effects of simulating assignment errors in phase 1

The initial tensor estimates for phase 2 are determined by the assignments made in phase 1. The quality

of these tensor estimates may be poor if any of the assignments made in phase 1 are incorrect, or if the assigned substructure poorly matches the RDCs. A poor fit between substructure and data may be caused either by differences between the solution and the crystal structure, or by internal dynamics. In this section we describe the results of experiments where we simulate making a) incorrect assignments in phase 1, and b) making assignments on substructures that poorly match the RDC data.

To test the effects of incorrect assignments made during phase 1, we ran 100 trials on each of the 20 models. On each trial, we made 5 randomly-selected assignments, x of which were incorrect, in lieu of running phase 1; the algorithm then proceeds to phase 2. These experiments were run for $x \in \{1, 2, 3\}$. Thus, we ran 6,000 trials (20 proteins $\times$ 100 random selections $\times$ 3 cases). NVR is very sensitive to incorrect assignments made during phase 1; the mean accuracies for $x = 1, 2$, and 3 were, 48%, 30%, and 23%, respectively.

To test the effects of making assignments during phase 1 to substructures that poorly match the RDC data, we ran 1 experiment on each of the 20 models. For each model, we identified the five residues whose RDCs most poorly matched the RDC data. These five residues were chosen by first back-computing RDCs from the model using the 'correct' alignment tensor. By 'correct' we mean the tensor constructed from the model using the SVD method and the correct assignments. Next, the five residues with the largest combined difference (i.e., in both media) between the experimental and back-computed values were identified. These five assignments were made manually in lieu of running phase 1; the algorithm then proceeds to phase 2. NVR is not very sensitive to the quality of the initial tensor; the mean and median accuracy for these 20 experiments were 90% and 100%, respectively.

In summary, NVR is very sensitive to the accuracy of the initial phase 1 assignments, but not as sensitive to the quality of the initial tensor estimates. Given the results of our experiments using corrupted d_{NN}'s, the sensitivity of NVR to inaccurate assignments is perhaps not surprising. That is, when a mis-assigned peak is involved in a d_{NN}, it forces additional peaks to be mis-assigned. In contrast, because NVR continues to use *all* experimental data during phase 2, deficiencies in the tensor estimates due to model error or dynamics can be overcome. The integration of different lines of evidence is the essence of the NVR method.

Given the sensitivity to establishing correct assignments early, one could incorporate additional experimental evidence to reduce the overall uncertainty in the data. Selective isotopic labelling is an obvious possibility. Additionally, NVR might be adapted to use the techniques of Grzesiek and Bax (1993) or Gemmecker et al. (1993) to identify solvent-accessible protons. This data would be complementary to the amide-exchange data already used by NVR: Zero or ϵ-edge weights would be formed between peaks associated with solvent-accessible protons and the non-solvent accessible residues from the model.

It is reasonable to consider alternative means for estimating alignment tensors, such that phase 1 might be skipped entirely. For example, structure-based tensor estimation methods (e.g., Zweckstetter and Bax, 2000) and a technique reported by Zweckstetter (2003), based on selective isotopic labelling, have been reported. To test this scenario, we skipped phase 1 (making no assignments) and instead passed the correct alignment tensor to phase 2, and subsequently assigned all the peaks. The mean and median accuracy for these experiments, over all 20 models, were 97% and 100%, respectively. This suggests that, in the case where NVR were to make some incorrect assignments, but the final tensor estimates were still reasonably accurate, one could increase assignment accuracies by a 'bootstrapping' procedure in which phase 2 was re-run using the final tensor estimates but discarding the assignments from the first run. To test this, we ran 100 trials on each of the 20 models. On each trial, we constructed tensors using the SVD method on the model and a set of assignments, $x\%$ of which were *incorrect*, in lieu of running phase 1; after discarding the assignments, the algorithm then proceeds to phase 2. These experiments were run for $x \in \{1, 3, 6, 12, 25, 50, 100\}$. Thus, we ran 14,000 trials (20 proteins $\times$ 100 random selections $\times$ 7 percentages). The bootstrapping procedure is very robust; the mean accuracies for $x = 1, 3, 6, 12, 25, 50$, and 100 were 97%, 97%, 98%, 98%, 98%, 98%, and 96%, respectively. The median accuracies for all values of x were all 100%.

These results again suggest that NVR is not very sensitive to the quality of the initial tensor estimates, because the additional lines of evidence (chemical shift prediction, amide-exchange, d_{NN}'s) can overcome these inaccuracies. NVR's voting algorithm (see p. 135) to integrate different lines of evidence is really just a means to increase a signal-to-noise ratio. Here the signal is the computed likelihood of the assignment

between a peak and the (correct) residue. Ideally, this probability would be 1. The noise is the uncertainty in the data such that the probability mass is distributed among multiple residues. Each line of evidence (i.e., experiment) has noise, but the noise tends to be random and thus cancels when the lines of evidence are combined. Conversely, the signals embedded in each line of evidence tend to reinforce each other, resulting in (relatively) unambiguous assignments. Hence, even if the two initial tensor estimates are poor, it is unlikely that they can conspire (by voting together) to force an incorrect assignment. More generally, given NVR's voting scheme (p. 135), any pair of lines of evidence is unlikely to outvote the majority.

Detecting incorrect assignments

It is reasonable to ask whether the difference between experimentally-observed and back-computed RDCs may be used to identify incorrect assignments. Unfortunately, the difference between experimentally observed and back-computed RDCs can be quite large due to dynamics, discrepancies between the idealized physics and the conditions in solution, and model error. In general, there may be no bound on the disparity between a experimentally-observed and a back-computed RDC. Thus, the difference between experimentally observed and back-computed RDCs cannot be used to used to identify any particular assignment as incorrect. However, the difference between experimentally observed and back-computed RDCs might be useful to assess the overall correctness of an ensemble of assignments. Recall that the alignment tensor for a given medium can be computed via SVD from a set of 5 or more assigned RDCs matched to a known substructure. In theory, *any* set of five assigned RDCs should yield the same tensor; in practice, noise and dynamics yield somewhat different tensors for different subsets of assignments. Consider a set of assignments for a set of $n \gg 5$ peaks (and their associated RDCs). We will denote as A_c a set of assignments for n peaks containing c incorrect assignments. For example, the set A_0 is the set of correct assignments, and A_2 is a set containing 2 erroneous assignments. At issue is whether tensors induced by subsets of A_0 are more consistent (similar) than tensors induced by subsets of A_c, when $c > 0$. To test this hypothesis, we ran a series of experiments. For each of the twenty models, we constructed sets A_c. Incorrect assignments were chosen randomly. Next, we constructed two tensors, one for each set of RDCs, for each A_c. Let $\mathbf{S}_{cm}$ be the tensor constructed using A_c and RDCs from medium m. Let B_c^k be a subset of A_c, such that $|B_c^k| = k$, where $k \geq 5$. Let $\mathbf{S}_{cm}^k$ be the tensor constructed using B_c^k using RDCs from medium m. We randomly selected 100 subsets of size k for each A_c and for each subset we then constructed two tensors, $\mathbf{S}_{c,1}^k$, and $\mathbf{S}_{c,2}^k$, one for each medium.

We then computed the similarity between the *all-residue tensor* $\mathbf{S}_{cm}$ and each *subset tensor* $\mathbf{S}_{cm}^k$ as follows. Saupe matrices are completely specified by their eigenvalues and eigenvectors. Following standard notation (Wedemeyer et al., 2002), we sort the eigenvectors by eigenvalue. We then compare eigenvalues and eigenvectors of the same rank. To quantify the similarity of the principal order frame (POF) of the subset tensor $\mathbf{S}_{cm}^k$ to the POF of the all-residue tensor $\mathbf{S}_{cm}$, we use the method of Yan et al. (2003) to compute a percentile that measures the fraction of all tensor orientations that fall within the angular deviations of the subset tensor ($\mathbf{S}_{cm}^k$) from the all-residue tensor ($\mathbf{S}_{cm}$). Suppose we randomly and isotropically rotate the subset tensor $\mathbf{S}_{cm}^k$. We compute the probability P_s, that the eigenvectors of the randomly-oriented tensor are simultaneously within the three angular errors measured between the eigenvectors of $\mathbf{S}_{cm}$ and the eigenvectors of $\mathbf{S}_{cm}^k$. By integrating isotropically over $SO(3)$, we compute an upper bound on P_s, which includes a 4-fold symmetry factor due to symmetry of the dipolar operator.[2] Thus, for each of the twenty models and a given pair of c and k, we obtain 200 similarity measurements (100 $\times$ two media). Let T_c^k be the set of these 4,000 (20 models $\times$ 200) measurements. We will denote the average, maximum and minimum of T_c^k as μ_c^k, κ_c^k, and φ_c^k, respectively. In our experiments, $c \in \{0, 2, 4, 8, 16, 32, n\}$ and $k \in \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$ and we computed T_c^k, μ_c^k and φ_c^k for all combinations of c and k.

