

# High-Throughput 3D Structural Homology Detection via NMR Resonance Assignment

Christopher James Langmead\*

Bruce Randall Donald †, ‡, §, ¶, ||

## Abstract

One goal of the structural genomics initiative is the identification of new protein folds. Sequence-based structural homology prediction methods are an important means for prioritizing unknown proteins for structure determination. However, an important challenge remains: two highly dissimilar sequences can have similar folds — how can we detect this rapidly, in the context of structural genomics? High-throughput NMR experiments, coupled with novel algorithms for data analysis, can address this challenge. We report an automated procedure, called HD, for detecting 3D structural homologies from sparse, unassigned protein NMR data. Our method identifies 3D models in a protein structural database whose geometries best fit the unassigned experimental NMR data. HD does not use, and is thus not limited by sequence homology. The method can also be used to confirm or refute structural predictions made by other techniques such as protein threading or homology modelling. The algorithm runs in  $O(pn + pn^{5/2} \log(cn) + p \log p)$  time, where  $p$  is the number of proteins in the database,  $n$  is the number of residues in the target protein and  $c$  is the maximum edge weight in an integer-weighted bipartite graph. Our experiments on real NMR data from 3 different proteins against a database of 4,500 representative folds demonstrate that the method identifies closely related protein folds, including sub-domains of larger proteins, with as little as 10-30% sequence homology between the target protein (or sub-domain) and the computed model. In particular, we report no false-negatives or false-positives despite significant percentages of missing experimental data.

\*Carnegie Mellon Dept. of Computer Science, Pittsburgh, PA 15213 USA.

†Dartmouth Computer Science Department, Hanover, NH 03755, USA.

‡Dartmouth Chemistry Department, Hanover, NH 03755, USA.

§Dartmouth Biological Sciences Department, Hanover, NH 03755, USA.

¶Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA.

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

*Abbreviations used:* NMR, nuclear magnetic resonance; RDC, residual dipolar coupling; 3D, three-dimensional; HSQC, heteronuclear single-quantum coherence;  $H^N$ , amide proton; SAR, structure activity relation;  $SO(3)$ , special orthogonal (rotation) group in 3D.

## 1 Introduction

An important goal of the structural genomics initiative [29] is the identification of all protein folds in nature. While a great many folds have been observed, it is believed that still more exist and have yet to be determined experimentally. Unfortunately, determining a protein's structure experimentally, via X-ray crystallography or Nuclear Magnetic Resonance (NMR), is very costly and time-consuming. Consequently, it is necessary to prioritize different proteins for structure determination based on the likelihood they will lead to new folds. It is known that proteins having high sequence homology are very likely to have high structural homology. Therefore, it is reasonable to prioritize those proteins with low ( $< 30\%$ ) sequence identity to known structures. However, it is not uncommon for two dissimilar amino acid sequences to fold to the "same" tertiary structure. For example, the RMSD between the human ubiquitin structure (PDB Id 1D3Z) and the structure of the Ubx Domain from human Faf1 (PDB Id 1H8C) is quite small (1.9 Å), yet they have only 16% sequence identity (Fig. 1). Predicting structural homology given low sequence identity poses a difficult challenge for sequence-based homology predictors. Detecting structural homology is relatively easy, once the structure is determined experimentally. We ask: is there a set of very fast, cheap experiments that can be analyzed to rapidly detect 3D structural homology, without resorting to full-blown structure determination?

This paper presents a new method for structural homology detection that takes advantage of high-throughput solution-state NMR. Our algorithm, called HD, computes the likelihood  $P(M|D)$ , where  $D$  is a set of sparse, unassigned NMR data, and  $M$  is a model taken from a database

This work is supported by grants to B.R.D. from the National Institutes of Health (R01 GM-65982), and the National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, and EIA-9802068, EIA-0305444).

of protein structures. Backbone resonance assignments are needed to compute this likelihood and we apply the technique of Nuclear Vector Replacement (NVR) [22, 24, 25] to perform backbone amide resonance assignments given the model. The key idea behind our method is that structurally homologous proteins give rise to similar data, regardless of sequence identity. An important advantage of HD is that the required experimental NMR data can be recorded in about 2 days, far less than the weeks of data acquisition required for full-blown structure determination via NMR. In this way, the algorithm can detect homology early on in the discovery process. HD may also be used in conjunction with techniques such as protein threading [26, 42], and computational homology modelling [6, 13, 14, 18, 33], providing experimental validation of the computational predictions.

HD is demonstrated on NMR data from 3 proteins against a database of 4,500 representative folds determined either by X-ray crystallography or by NMR. We report no false positives or false negatives in detecting structural homologies between proteins with less than 30% sequence identity. We also report the successful detection of homology to a sub-domain of a larger protein.

## 1.1 Organization of paper

We begin, in Section 2, with a review of the specific NMR experiments used in our method, highlighting their information content. Section 3 describes existing techniques for homology detection. In section 4, we give the details of the HD algorithm and analyze its computational complexity. Section 5 reports the results of HD on real NMR data from three different proteins.

## 2 Background

Atomic nuclei having the quantum property of spin  $> 0$  resonate when subjected to radio-frequency energy in a strong magnetic field. The resonant frequency (or *chemical shift*) is determined by a number of factors including the atom type ( $^1\text{H}$ ,  $^{15}\text{N}$ ,  $^{13}\text{C}$ , etc.) and the local electronic environment surrounding the nucleus. An NMR spectrometer records these resonant frequencies as time-domain signals. These time-domain signals are almost always analyzed and interpreted in the frequency-domain, where resonances manifest as peaks in a spectrum. NMR data capture interactions between spin systems (tuples of atomic nuclei) in  $\mathbb{R}^2$ ,  $\mathbb{R}^3$ , or  $\mathbb{R}^4$ , where the axes are the chemical shifts of the constituent nuclei. For example, our algorithm processes the 2-dimensional  $^{15}\text{N}$ -edited Heteronuclear Single-Quantum Coherence (HSQC) spectrum, where each peak identifies an amide (bonded  $\text{H}^{\text{N}}$  and  $^{15}\text{N}$  atoms) pair. Proteins are linear-polymers of amino acids and the

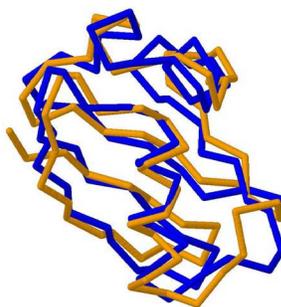
backbone of every amino acid (except proline), has a single amide group. Thus, in an ideal HSQC spectrum, each residue (amino acid) in the protein gives rise to a single, well-defined peak.\*

The process of mapping each peak to the spin-system that generated it is known as *assignment*. For the purposes of exposition, we will equate *spin-system* with *residue* as per the particular set of NMR data upon which our algorithm operates. Hence, we will (re)define assignment as the mapping of peaks to residues. The resonance assignment problem is, in fact, equivalent to the well-studied assignment problem from combinatorial optimization [19]. Let  $R$  be the set of residues in the primary sequence of the protein (except prolines and the  $N$ -terminus). Let  $K$  be the set of peaks in the HSQC. One can imagine constructing a bipartite graph on  $R$  and  $K$  as follows:  $B = \{K \cup R, E\}$ , where  $E = K \times R$ . Given some suitable means for computing a weight for each edge  $e \in E$ ,  $w : K \times R \rightarrow \mathbb{R}$ , one could imagine applying a standard algorithm for computing maximum bipartite matchings, such as the Kuhn-Munkres algorithm [19], for solving the resonance assignment problem. Indeed, maximum bipartite matching has been applied to resonance assignment by a number of algorithms (e.g., [17, 43, 22]).

Our method for detecting homology works as follows. Each model in a database of structures is used to compute a (different) set of assignments that correlate the experimental NMR data to the model. These assignments are made within a probabilistic framework, which we describe in Section 2.1. Consequently, we can compute the likelihood of the assignments, and therefore the model. The intuition is that a homologous structure should “fit” the data better than an unrelated structure.