As shown in Figure 5, average tensor consistency (μ_c^k) is inversely correlated with the number of incorrect assignments. That is, for any fixed k, if $x > y$, then $\mu_x^k < \mu_y^k$. The average tensor consistency increases with k. This is to be expected; suppose that a single peak p in subset B_c^k is incorrectly assigned to a residue whose backbone amide bond vector is oriented in a different direction than the bond vector associated with the *correct* assignment. The SVD method for computing alignment tensors finds the tensor that minimizes the sum of the squares of the differences between the experimentally observed RDCs and the set of back-computed RDCs. That is, it is governed by a quadratic error function. Conceptually, the incorrect

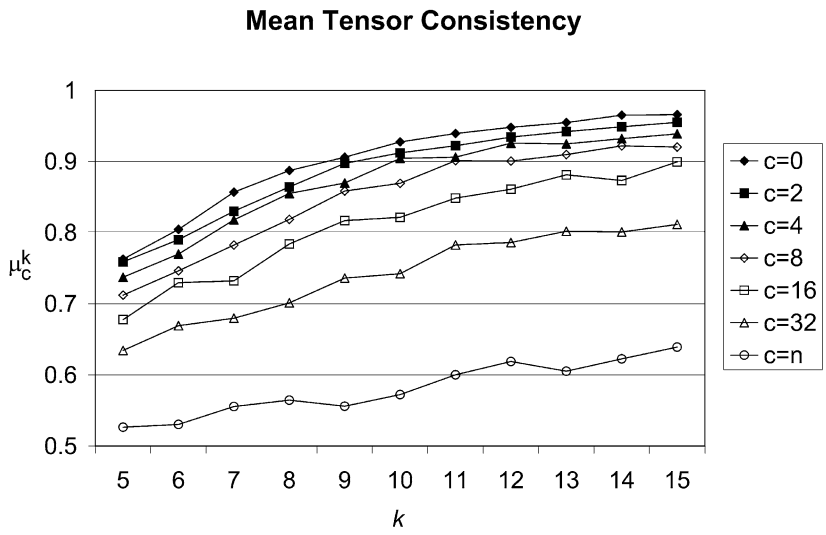

Figure 5. Mean tensor consistency and incorrect assignments: Average tensor consistency, μ_c^k, is inversely correlated with the number of incorrect assignments, c, and positively correlated with the size of the subset, k, (see text for details).

assignment represents an outlier that the SVD method is forced to fit. The larger an outlier p is, the more $\mathbf{S}_{cm}^k$ differs from $\mathbf{S}_{cm}$. However, if we increase the size of k, the effect that any single incorrect assignment, p, has on $\mathbf{S}_{cm}^k$ decreases, and the consistency between $\mathbf{S}_{cm}$ and $\mathbf{S}_{cm}^k$ increases.

In contrast, as shown in Figure 6, minimum tensor consistency (φ_c^k) is not a good indicator of the relative amount of incorrect assignments. This is also to be expected; as argued above, an incorrect assignment tends to generate an 'incorrect' tensor. Thus, if a given randomly chosen subset has a significant number of incorrect assignments, the tensor associated with that subset will be inconsistent with $\mathbf{S}_{cm}$. Perhaps more important, φ_0^k, the minimum tensor consistency value observed when the subsets are drawn from A_0 (the set of correct assignments), can also be very low, even for large k. As previously mentioned, it is not possible to distinguish incorrect assignments from anomalous RDCs or model error. Thus, it is not unexpected that we observe low values of φ_0^k. Maximum tensor consistency (κ_c^k) is also not a good indicator of the relative number of incorrect assignments (data not shown). In all of our experiments, regardless of the sizes of c and k, we always observed at least one pair, ($\mathbf{S}_{cm}$, $\mathbf{S}_{cm}^k$), whose consistency fell into the 100$^{\text{th}}$ percentile of accuracy. That is, when randomly selecting subsets of A_c, it is likely that at least one subset will yield a tensor consistent with $\mathbf{S}_{cm}$.

As we have seen, no simple threshold applied to either φ_0^k or κ_c^k is sufficient to determine whether a set of assignments, A, contains errors. Average tensor consistency (μ_c^k), on the other hand, is correlated with the number of incorrect assignments. As seen in Figure 5, for any given value of k, a threshold can be chosen that perfectly separates A_0 from A_2, A_4, A_8, A_{16}, A_{32}, and A_n. However, this threshold varies with k. It is doubtful that the threshold might be chosen analytically; the specific value of μ_c^k almost certainly depends on the structural homology between the model and the target protein, which is unknown *a priori*. However, the correlation between tensor consistency and correctness of assignments does suggest a new strategy for performing resonance assignment wherein the goal is to maximize μ_c^k. This is an interesting area for future work.

Finally, it is worth considering the sensitivity of μ_c^k to the absolute number of errors. To test this, we computed t-tests between T_0^k and T_c^k, for all k and all $c > 0$. That is, we wish to determine for which value of k do μ_0^k and μ_c^k become statistically significantly different. Table 16 reports the p-values from a one-tailed t-test at a significance level of 0.05. Statistically significant differences between T_0^k and T_c^k, for all k and all $c \geq 4$ are seen. That is, μ_c^k is sensitive to 4 or more incorrect assignments in subsets of $k = 5$ and greater. Statistically significant differences between T_0^k and T_c^k, for all $k \geq 6$ and all $c \geq 4$ are seen. Thus, using subsets of size 6, or greater, one could, in principle, distinguish between a correct set of assignments and an incorrect one using tensor consistency alone. Unfortunately, as previously argued,

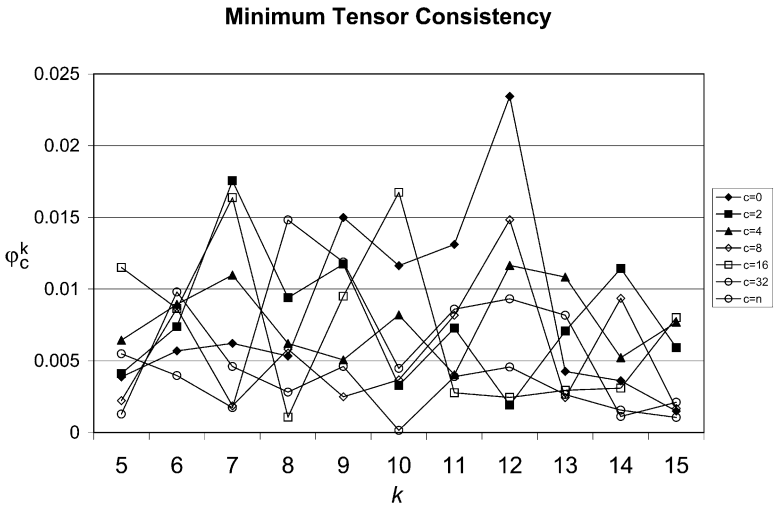

Figure 6. Minimum tensor consistency and incorrect assignments: Minimum tensor consistency, φ_c^k, is not correlated with the either number of incorrect assignments, c, or with the size of the subset, k, (see text for details).