The experimental inputs to HD are detailed in Table 1. HD calls the NVR algorithm as a subroutine. NVR computes assignments by correlating geometric constraints to a given model of the protein’s structure. These constraints are extracted directly from the NMR data. We will summarize these constraints here. All NMR data input to NVR is unassigned. An assay for measuring amide-exchange rates serves to identify resonant peaks associated with labile, solvent-accessible amide protons. NVR also uses  $\text{H}^{\text{N}}\text{-H}^{\text{N}}$  NOE’s ( $d_{\text{NNS}}$ ) and residual dipolar couplings (RDCs).  $d_{\text{NNS}}$  may be observed between pairs of amide protons that are within approximately 5 Å of each other.  $d_{\text{NNS}}$  are *local* measurements. In contrast, RDCs [37, 38] provide *global* orientational restraints on internuclear bond vectors. We note that RDCs are the *only* global measurement that can be made using NMR and that the techniques for recording

\*In practice however, peaks often overlap and some may not appear at all due to intra-molecular dynamics. These issues are just some of the challenges faced when analyzing NMR data. Prolines and the  $N$ -terminus do not, of course, generate amide peaks.



**Figure 1.** The backbone alignment of PDB Id 1D3Z (light grey/gold) and 1H8C (dark grey/blue). The RMSD between the proteins is only 1.9 Å, yet they have only 16% sequence identity.

RDCs from proteins in solution were first reported in 1995 [38]. For good introductions to RDCs see [34, 27, 37]. For each RDC  $d$ , we have

$$d = d_{\max} \mathbf{v}^T \mathbf{S} \mathbf{v}, \quad (1)$$

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  $3 \times 3$  *Saupe order matrix* [34].  $\mathbf{S}$  is a symmetric, traceless, rank 2 tensor with 5 degrees of freedom, which describes the average substructure alignment in the dilute liquid crystalline phase. We will refer to  $\mathbf{S}$  as the *alignment tensor*.  $\mathbf{S}$  is initially an unknown; various methods for estimating  $\mathbf{S}$  from unassigned data exist (e.g., [25, 45, 44]). Given assignments of five or more RDCs in substructures of known geometry,  $\mathbf{S}$  can be determined using singular value decomposition [27].

Once  $\mathbf{S}$  has been determined, RDCs may be simulated (back-calculated) given any other internuclear vector  $\mathbf{v}_i$ . In particular, suppose an ( $^1\text{H}^{\text{N}}, ^{15}\text{N}$ ) peak  $i$  in an 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). In this way, back-computed RDCs can be used to generate constraints on assignment.

## 2.1 Nuclear Vector Replacement

In this section, we will briefly summarize the NVR algorithm [22]. NVR is divided into two phases, *Tensor Determination* and *Resonance Assignment*. In the first phase, chemical shift predictions,  $d_{\text{NNS}}$ , and amide exchange rates are correlated with a given structural model of the protein to make a small number of assignments using Expectation/Maximization (EM). Specifically, this phase attempts to assign at least 5 peaks for the purpose of determining the alignment tensors directly. The tensors are then used to con-

vert 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).

### 2.1.1 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. EM has been described previously [11]. 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 [9]. In the EM framework there are both observed and hidden (i.e., unobserved) random variables. In the context of NVR, the observed variables are the chemical shifts,  $d_{\text{NNS}}$ , 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 tensor. Of course, the values of the hidden variables are unknown. Specifically,  $Y_G$  is the set of edge weights of a bipartite graph. The weights  $Y_G$  represent *correct* assignments, and 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  $\Theta$  of all assignments made so far by the algorithm. Initially,  $\Theta$  is empty. As EM makes more assignments,  $\Theta$  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  $E(\Theta \cup \Theta' | \Theta) = E(\log \mathbf{P}(X, Y | \Theta \cup \Theta'))$ . Here,  $\Theta'$  is a non-

Experiment/Data	Information Content	Role
$H^N-^{15}N$ HSQC	$H^N, ^{15}N$ Chemical shifts	Backbone resonances, Cross-referencing NOESY
$H^N-^{15}N$ RDC (in 2 media)	Restrains 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
$^{15}N$ TOCSY	Side-Chain Chemical Shifts	Tensor Determination, Resonance Assignment
Backbone Structure	Tertiary Structure	Tensor Determination, Resonance Assignment
Chemical Shift Predictions	Restrains on Assignment	Tensor Determination, Resonance Assignment