Table 16. Mean tensor consistency sensitivity. p-values, as computed using a one-tailed t-test, with a significance level of 0.05, between T_0^k and T_c^k, for all k and all $c > 0$ (see text for definitions). Only one combination ($c = 2$,$k = 5$) yields a statistically insignificant difference in the means of the two populations

	c					
k	2	4	8	16	32	n
5	**0.27**	4.1×10^{-5}	0	0	0	0
6	8.8×10^{-3}	0	0	0	0	0
7	1.0×10^{-6}	0	0	0	0	0
8	2.0×10^{-6}	0	0	0	0	0
9	0.03	0	0	0	0	0
10	4.2×10^{-5}	0	0	0	0	0
11	2.0×10^{-6}	0	0	0	0	0
12	2.5×10^{-5}	0	0	0	0	0
13	2.4×10^{-5}	0	0	0	0	0
14	0	0	0	0	0	0
15	2.8×10^{-5}	0	0	0	0	0

tensor consistency alone cannot be used to identify *which* assignments are incorrect.

Discussion

It is reasonable, in principle, to cast the problem of resonance assignment of a known structure using RDCs, into a combinatorial optimization framework (Hus et al., 2002). Hence, initially, we attempted to treat the problem as an optimal bipartite matching problem. NVR operates on bipartite graphs between peaks and residues. The edge weights from each peak to all residues form a probability distribution. The probabilities are derived from 1) amide-exchange experiments, 2) d_{NN}s, 3) chemical shift predictions based on average chemical shifts from the BMRB (Seavey et al., 1991), 4) chemical shift predictions made by the program SHIFTS (Xu and Case, 2001), 5) chemical shift predictions made by the program SHIFTX (Neal et al., 2003), and 6-7) constraints from RDCs in two media. As shown in Figure 7, maximum bipartite matching (Kuhn, 1955) does not yield satisfactory accuracies on various combinations of these graphs. No combination achieves higher than 53% accuracy and the mean accuracy is only 11%. In other words, neither amide RDCs, nor any of the chemical shift prediction methods provide enough constraint to yield accurate assignments using maximum bipartite matching alone. Of course, neither SHIFTX nor SHIFTS are intended to perform resonance assignment directly. Brüschweiler and co-workers (Hus et al., 2002) have successfully applied maximum bipartite matching to resonance assignment, but that technique requires RDCs from several different bond types which, in turn, requires ^{13}C-labelling of the protein and triple resonance experiments. One of the initial goals of NVR was to answer the question, are backbone amide RDCs and d_{NN}s sufficient for performing resonance assignment. Figure 7 does not imply that (Hus et al., 2002) performs poorly. Rather, it implies that the EM method

may be more effective than bipartite matching alone on the different, and sparser set of experimental data employed in NVR.

RDCs may exhibit both *degeneracy*, when two or more residues have the same RDC, and *aberrations*, when experimentally recorded residual dipolar couplings deviate significantly from their predicted values. RDC degeneracies arise due to the geometry of the protein and the dipolar operator. Aberrant RDCs 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.

Chemical shifts, similarly, are also subject to degeneracy (peak overlap) and aberrations, when a given residue's amide chemical shifts fall outside their expected range. However, as also observed in (Hus et al., 2002), by combining RDCs and chemical shifts, both degeneracy and aberrations may be partially overcome. For example, in Figure 7, a maximum bipartite matching on the joint probabilities of chemical shift prediction and RDCs does better than chemical shift prediction or RDCs alone.

The advantage of NVR over maximum bipartite matching lies in its iterative nature. The algorithm takes a conservative approach, making only likely assignments, given the current information. After making these assignments, the edge weights between the remaining unassigned peaks and residues are updated. Suppose that, during the ith iteration of the algorithm, peak k is assigned to residue r. The edge weight between peak k and residue r is then set to 1, indicating the certainty of that assignment. As previously stated, the edge weights form a probability distribution. Accordingly, the edge weight between peak k and any other residue is set to 0. Similarly, the weights on the edges from any *other* peak to r are immediately set to 0. The (non-zero) edge weights from each remaining unassigned peak are re-normalized prior to the next iteration. Thus, a peak whose assignment may be ambiguous in iteration i may become unambiguous in iteration $i + 1$. The unassigned d_{NN}s play an especially important role in disambiguating competing assignments (Figure 8).

Limitations

There are a number of limitations to our algorithm worth noting. The first is that we have only tested NVR on models with both high sequence and structural homology. Consequently, the present form of the algorithm may be best applied to scenarios where a crystal structure of the same protein is available, as may be the case in a SAR by NMR study. Models with significantly less homology will likely have somewhat different networks of hydrogen bonds and NOEs, as well as different amide bond-vector orientations. The probabilistic framework in which RDCs are interpreted will likely be robust to reasonable amounts of variation. In contrast, the hard constraints employed by NVR in the interpretation of amide-exchange rates and d_{NN}'s will likely force assignment errors in these cases. In terms of differences in hydrogen bonding patterns, NVR is not very sensitive to the total number of slow-exchanging peaks (as seen in Table 14). The difficulty is when a peak which *should* be fast exchanging (with respect to the model) is identified as being slow-exchanging. Our followup experiment (page 123) wherein we set the edge-weights between slow-exchanging peaks and labile protons to a small value ϵ (instead of 0), performed well (median accuracy 98%). Thus, a modified form of NVR might adopt this approach. In terms of differences in NOE networks, the majority of all d_{NN}'s processed in our experiments were sequential. Thus, little or no compensation is required for low-homology models. Long range d_{NN}'s, in contrast, might be handled by using a larger tolerance radius (see Methods section).

A more comprehensive analysis of the performance of the algorithm under varying amounts of homology (both structural and sequential) remains an important goal. Computational modelling could be used to construct a variety of alternative models having strictly controlled amounts of homology. These models may inform, for example, the minimum amount of homology required by NVR for a given set of experimental data. NVR is perhaps best seen as a framework; one instance of that framework is defined in Table 1. NVR, could, however, be easily adapted to include more, and different kinds of NMR data. Homology studies might be used to define, for a given level of homology, which experimental data are required.

Finally, NVR relies on the ability to make an initial set of assignments using chemical shift predictions. Clearly, the ability to do so is a function of which amino acids comprise the protein and the peak density

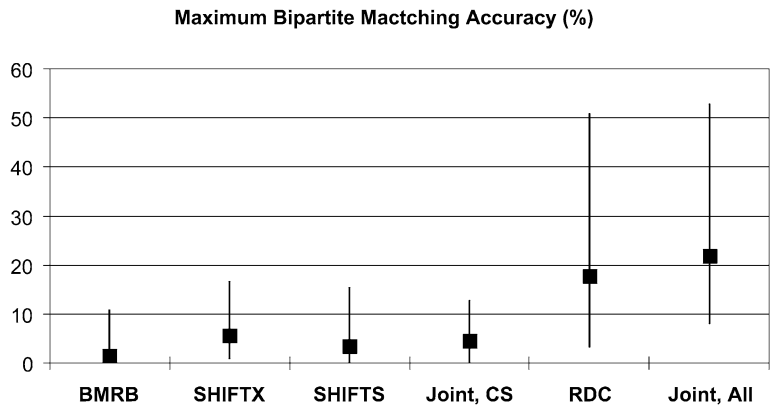

Figure 7. Maximum bipartite matching accuracy: Comparison of the accuracy of the Kuhn–Munkres maximum bipartite matching algorithm on various combinations of the data processed by NVR. Column 1, BMRB, is a bipartite graph with edge weights computed using chemical shift prediction based on statistics from the BMRB. Columns 2 and 3, SHIFTX and SHIFTS, are bipartite graphs with edge weights computed using the chemical shift prediction programs SHIFTX and SHIFTS, respectively. Column 4, JOINT, CS, is a bipartite graph with edge weights computed by taking the joint probability (Eq.(7)) over BMRB, SHIFTX and SHIFTS graphs. Column 5, RDC, is a bipartite graph with edge weights computed by taking the joint probability of both sets of RDCs. Column 6, JOINT, ALL, is a bipartite graph with edge weights computed by taking the joint probability over BMRB, SHIFTX, SHIFTS, and both sets of RDCs using the correct alignment tensors. Squares are the mean accuracy over the 3 proteins and 20 test models. Vertical bars indicate the range of the maximum and minimum accuracies.

in various regions of the HSQC spectrum, neither of which can be controlled. Should chemical shift prediction incorrectly assign one or more peaks in phase 1, the tensors constructed using these assignments would be wrong, causing still more incorrect assignments to be made. We are presently exploring the incorporation of data from an ^{15}N-HSQC-TOCSY to provide more robust identification of amino acid type from chemical shift data. A related issue is that the NVR algorithm has no means for controlling *which* residues are assigned during phase 1. In particular, it is possible that the set of residues assigned during phase 1 may correspond to a set of bond vectors which lack sufficient independence to construct an accurate tensor. We note, however, that during phase 2, chemical shift predictions are also used, compensating for inadequacies in the tensors. Moreover, the tensor estimates are refined on every iteration, as more assignments are made. Hence, if the bond vectors from the model are reasonably independent, the algorithm is guaranteed to eventually construct a tensor using a set of vectors such that the tensor is determined accurately. Moreover, the algorithm could be modified such that it exits phase 1 only when the independence condition is met.