**Table 1. Experiment Suite:** The 6 *unassigned* NMR spectra used by our algorithm to perform homology detection. The HSQC provides the backbone resonances to be assigned.  $H^N-^{15}N$  RDC data in two media provide independent, global restraints on the orientation of each backbone amide bond vector. The H-D exchange HSQC identifies fast exchanging amide protons. These amide protons are likely to be solvent-exposed and non-hydrogen bonded and can be correlated to the structural model. A sparse number ( $< 1$  per residue, on average) of unassigned  $d_{NNS}$  can be obtained from the NOESY. These  $d_{NNS}$  provide distance constraints between spin systems which can be correlated to the structural model. The  $^{15}N$  TOCSY is used to measure  $^1H$  side-chain resonances. These resonances are useful in eliminating certain amino acid types from consideration when performing assignments. Chemical shift predictions are used as a probabilistic constraint on assignment.

empty set of candidate new assignments that is disjoint from  $\Theta$ . The  $M$  step computes the maximum likelihood new assignments  $\Theta^*$ ,  $\Theta^* = \underset{\Theta'}{\operatorname{argmax}} E(\Theta \cup \Theta' | \Theta)$ . The master list of assignments is then updated,  $\Theta \leftarrow \Theta \cup \Theta^*$ . Thus, on each iteration, the EM algorithm makes the most likely assignments. The algorithm terminates when each peak has been assigned. NVR runs in time  $O(n^{5/2} \log(cn))$ , where  $n$  is the number of amino acids in the protein and  $c$  is the maximum edge weight in an integer-weighted bipartite graph [23]. For reference,  $c$  is a constant and dictated by the resolution of the NMR data. The NVR algorithm is described in detail in [22].

### 3 Related Work

Xu and co-workers [42] have also attacked homology detection using sparse NMR data. Their method extends protein threading by incorporating a sparse set of NOE data. Their method requires assigned NOEs while our method works on unassigned NMR data.

Assigned RDCs have been used for homology detection [1, 3]. In contrast, our algorithm processes unassigned NMR data. This is a significant distinction. Assigning NMR data typically requires many days of data acquisition followed by several days of data analysis. Moreover, assigned RDCs can be used to compute the alignment tensor,  $S$ , directly. Our method calls NVR as a subroutine, and

thus simultaneously performs tensor estimation, resonance assignment, and homology detection.

Unassigned NOEs [12] and unassigned RDCs [21, 39] have also been used for homology detection. [21] and [39] both estimate the alignment tensor  $S$  by first estimating the eigenvalues of  $S$  and subsequently, its eigenvectors via a discrete rotation search over  $SO(3)$ . In both cases,  $S$  can be estimated in time  $O(pnk^3 + p \log p)$  where  $p$  is the number of proteins in the database,  $n$  is the number of amino acids in the protein, and  $k$  is the resolution of a grid over  $SO(3)$ . The two techniques measure structural homology by computing the similarity of the distribution of (unassigned) experimentally recorded RDCs to an expected set of RDCs. The technique presented in [39] takes advantage of the fact that the backbone amide bond vectors in an  $\alpha$ -helix are roughly parallel, and therefore generate (approximately) the same RDC value. Their method counts the number of  $\alpha$ -helices in the protein, and then estimates their relative size and orientation. The number, sizes and orientation of these helices are compared to putative homologs. Their method does not attempt to estimate  $\beta$ -structure and therefore does not generalize to all protein structures. Indeed, their method was only demonstrated on one protein (F1Fo ATP Synthase) which is 84%  $\alpha$ -helical and contains no  $\beta$  structure. In contrast, the method we presented in [21], called GD, imposes no bias on the secondary structure characteristics of the protein and was demonstrated on a variety of different proteins with different secondary structure

characteristics. GD reported no false negatives in detecting homology against a database of 2,500 structures. Unfortunately, GD did report a number of false-positives at a rate of about 2%. These false-positives are the main motivation for the algorithm presented in this paper (HD). HD is very different than GD. HD processes more NMR data than GD and requires no rotation search over  $SO(3)$ . The computational complexity of HD is  $O(pn + pn^{5/2} \log(cn) + p \log p)$  vs.  $O(pnk^3 + p \log p)$  for GD. Thus, HD is faster (when  $k \geq O(n^{2/3})$ ) and has no dependency on  $k$ . In practice it takes only minutes to process each protein. Conservative heuristics are also employed to eliminate much of database from consideration early such that the entire database can be processed in under an hour. More important, HD generated no false positives or false negatives in our experiments.