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. The 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 sparse d_{NN}s of the 3D ^{15}N-NOESY spectrum, in $O(n^3)$ time. We tested NVR on 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 100% assignment accuracy. Similarly good results were obtained in experiments with streptococcal protein G (99%) and hen lysozyme (100%)

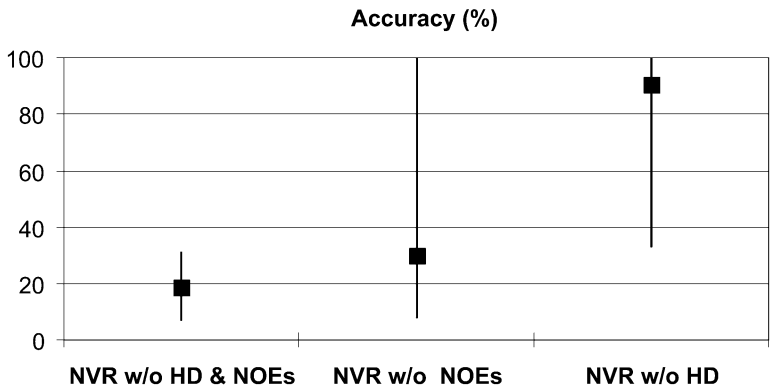

Figure 8. Relative importance of unassigned d_{NN}s in NVR: Comparison of the accuracy of NVR on various combinations of the data processed by NVR. Column 1, the result of running NVR with chemical shift prediction and RDCs, but without H-D exchange rates or d_{NN}s. Column 2, the result of running NVR with chemical shift prediction, RDCs, and H-D exchange rates, but without d_{NN}s. Column 3, the result of running NVR with chemical shift prediction, RDCs, and d_{NN}s, but without H-D exchange rates.

when they were matched by NVR to a variety of 3D structural models.

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 (Altschul et al., 1990), protein threading (Lathrop and Smith, 1996; Xu et al., 2000), or homology modelling (Blundell et al., 1987; Fetrow and Bryant, 1993; Greer, 1991; Johnson et al., 1994; Sali et al., 1990), one would modify NVR to perform assignments by matching RDCs to an homologous structure. Thus, NVR could play a role in structural genomics.

Software

The NVR software is available by contacting the authors, and is distributed under the Gnu Public License (Gnu, 2002).

Methods

Data and preprocessing

Atomic coordinates for the 20 trial structures (Tables 2–4), amide exchange data and residual dipolar coupling data for 1D3Z, 1E8L and 3GB1 were obtained from the PDB (Berman et al., 2000). The unassigned ^{15}N-HSQC and ^{15}N-edited HSQC-NOESY peak list for ubiquitin were obtained from the Driscoll lab (Harris, 2002). ^{15}N-HSQC peaks were cross-referenced to diagonal NOESY peaks manually. NVR uses only *unambiguous* d_{NN}'s; unambiguous d_{NN}s were extracted from the unassigned NOESY peak list. Unambiguous was defined as a NOESY cross-peak whose two ^{1}H and one ^{15}N chemical shifts could be cross-referenced to a unique pair of peaks in the HSQC. Peaks whose ^{1}H shifts were less than 0.01 ppm apart, or whose ^{15}N shifts were less than 0.1 ppm apart were said to be *ambiguous*. We obtained 42 unassigned d_{NN}s, for ubiquitin. This amounts to fewer than one NOE per residue. It is interesting to ask what percentage of the NOEs used by NVR are sequential NOEs. Of the unambiguous d_{NN}'s used by NVR for ubiquitin, 64% are sequential. For SPG and Lysozyme, the percentages are 60% and 53%, respectively. For ubiquitin, the unambiguous d_{NN}'s used by NVR comprise 26% of all sequential NOEs. For SPG and Lysozome, the percentages are 30% and 31%, respectively.

Amide exchange rates were obtained from the restraints files of the PDB structure 1D3Z. Specifically, the backbone amides identified as hydrogen-bonded restraints were defined to be slow-exchanging. Table 17 summarizes the number of slow-exchanging peaks processed by NVR. The use of these hydrogen-bonds represent something of an idealized scenario. The implicit assumption is that the protection factors computed from amide-exchange experiments include these residues. We note, however, that it is also possible that the hydrogen bonded restraints in the '.mr' file may, in fact, be a *subset* of the peaks that might be identified as slow-exchanging in a standard amide-exchange study. This is not a problem; the results of our perturbation studies (Tables 14 and 15, pp. 123–124) indicate that NVR is only sensitive to mislabelling labile protons as

Table 17. Slow-exchanging peaks processed by NVR

Protein	Number slow-exchanging peaks
Ubiquitin	27
SPG	34
Lysozyme	27

slow-exchanging. Protons involved in hydrogen bonds (with other residues) are, by definition, not labile. Conversely, NVR is insensitive to non-labile protons being labelled as fast-exchanging. Thus, by using the hydrogen-bond restraints as a proxy for amide exchange rates, NVR is arguably using *less* information than might be obtained from an amide exchange study.

Chemical shifts for SPG and Lysozyme were obtained from the BMRB (Seavey et al., 1991). NOEs for SPG and Lysozyme were obtained from the restraints files from the PDB structures 3GB1 (Kuszewski et al., 1999) and 1E8L (Schwalbe et al., 2001). Amide exchange rates for SPG and Lysozyme were obtained from the restraints files from the PDB structures 1GB1 (Gronenborn et al., 1991)[3] and 1E8L (Schwalbe et al., 2001). Once again, peaks whose ^{1}H shifts were less than 0.01 ppm apart, or whose ^{15}N shifts were less than 0.1 ppm apart were said to be ambiguous, and their corresponding NOEs were not used in the experiments.

Pseudocode

Pseudocode for our algorithm is given in Algorithm listings 1–4, and described formally in the following sections. The pseudocode uses the symbols defined in the sections below.

Initialization

During the initialization step, 5 weighted bipartite graphs are constructed. Let R be the set of residues in the model (removing prolines and the N-terminus). For each residue $r \in R$, $\texttt{AAType}(r)$ returns the amino acid type of residue r, and $\texttt{SSType}(r)$ returns the secondary structure type of r. The amino acid and secondary structure types are used to predict the backbone amide chemical shifts. Let K be the set of peaks in the HSQC. The chemical shifts of each peak $k \in K$ are given by $\omega(k) = (\omega_H(k), \omega_N(k))$, where $\omega_H(k)$ and $\omega_N(k)$ are the amide proton and nitrogen chemical shifts, respectively. The difference between these experimentally determined chemical shifts and the set of predicted chemical shifts are converted into assignment probabilities.

Each bipartite graph is defined as follows: $B = \{K, R, E\}$, where $E = K \times R$. Each edge $e \in E$ is weighted, $w : K \times R \rightarrow \mathbb{R}^+ \cup \{0\}$. The edge weights from each peak $k \in K$ are normalized so that they form a probability distribution. If there are missing peaks in the HSQC then $|K| < |R|$. In this case *dummy* peaks are added to the set K until $|K| = |R|$. Finally, the program MOLMOL (Koradi et al., 1996) was used to add amide protons to the x-ray structures and to identify any amide protons that are involved in hydrogen bonds or not solvent accessible. The hydrogen bonded and non-solvent accessible amide protons are correlated with H-D exchange rates.

Amide exchange constraints

Amide exchange rates are treated as binary geometric constraints. Specifically, non-hydrogen bonded surface residues (as determined by the program MOLMOL on the input model) are assigned a zero edge weight to any slow-exchanging peak. A uniform probability is given to any non-zero edges from a given peak. Let B_{HD} be the bipartite graph constructed using the amide exchange data. Constructing B_{HD} takes $O(n^2)$ time. B_{HD} is only used during the initialization step of NVR to restrict the set of possible assignments.