The HD algorithm is an example of ‘structural homology detection by NMR resonance assignments’, which was suggested in [21]. Recently [28], Meiler and Baker have introduced a technique for homology detection and fold determination from unassigned  $^{15}\text{N}$  and  $^{13}\text{C}$  labeled data. HD requires only  $^{15}\text{N}$  labeling. This confers a significant advantage in terms of cost, as  $^{13}\text{C}$  labeling is an order of magnitude more expensive than  $^{15}\text{N}$  labeling. Moreover, Meiler and Baker’s method are based on Neural Networks, fuzzy-logic, and Monte Carlo sampling. HD, in contrast, is built upon combinatorially precise algorithms. However, the pre-processing before HD currently uses bioinformatics tools such as HNN [15], which is not combinatorially precise.

## 4 Details of the HD algorithm

### 4.1 Preprocessing

We first assembled a database of 4,496 structural models from the Protein Data Bank (PDB [5]) representing a variety of different fold-families. Let  $t$  be the target protein. That is,  $t$  is the protein whose structure we are trying to determine via HD. Let  $M$  be the set of protein models in the database. Let  $m \in M$  be a model in the database. Let  $s(m)$  be the primary sequence of  $m \in M$ , and let  $s(t)$  be the primary sequence of  $t$ .

Using the program HNN [15], we estimate the secondary structure of the target protein using its primary sequence,  $s(t)$ . HNN was chosen specifically because it performs secondary structure predictions *without* performing a sequence alignment to known structures. HNN makes predictions using a neural network. We note that none of our test proteins were present in the training set used to train HNN. Therefore, our algorithm does not gain an unfair advantage based on sequence similarity to known structures.

The database is then filtered using the secondary structure prediction and the length of  $t$ . Briefly, structures are discarded that have very different secondary structure com-

position or are significantly longer or smaller than  $t$ . The interested reader is directed to Appendix A.1 for details of the filtering criteria. Let  $W \subseteq M$  be the set of proteins that satisfy the constraints of the filters.

Next, for each model  $m \in W$ , we use the homology modelling program MODELLER [32] to perform both sequence alignment between  $s(t)$  and  $s(m)$ , and subsequently build a backbone model for  $t$  based on the backbone structure of  $m \in W$ .

### 4.2 Using NVR

Let  $T'$  be the set of models constructed by MODELLER from  $m \in W$ . The function  $\text{NVR}(t', D)$  takes as input a model  $t' \in T'$  and a sparse set of NMR data,  $D$ , and returns an assignment  $A$ . We run NVR for each model in  $T'$ .

We next compute the likelihood of an assignment,  $A$ , given the data. To do this, we must describe in more detail the inner mechanisms of NVR. The algorithm in this paper uses a slightly modified form of the NVR algorithm in [22]. We will indicate the changes we made in the following paragraphs. In this paper, we equate the likelihood of the assignment with the likelihood of the model. That is,  $P(A|D) = P(m|D)$ .

As discussed in Section 2.1, NVR uses a probabilistic framework to assign the peaks in the HSQC spectrum. In particular, NVR constructs seven weighted bipartite graphs encoding seven different probability distributions on assignment. Let  $R$  be the set of residues in the model (as constructed by the program MODELLER). Let  $K$  be the set of peaks in the HSQC. Each bipartite graph is defined as follows:  $B = \{K \cup R, E\}$ , 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  $|K| = |R|$ .