We note that it is not necessary to compute exact amide-exchange rates. Rather, NVR requires only that slow-exchanging peaks be identified. These could be obtained, for example, as follows: While the HSQC, 2 RDC and NOESY spectra are being recorded, a second sample of the protein can be, in parallel, suspended in 100% D_2O allowing amide exchange to occur. The recording, processing, peak picking and cross-referencing of the HSQC, 2 RDC, and NOESY spectra (in preparation for the NVR algorithm) will certainly require at least a day. Thus, the second lyophilized sample of the protein (in D_2O) will have ample time for exchange to occur. A final HSQC spectrum can then be recorded on the D_2O sample. It is to be expected that all labile protons will have exchanged by this point. Thus, only peaks from slow-exchanging protons will remain in the HSQC of the D_2O sample. One may simply choose a threshold to identify slow-exchanging peaks. For example, if a given peak in the D_2O sample has at least 25% of the volume of the corresponding peak in the first HSQC, then that peak

Algorithm 1 Pseudocode for NVR. All symbols are defined in the Methods section.

Input:

$K \leftarrow$ Peak-list of the H^N-^{15}N HSQC Spectrum
$NOES \leftarrow$ Peak-list of the H^N-^{15}N HSQC-NOESY Spectrum
$RDC_1 \leftarrow$ Dipolar couplings in the RDC Spectrum (medium 1)
$RDC_2 \leftarrow$ Dipolar couplings in the RDC Spectrum (medium 2)
$HD \leftarrow$ Slow-exchanging HSQC peaks, as determined by H-D exchange experiments
$D \leftarrow$ structural model of the protein.

/ Preprocessing Steps */*
$R \leftarrow$ `Extract-Residues(D)` */* excluding prolines and N-terminus */*
$T \leftarrow$ `Extract-Residue-Type-and-Secondary-Structure(D)`
$ABV \leftarrow$ `Extract-and-Normalize-Backbone-Amide-Vectors(D)`
$U \leftarrow$ `Extract-Non-Hydrogen-Bonded-Surface-Residues(D)`
$d_{NN}s \leftarrow$ `Extract-Unambiguous-DNNs(NOES, K)`

/ Initialization Steps */*
$B_{HD} \leftarrow$ `Build-Bipartite-Graph(K,R,U,HD)`
$B_{NOE} \leftarrow$ `Build-Bipartite-Graph(K,R,ABV,`d_{NN}`s)`
$B_{BMRB} \leftarrow$ `Build-Bipartite-Graph(K,R,Chemical-Shift-Prediction(T,BMRB))`
$B_{SHIFTS} \leftarrow$ `Build-Bipartite-Graph(K,R,Chemical-Shift-Prediction(D,SHIFTS))`
$B_{SHIFTX} \leftarrow$ `Build-Bipartite-Graph(K,R,Chemical-Shift-Prediction(D,SHIFTX))`
`Synchronize-Graphs(`$B_{HD}, B_{NOE}, B_{BMRB}, B_{SHIFTS}, B_{SHIFTX}$`)`
$\Theta \leftarrow \varnothing$
$n \leftarrow |R|$

/ Phase 1 */*
while $|\Theta| < 5$ **do**
$V \leftarrow$ `E-Step(`$B_{BMRB}, B_{SHIFTS}, B_{SHIFTX}$`)`
$\Theta \leftarrow \Theta \cup$ `M-Step(`V`)`
$B_{NOE} \leftarrow$ `Apply-NOE(`B_{NOE}`)`
`Synchronize-Graphs(`$B_{NOE}, B_{BMRB}, B_{SHIFTS}, B_{SHIFTX}$`)`

/ Phase 2 */*
while $|\Theta| < n$ **do**
/ Compute Tensors and Build RDC-based bipartite graphs */*
$S_1 \leftarrow$ `Compute-Tensor(`Θ`, ABV,` RDC_1`)`
$S_2 \leftarrow$ `Compute-Tensor(`Θ`, ABV,` RDC_2`)`
$M_1 \leftarrow$ `Build-Bipartite-Graph(K,R,ABV,`RDC_1, S_1`)`
$M_2 \leftarrow$ `Build-Bipartite-Graph(K,R,ABV,`RDC_2, S_2`)`

/ Make Assignments using E/M */*
$V \leftarrow$ `E-Step(`$B_{BMRB}, B_{SHIFTS}, B_{SHIFTX}, M_1, M_2$`)`
$\Theta \leftarrow \Theta \cup$ `M-Step(`V`)`
$B_{NOE} \leftarrow$ `Apply-NOE(`B_{NOE}`)`
`Synchronize-Graphs(`$B_{NOE}, B_{BMRB}, B_{SHIFTS}, B_{SHIFTX}$`)`

return Θ

Algorithm 2 Pseudocode for E step. All symbols are defined in the Methods section. Note that $|\mathcal{B}|$ is at most 5, so $|2^{\mathcal{B}}|$ is a constant.

Input:
$B_1, B_2, ..., B_m$ /* *m bipartite graphs* */
K /* *set of peaks in the HSQC* */
R /* *set of residues in the protein* */

/* *Initialization Steps* */
Let $\mathcal{B} = \{B_1, B_2, ..., B_m\}$
Let 2^B be the power set (set of all subsets) of $\mathcal{B}$
Let $E = K \times R$
$V = \{K, R, E\}$ /* *build a bipartite graph* */
for all $e \in E$ **do**
$w(e) \leftarrow 0$ /* *initialize all edge-weights to 0*/

/* *Compute Expectation Graph* */
for all $\mathcal{B}_i \in 2^{\mathcal{B}}$ **do**
E_i =`Kuhn-Munkres(Combine-Graphs(`$\mathcal{B}_i, K, R$`))` /* $E_i \subset E$ */
for all $e \in E_i$ **do**
$w(e) \leftarrow w(e) + 1$ /* *increment edge weight for V* */

return V

Algorithm 3 Pseudocode for Combine-Graphs All symbols are defined in the Methods section.

Input:
$\mathcal{B}$ /* *A set of bipartite graphs* */
K /* *set of peaks in the HSQC* */
R /* *set of residues in the protein* */

/* Initialization Steps */
$C = \{K, R, E\}$ /* build a bipartite graph */
for all $e \in E$ **do**
$w(e) \leftarrow \prod_{b \in \mathcal{B}} w_b(e)$ /* compute joint probability for C */

return C

Algorithm 4 Pseudocode for M Step. All symbols are defined in the Methods section.

Input:
V /* *Expectation Graph* */

/* *Make assignments* */
Let w_{max} =`Find-Maximum-Edge-Weight(`V`)`
$\Theta^* \leftarrow w^{-1}(w_{max})$

return Θ^*

is labelled as being slow-exchanging. More conservative estimates can be obtained by setting this threshold higher. Alternatively, NVR already uses MOLMOL to predict which backbone amide protons are likely to be slow exchanging. Thus, the threshold could be set such that the expected number of peaks is (roughly) equal to the number of slow-exchanging amide protons, as predicted by MOLMOL. Furthermore, two thresholds might be used. The first (conservative) threshold might be set to 50%. Any peak in the deuterated HSQC that has 50%, or more, of the volume in the non-deuterated HSQC will be processed as stated above. The second threshold, say 25%, could be used as follows: For any peak volume that falls between the 25% and 50% threshold, we label it as being slow-exchanging, but apply a 'soft-threshold'. That is, rather than setting the edge probability between these peaks and labile protons to 0, we instead set the probability to some small value $\epsilon > 0$. Our perturbations studies on corrupted amide exchange data (page 123) suggest that NVR is robust to significant noise and error in the amide exchange data when ϵ-weights are used.

d_{NN} *Constraints*

d_{NN}s are also interpreted as binary 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 δ Å of r in the model. In our experiments, δ varied by the size of the protein, ranging from 5 Å to 8 Å. The larger tolerance for the largest protein (Lysozyme), compensates for crystal contacts and conformational differences between the protein in solution versus in the crystal. Larger proteins may have larger violations of d_{NN}s, particularly non-sequential d_{NN}s. The average number of non-sequential residues in our δ-radius were 2.0, 3.5, and 7.9 for SPG, Ubiquitin, and Lysozyme, respectively. Let B_{NOE} be the bipartite graph built constructed using the NOESY data. Constructing B_{NOE} takes $O(n^2)$ time.