The first bipartite graph is constructed using amide exchange rates experimentally measured by NMR. Amide exchange rates are indicative of solvent accessibility and hydrogen bonding. The program RASMOL is used on the input model to identify the residues with hydrogen-bonded backbone amides. A uniform probability is given over any slow exchanging peak and these residues. Edges from non-hydrogen bonded surface residues to slow-exchanging peaks are given a default probability of  $\epsilon$ . This is the first difference between the algorithm in this paper and the NVR algorithm as presented in [22, p. 131]. In the unmodified NVR algorithm, we immediately set  $\epsilon = 0$ , effectively disallowing such an assignment. This is appropriate when the input model has very high ( $> 90\%$ ) sequence homology to the target protein. In that case, we expect the hydrogen bonding patterns to be the same. In a homologous

protein, however, the structures are somewhat different and may have different hydrogen bonding patterns. This is especially true of the models we constructed using the MODELLER program. In this case, it is appropriate to set the assignment probability to a low, but non-zero value. Let  $B_{HD}$  be the bipartite graph constructed using the amide exchange data. Constructing  $B_{HD}$  takes  $O(n^2)$  time.

Next, the model is used to predict the chemical shifts of the backbone amide protons and nitrogens. 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 used to compute probabilities which, in turn, become edge weights on a bipartite graph:  $w(k, r) = \mathbf{P}(k \mapsto r) = f(k, r)$ , where  $k \in K$  and  $r \in R$ . Here,  $f(k, r) = \mathcal{N}(\omega_H(k) - \mu_H(r), \sigma_H(r)) \mathcal{N}(\omega_N(k) - \mu_N(r), \sigma_N(r))$ . The function  $\mathcal{N}(x - \mu, \sigma)$  is the probability of observing the difference  $x - \mu$  in a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . That is,

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

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.

Three different methods for backbone chemical shift prediction are used resulting in three different bipartite graphs. The first method applies statistics from the BioMagRes-Bank (BMRB) [36]. Let  $B_{bmrB}$  be the bipartite graph whose edges are computed using the statistics from the BMRB. The programs SHIFTS [41] and SHIFTX [30] are also used to compute chemical shift predictions using the model. 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  $B_{bmrB}$ ,  $B_{shifts}$  and  $B_{shiftx}$  takes  $O(n^2)$  time. A detailed explanation of the construction of these graphs, including the estimating of the various  $\mu$ 's and  $\sigma$ 's, is presented in [22].

The final modification we have made to the NVR algorithm is the incorporation of data from the  $^{15}\text{N}$  TOCSY [7]. The  $^{15}\text{N}$  TOCSY measures the chemical shifts of side-chain protons. These side-chain chemical shifts are mapped to the peaks in the HSQC. Differences between the native structure and the model built using MODELLER will perturb the chemical shift predictions. The  $^{15}\text{N}$  TOCSY partially compensates for these differences. The number of observed

side-chain peaks is indicative of amino acid type. For example, modulo noise, if a given residue gives rise to more than two side-chain peaks, it cannot possibly be a Glycine, because Glycine has only two  $\text{H}_\alpha$  protons. The actual frequencies of the side-chain resonances are also indicative of the amino acid type. Note, the  $^{15}\text{N}$  TOCSY does not uniquely identify the amino acid type associated with each peak. For example, it is often the case that some of the side-chain resonances are missing from the data. Thus, the  $^{15}\text{N}$  TOCSY is more useful for ruling out certain amino acid types.

The fifth bipartite graph,  $B_{tocsy}$ , is constructed as follows. If the number of side chain resonances for  $k \in K$  exceeds the number of side-chain protons for a given amino acid type  $t$ , the edge-weights between  $k$  and any instance of amino acid type  $t$  in the primary sequence of the protein is set to 0. Otherwise, the edge-weight is computed as the joint probability of the individual side chain resonances. The marginal probabilities are computed using Gaussian distributions. More sophisticated means for modeling assignment likelihoods using TOCSY resonances have been employed in the JIGSAW algorithm [4]. It would be interesting to include these techniques into NVR.