Initially, B_{NOE} is a complete bipartite graph, with all edge weights set to $1/n$ where n is the number of residues in the model. However, edges with weight 0 in B_{HD} are immediately set to 0 in B_{NOE} as well, because the H-D exchange exchange data indicates that assignment is impossible. These H-D exchange-derived constraints are propagated throughout B_{NOE} as follows. Let $k \in K$ be a peak with one or more zero-weight edges (due to H-D exchange-derived constraints). Let $Q \subset R$ be the set of residues with non-zero edge weights to k. That is, Q is the set of possible assignments for k. Let $T \subset R$ be the set of residues within δ Å of any element of Q. Let $U \subset K$ be the set of peaks that have a d_{NN} with k. For each $u \in U$, let $V_u \subset R$ be the set of residues with non-zero edge weights to u. We then set $V_u \leftarrow V_u \cap T$.

This procedure for propagating assignment constraints via the d_{NN}s is run after each iteration of the Expectation/Maximization algorithm (described below). We will call this procedure `ApplyNOE`(B_{NOE}) which takes as input B_{NOE} and updates B_{NOE} in the manner just described. `ApplyNOE`(B_{NOE}) takes $O(n^2)$ time because each peak is considered once, in some fixed order. If a peak has a d_{NN}, then it forces an update of at most a constant number of other peaks, due to geometric constraints and the sparseness of the unambiguous d_{NN}s. Each update may require updating $O(n)$ edges.

Chemical shift constraints

A training set of 457 different protein structures (solved by NMR) and their associated chemical shifts as deposited in the BMRB (Seavey et al., 1991) were obtained. None of the 3 proteins in our test set (Ubiquitin, Lysozyme, SPG) were present in this set. The average and standard deviation of the amide proton and nitrogen chemical shift was computed for each amino acid, in each secondary structure type (α, β, coil). In addition, the maximum number of standard deviations above and below the mean was recorded for the training set. These statistics were used to construct edge weights on a bipartite graph:

$$w(k, r) = \mathbf{P}(k \mapsto r) = f(k, r) \tag{4}$$

where $k \in K$ and $r \in R$. Here,

$$\begin{aligned} f(k, r) \;=\; & \mathcal{N}(\omega_H(k) - \mu_H(r), \sigma_H(r))\mathcal{N}(\omega_N(k) \\ & - \mu_N(r), \sigma_N(r)). \end{aligned} \tag{5}$$

Consider the distribution of a subset of all chemical shifts in our training set with the same amino-acid and secondary structure type as r. $\mu(r)$ is a pair, $(\mu_H(r), \mu_N(r))$, corresponding to the mean amide proton and nitrogen chemical shifts observed in that distribution. $\sigma(r)$, similarly, is defined as the standard deviations of the amide proton and nitrogens of that distribution. The function $\mathcal{N}(x - \mu, \sigma)$ is the probability of observing the difference $x - \mu$ in a normal

distribution with mean μ and standard deviation σ. That is,

$$\mathcal{N}(x-\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right). \quad (6)$$

Thus, the probabilities are computed using two one-dimensional Gaussian distributions (one for proton shifts, one for nitrogen shifts) with means $\mu(r)$ and standard deviation $\sigma(r)$. We are thus implicitly assuming that the two dimensions are independent. More sophisticated treatments that model the covariance between the two dimensions are worth investigating. If a given amide proton or nitrogen shift is beyond the maximum number of standard deviations (as computed in the training set), its weight is set to zero. Let B_{bmrb} be the bipartite graph whose edges are computed using the statistics from the subset of the BMRB. Constructing B_{bmrb} takes $O(n^2)$ time.

The same training set was used to compute statistics on the accuracies of the programs SHIFTS (Xu and Case, 2001) and SHIFTX (Neal et al., 2003). The mean and standard deviation between the predicted amide proton and nitrogen shift, as well as the maximum number of standard deviations above and below the mean was computed for the training set. These statistics were used to construct two additional bipartite graphs, in the manner just described, between the peaks and residues in our test set. Let B_{shifts} and B_{shiftx} be the bipartite graphs whose edges are computed using the statistics computed from the programs SHIFTS and SHIFTX, respectively. Constructing B_{shifts} and B_{shiftx} takes $O(n^2)$ time.

When the sequence of the target protein is not 100% identical to the structural model, chemical shift predictions are made based on the amino acid type of the target protein, but the secondary structure type of the homologous model. For example, 1UD7 is mutant form of ubiquitin where 7 hydrophobic core residues have been altered (I3V, V5L, I13V, L15V, I23F, V26F, L67I). Chemical shift predictions were made using the BMRB statistics of the target protein's (1D3Z's) native sequence (I3, V5, I13, L15, I23, V26, L67) but the secondary structure types dictated by 1UD7.

Next, the graphs B_{NOE}, B_{bmrb} B_{shifts} and B_{shiftx} are *synchronized*. That is, any edge whose weight is zero in one graph is set to zero in *all* graphs. The NVR algorithm always ensures that all graphs are synchronized prior to any iteration. Using B_{bmrb} B_{shifts} and B_{shiftx}, the top γ candidate assignments for each peak are then extracted by first selecting, for each peak, the edge whose weight is larger than $n-\gamma$ of the remaining edges. This edge-weight is used as a threshold and the $n-\gamma$ edges whose edge-weights are smaller than the threshold are set to 0. Selecting the threshold can be done in $O(n)$ time ((Cormen et al., 2001), pp 189-192). In all our experiments we took $\gamma = 30$. Each of the γn selected edges selects a residue in R. Let $Q_\gamma \subset R$ be the subset of residues selected in this manner. Clearly the size of Q_γ is $O(n)$. The reduction from a quadratic to a linear number of edges makes each bipartite graph sparse. NVR repeatedly calls maximum bipartite matching as a subroutine. We used an implementation of the Kuhn–Munkres algorithm for maximum bipartite matching (Kuhn, 1955). On a complete graph, that algorithm runs in time $O(n^3)$, where n is the number of vertices in the graph (here, n is also the number of residues in the protein, excluding prolines and the N-terminus). Kuhn–Munkres considers every edge in the bipartite graph, performing $O(n)$ work per edge. A complete graph has $O(n^2)$ edges, so Kuhn–Munkres requires $O(n^3)$ time on a complete graph. In our algorithm, however, there are only $O(n)$ edges, thus, this same algorithm runs in time $O(n^2)$ on a sparse graph. The Kuhn–Munkres algorithm does assume that a complete matching is possible in the resulting graph. That is, each peak $k \in K$ must be assigned to a unique residue $r \in R$. In principle, if only the top γn candidate assignments are used then it is possible that $|Q_\gamma| < |R|$, thereby violating this assumption. However, it can be shown using a probabilistic argument (Cormen et al., 2001, pp. 109–110) that the constant $\gamma = 30$ will suffice for any $n \leq 10^{13}$, which more than adequately covers range of protein sizes accessible to NMR. Furthermore, from a purely complexity-theoretic point of view, the same argument says that the constant $\gamma = 3$ would suffice for every protein we tested; in practice we used the constant $\gamma = 30$ for robustness: since the NVR algorithm runs in only seconds to minutes, there is no practical performance distinction between using 3 versus 30. Finally, we note that the complete graph could be used as input, and that this would result in a slow-down by only a linear factor ($O(n)$).

Tensor determination (phase 1)

The input to the first phase are the bipartite graphs B_{NOE}, B_{bmrb}, B_{shifts} and B_{shiftx}. Let Θ be the master list of assignments. Initially, Θ is empty. Expectation/Maximization is used to make the first few assignments using B_{bmrb}, B_{shifts} and B_{shiftx}.

The E step is computed as follows. We construct a new bipartite graph V, called the *expectation graph*, whose edge weights are initialized to 0. Let $\mathcal{B} = \{B_{bmrb}, B_{shifts}, B_{shiftx}\}$ and $2^{\mathcal{B}}$ be the powerset of $\mathcal{B}$. Thus, each element $b \in 2^{\mathcal{B}}$ is a subset of $\mathcal{B}$. The size of $2^{\mathcal{B}}$ is $\sum_i^3 \binom{3}{i} = 7$. `Combine-Graphs` is a function that takes as input a set of bipartite graphs and returns a new bipartite graph. The edge weights of the output graph are the joint probabilities of the edges in the input graphs:

$$w(k, r) = \prod_{i \in b} w_i(k, r). \tag{7}$$

Let $C = \{\texttt{Combine-Graphs}(b) \mid b \in 2^{\mathcal{B}}\}$. Clearly, $|C| = |2^{\mathcal{B}}| = 7$. Furthermore, each $c \in C$ is sparse because the edges in c are the same as the edges in B_{bmrb}, B_{shifts}, and B_{shiftx}, which are each sparse. Only the weights are different (Equation 7).

For each combined graph $c \in C$, we compute a maximum bipartite matching. Let $H_c \subset E$ be the matching computed on c. For each edge $e \in H_c$ we increment the weight on the same edge in the expectation graph, V, by 1. This is done for all 7 bipartite matchings. Informally, each bipartite matching 'votes' for a set of edges. Thus, edge weights in V record the number of times a particular edge was part of a maximum bipartite matching. Note that the edge weights are probabilities in the bipartite graphs. Thus, a bipartite matching gives a maximum likelihood solution on that graph, which in turn maximizes the expected log-likelihood of the average edge weight.

Let w_{max} be the largest edge weight in V. The M step is computed, and assignments are made by $\Theta \leftarrow \Theta \cup w^{-1}(w_{max})$. As previously stated, each of the constituent votes used to construct V is a maximum likelihood solution – it maximizes the expected edge weight in the matching. Therefore, in the bipartite graph V, those edges with maximum votes have the highest expected values over all combinations of the data, thus satisfying the condition in Eq. (3). In our 20 experiments w_{max} was always 7. That is, all possible combinations of B_{bmrb}, B_{shifts} and B_{shiftx} voted for the same edge.

When an assignment is made, the associated nodes $k \in K$ and $r \in R$ are removed from B_{NOE}. Next, the procedure `ApplyNOE`(B_{NOE}) is applied. Finally, B_{bmrb}, B_{shifts}, and B_{shiftx} are synchronized with B_{NOE} so that all graphs have the same set of zero-weight edges. Each graph is then re-normalized so that the edge weights from each peak form a probability distribution. This completes a single iteration of the EM algorithm.

The E step takes $O(n^2)$ time. Computing the 7 maximum bipartite matchings takes $O(n^2)$ because each graph is sparse. The M step also takes $O(n^2)$ time to identify the maximum weight edges and $O(n^2)$ time to run `ApplyNOE`(B_{NOE}). Thus, each iteration of the EM algorithm takes $O(n^2)$. In the first stage, this algorithm is repeated until at least 5 assignments have been made. In our 20 trials, this never required more than 2 iterations. The algorithm then proceeds to phase 2.

Resonance assignments (phase 2)

The input to the second phase are the current bipartite graphs B_{NOE}, B_{bmrb}, B_{shifts} and B_{shiftx} and Θ, the master list of assignments. There are at least 5 assignments in Θ, thus we can determine the alignment tensors for the two RDC media using SVD (Losonczi et al., 1999). Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be the alignment tensors computed using the assignments in Θ for media 1 and 2, respectively. Each order matrix is used to back-compute a set of expected RDCs from the model using Eq. (1). Let D_m be the set of observed RDCs in medium m, and F_m be the set of back-computed RDCs using the model and $\mathbf{S}_m$. Two bipartite graphs M_1 and M_2 are constructed on the peaks in K and residues in R. The edge weights are computed as probabilities as follows:

$$w(k, r) = \mathbf{P}(k \mapsto r | S_m) = g(k, r), \tag{8}$$

where $k \in K$ and $r \in R$. Here,

$$g(k, r) = \mathcal{N}(d_m(k) - b_m(r), \sigma_m), \tag{9}$$

where $d_m(k) \in D_m$, $b_m(r) \in F_m$. Thus, the probabilities are computed using a 1 dimensional Gaussian distribution $\mathcal{N}$ (Equation 6) with mean $d_m(k) - b_m(r)$ and standard deviation σ_m. We used $\sigma = L/8$ Hz in all our trials, where L is the range of the RDCs in that medium (the maximum-valued RDC minus the minimum valued RDC). If an RDC is missing in medium i for a peak k, then we set the weight $w(k, r) = 1/n_0$ in graph M_i, for each residue r of the n_0 remaining (i.e., *unassigned*) residues. The bipartite graphs M_1 and M_2 are synchronized with B_{NOE} and then re-normalized so that the edge weights are probabilities.

The second phase uses the same EM algorithm as in phase 1, except that there are now two additional

bipartite graphs (M_1 and M_2) used to compute the expectation graph, V. Thus, there are $\sum_i^5 \binom{5}{i} = 31$ graph combinations used to construct V. At the end of each iteration, the alignment tensors $\mathbf{S}_1$ and $\mathbf{S}_2$ are updated (refined) using the master set of assignments, Θ, and M_1 and M_2 are recomputed using Equation 8. As in the first phase, each iteration of the EM algorithm takes $O(n^2)$. Thus making the remaining $O(n)$ assignments, takes $O(n^3)$ time. In all our trials, no more than 10 iterations were ever required to make all remaining assignments, since in practice, multiple assignments are made on each iteration.

A performance enhancement can be gained at the end of each iteration by taking M_1 and M_2 and computing a maximum bipartite matching on each. Let $H_1 \subset K \times R$ and $H_2 \subset K \times R$ be the maximum bipartite solutions for M_1 and M_2, respectively. Let $H = H_1 \cap H_2$. If $H \neq \varnothing$, then the assignments in H are made. This saves a constant factor in the runtime because only two bipartite matchings are needed. The use of this heuristic is justified because in general, RDCs are better predictors of assignment than are chemical shift statistics (see Figure 7).

Notes

1. In our experiments, we used the RDCs listed in the PDB restraints files; RDCs for residues 73-76 are available (Ottiger and Bax, 1998), but were omitted from the restraints file due to the flexibility of the C-terminus.
2. There is an inversion symmetry for each of the three eigenvectors. Therefore, there are eight isometries which leave the Saupe matrix unchanged. However, only four of those isometries are pure rotations ($SO(3)$). The other four are perversions in $O(3) - SO(3)$ (rotations composed with a reflection) and hence are not used to integrate over $SO(3)$.
3. The restraints file for PDB structure 3GB1 (Kuszewski et al., 1999) did not list any explicit hydrogen bonds. Therefore, the hydrogen bonds listed in PDB structure 1GB1 (Gronenborn et al., 1991) were used instead.

Acknowledgements

We thank Drs Hany Farid, J.C. Hoch, Ramgopal Mettu, Mr Anthony K. Yan, Ms Elisheva Werner-Reiss and all members of Donald Lab for helpful discussions and comments on drafts.

References

Al-Hashimi, H., Gorin, A., Majumdar, A., Gosser, Y. and Patel, D. (2002) *J. Mol. Biol.*, **318**, 637–649.
Al-Hashimi, H. and Patel, D. (2002) *J. Biomol. NMR*, **22**, 1–8.
Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) *J. Mol. Biol.*, **215**, 403–410.
Andrec, M., Du, P. and Levy, R. (2001) *J. Biomol. NMR*, **21**, 335–347.
Annila, A., Aitio, H., Thulin, E. and T., D. (1999) *J. Biomol. NMR*, **14**, 223–230.
Artymiuk, P.J., Blake, C.C.F., Rice, D.W. and Wilson, K.S. (1982) *Acta Crystallogr. B Biol. Crystallogr.*, **38**, 778.
Babu, C. R., Flynn, P.F. and Wand, A.J. (2001) *J. Am. Chem. Soc.*, **123**, 2691.
Bailey-Kellogg, C., Widge, A., Kelley III, J.J., Berardi, M., Bushweller, J. and Donald, B. (2000) *J. Comput. Biol.*, **7**, 537–558.
Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P. (2000) *Nucl. Acids Res.*, **28**, 235–242.
Blundell, T., Sibanda, B., Sternberg, M. and Thornton, J. (1987) *Nature*, **326**, 347–352.
Chen, Y., Reizer, J., Saier Jr., M.H., Fairbrother, W.J. and Wright, P.E. (1993) *Biochemistry*, **32**, 32–37.
Chou, J., Li, S. and Bax, A. (2000) *J. Biomol. NMR*, **18**, 217–227.
Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C. (2001) *Introduction to Algorithms*, 2nd edn., The MIT Press, Cambridge, MA, chapter 5, pp. 109–110.
Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.
Cross, A.D.J. and Hancock, E.R. (1998) *IEEE Trans. Pattern Anal. Machine Intell.*, **20**, 1236–1253.
Delaglio, F., Kontaxis, G. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 2142–2143.
Dempster, A., Laird, N. and Rubin, D. (1977) *J. Roy. Stat. Soc., Ser. B*, **39**, 1–38.
Diamond, R. (1974) *J Mol. Biol.*, **82**, 371–391.
Fejzo, J., Lepre, C., Peng, J., Bemis, G., Ajay, Murcko, M. and Moore, J. (1999) *Chem. Biol.*, **6**, 755–769.
Fetrow, J. and Bryant, S. (1993) *Bio/Technology*, **11**, 479–484.
Fiaux, J., Bertelsen, E.B., Horwich, A.L. and Wüthrich, K. (2002) *Nature*, **418**, 207–211.
Fowler, C., Tian, F., Al-Hashimi, H.M. and Prestegard, J.H. (2000) *J. Mol. Biol.*, **304**, 447–460.
Gallagher, T., Alexander, P., Bryan, P. and Gilliland, G.L. (1994) *Biochemistry*, **33**, 4721–4729.
Gemmecker, G., Jahnke, W. and Kessler, H. (1993) *J. Am. Chem. Soc.*, **115**, 11620–11621.
Girard, E., Chantalat, L., Vicat, J. and Kahn, R. (2001) *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 1–9.
Gnu (2002) The GNU General Public License, http://www.gnu.org/licenses/licenses.html
Greer, J. (1991) *Meth. Enzymol.*, **202**, 239–252.
Grishaev, A. and Llinas, M. (2002) *PNAS*, **99**, 6707–6712.
Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T. and Clore, G.M. (1991) *Science*, **253**, 657.
Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 627–638.
Harris, R. (2002) The Ubiquitin NMR Resource Page, BBSRC Bloomsbury Center for Structural Biology http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/index.html
Hoch, J., Burns, M. M. and Redfield, C. (1990) In *Frontiers of NMR in Molecular Biology*, Alan R. Liss, Inc., NY, pp. 167–175.
Hus, J., Marion, D. and Blackledge, M. (2000) *J. Mol. Biol.*, **298**, 927–936.

Hus, J., Prompers, J. and Bruschweiler, R. (2002) *J. Magn. Reson.*, **157**, 119–125.
Johnson, E., Lazar, G.A., Desjarlais, J.R. and Handel, T.M. (1999) *Struct. Fold Des.*, **7**, 967–976.
Johnson, M., Srinivasan, N., Sowdhamini, R. and Blundell, T. (1994) *Mol. Biochem.* **29**, 1–68.
Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graph.*, **14**, 51–55.
Kuhn, H. (1955) *Nav. Res. Logist. Quart.*, **2**, 83–97.
Kurinov, I.V. and Harrison, R.W. (1995) *Acta Crystallogr. D Biol. Crystallogr.*, **51**, 98–109.
Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1999) *J. Am. Chem. Soc.*, **121**, 2337–2338.
Langmead, C.J. and Donald, B.R. (2003) In *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB), Stanford University, Palo Alto, CA (August 11-14)*, pp. 209–217.
Langmead, C.J., Yan, A.K., Wang, L., Lilien, R.H. and Donald, B.R. (2003) In *Proceedings of the 7th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB) Berlin, Germany, April 10-13*, pp. 176–187.
Lathrop, R. and Smith, T. (1996) *J. Mol. Biol.*, **255**, 641–665.
Lim, K., Nadarajah, A., Forsythe, E.L. and Pusey, M.L. (1998) *Acta Crystallogr. D Biol. Crystallogr.*, **54**, 899–904.
Losonczi, J., Andrec, M., Fischer, W. and J.H., P. (1999) *J Magn. Reson.*, **138**, 334–342.
Meiler, J., Peti, W. and Griesinger, C. (2000) *J. Biomol. NMR*, **17**, 283–294.
Mueller, G., Choy, W., Yang, D., Forman-Kay, J., Venters, R. and Kay, L. (2000) *J. Mol. Biol.*, **300**, 197–212.
Neal, S., Nip, A.M., Zhang, H. and Wishart, D.S. (2003) *J. Biomol. NMR*, **26**, 215–240.
Oki, H., Matsuura, Y., Komatsu, H. and Chernov, A. A. (1999) *Acta Crystallogr. D Biol. Crystallogr.*, **55**, 114.
Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 12334–12341.
Palmer, A.G. (1997) *Curr. Opin. Struct. Biol.*, **7**, 732–737.
Ramage, R., Green, J., Muir, T.W., Ogunjobi, O.M., Love, S. and Shaw, K. (1994) *J. Biochem.*, **299**, 151–158.
Redfield, C., Hoch, J. and Dobson, C. (1983) *FEBS Lett.*, **159**, 132–136.
Rohl, C. and Baker, D. (2002) *J. Am. Chem. Soc.*, **124**, 2723–2729.
Rossman, M. and Blow, D. (1962) *Acta Crystallogr.*, **15**, 24–31.
Sali, A., Overington, J., Johnson, M. and Blundell, T. (1990) *Trends Biochem. Sci.*, **15**, 235–240.
Saupe, A. (1968) *Angew. Chem.*, **7**, 97–112.
Schneider, D., Dellwo, M. and Wand, A.J. (1992) *Biochemistry*, **31**, 3645–3652.
Schwalbe, H., Grimshaw, S.B., Spencer, A., Buck, M., Boyd, J., Dobson, C.M., Redfield, C. and Smith, L.J. (2001) *Protein Sci.*, **10**, 677–688.
Seavey, B., Farr, E., Westler, W. and Markley, J. (1991) *J. Biomol. NMR*, **1**, 217–236.
Shuker, S.B., Hajduk, P.J., Meadows, R.P. and Fesik, S.W. (1996) *Science*, **274**, 1531–1534.
Tian, F., Valafar, H. and Prestegard, J. H. (2001) *J. Am. Chem. Soc.*, **123**, 11791–11796.
Tjandra, N. and Bax, A. (1997) *Science*, **278**, 1111–1114.
Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 9279–9283.
Vaney, M.C., Maignan, S., Ries-Kautt, M. and Ducruix, A. (1996) *Acta Crystallogr. D Biol. Crystallogr.*, **52**, 505–517.
Vijay-Kumar, S., Bugg, C.E. and Cook, W.J. (1987) *J. Mol. Biol.*, **194**, 531–544.
Wang, L. and Donald, B.R. (2004) *J. Biomol. NMR*, in press.
Weber, P.L., Brown, S.C. and Mueller, L. (1987) *Biochemistry*, **26**, 7282–7290.
Wedemeyer, W.J., Rohl, C.A. and Scheraga, H.A. (2002) *J. Biomol. NMR*, **22**, 137–151.
Xu, X. and Case, D. (2001) *J. Biomol. NMR*, **21**, 321–333.
Xu, Y., Xu, D., Crawford, O.H., Einstein, J.R. and Serpersu, E. (2000) *Proc. RECOMB*, pp. 299–307.
Yan, A., Langwead, C. and Donald, B.R. (2003) *A Probability-Based Similarity Measure for Saupe Alignment Tensors with Applications to Residual Dipolar Couplings in NMR Structural Biology.* Technical Report No. TR2003-474, Dartmouth Computer Science Department, http://www.cs.dartmouth.edu/reports/abstracts/TR2003-474/
Zweckstetter, M. (2003) *J. Biomol. NMR*, **27**, 41–56.
Zweckstetter, M. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 3791–3792.
Zweckstetter, M. and Bax, A. (2001) *J. Am. Chem. Soc.*, **123**, 9490–9491.