The final two bipartite graphs are constructed using the experimentally determined RDCs. As previously mentioned, the NVR algorithm has two phases. In the first phase, a small number of assignments are made using the probability distributions encoded in  $B_{HD}$ ,  $B_{bmrB}$ ,  $B_{shifts}$ ,  $B_{shiftx}$  and  $B_{tocsy}$ . Once 5 assignments are made, we can determine the alignment tensors for the two RDC media using SVD. Let  $\mathbf{S}_1$  and  $\mathbf{S}_2$  be the alignment tensors computed using the assignments in  $\Theta$  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  $B_{S_1}$  and  $B_{S_2}$  are constructed on the peaks in  $K$  and residues in  $R$ . The edge weights are computed as probabilities as follows:  $w(k, r) = \mathbf{P}(k \mapsto r | S_m) = g(k, r)$  where  $k \in K$  and  $r \in R$ . Here,  $g(k, r) = \mathcal{N}(d_m(k) - b_m(r), \sigma_m)$ , where  $d_m(k) \in D_m$ ,  $b_m(r) \in F_m$ . Thus, the probabilities are computed using a 1 dimensional Gaussian distribution  $\mathcal{N}$  (Eq. (2)) with mean  $d_m(k) - b_m(r)$  and standard deviation  $\sigma_m$ . 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 bipartite graph  $B$ , for each residue  $r$  of the  $n_0$  remaining (i.e., *unassigned*) residues. Constructing  $B_{S_1}$  and  $B_{S_2}$  takes  $O(n^2)$  time.

In summary, there are seven bipartite graphs encoding seven different probability distributions on assignment. Constructing these graphs takes  $O(n^2)$  time. NVR uses these bipartite graphs and iteratively assigns each peak in

Protein	Homolog	Sequence Identity	RMSD	HD-score
Ubiquitin	1H8C:A	26.8%	1.8Å	-7.65
	1RFA	15.9%	2.2Å	-8.69
	1VCB:B	11.8%	1Å	-8.31
	1EF1:A [4-84]	10%	1.6Å	-8.50
GαIP	1DK8:A	28.7%	1.8Å	-7.18
SPG	1JML:A	12.8%	1.8Å	-9.27
	1HEZ:E	12.7%	2.0Å	-9.65

**Table 2. Homologous Structures with low sequence similarity detected by HD.** Column one lists the three test proteins. Column 2 lists the PDB Id and chain Id for the structures detected by HD. Column 3 lists the sequence identity of the proteins in column 1 and the primary sequence of the structure in column 2. Column 4 lists the backbone RMSD between the structures in column 2 and the native structures of the the proteins in column 1. The native structures for Ubiquitin, GαIP and SPG were taken as 1D3Z, 1CMZ and 3GB1, respectively. Column 5 lists the score computed by our algorithm HD. Higher-HD scores (closer to 0) indicate closer structural similarity. Note that this table does not include those structures detected by HD which have more than 30% sequence identity (see text).

the HSQC. Following assignment, we can then go back and compute the likelihood of the assignments. Let  $\mathcal{B} = \{B_{HD}, B_{bmr}, B_{shifts}, B_{shiftx}, B_{tocsy}, B_{S_1}, B_{S_2}\}$ . The edge and vertices are identical for every element of  $B_i$ , only the edge-weights differ. Combine-Graphs is a function that takes as input a set of bipartite graphs and returns a new bipartite graph [22]. The edge weights of the output graph are the joint probabilities of the edges in the input graphs:

$$w(k, r) = \prod_{i \in \mathcal{B}} w_i(k, r). \quad (3)$$

Let  $H = \text{Combine-Graphs}(\mathcal{B})$ . Given an assignment  $A \subset K \times R$ , as computed by NVR, we can compute the expected log-likelihood of that assignment using the bipartite graph  $H$ :

$$\mathcal{L}(A|H) = \frac{1}{|A|} \sum_{i \in A} \log w_i, \quad (4)$$

We have found that the expected log-likelihood is more robust than the total log-likelihood to small differences between the protein's native structure and the homologous structure.

Recall that  $W$  is the set of putative homologous structures from our database. We compute an assignment  $A$  and the bipartite graph  $H$  for each  $m \in W$  using NVR. We then rank each  $m \in W$  by  $\mathcal{L}(A|H)$ .

We now analyze the computational complexity of HD. Let  $p$  be the number of proteins in the database  $M$ . Applying the various filters takes  $O(n)$  time for each  $m \in M$ . Now, let  $q = |W|$ , that is, the number of proteins that pass the filter. In our experiments,  $q$  is typically very small ( $< 10$ ). The HD algorithm calls NVR as a subroutine on each  $m \in W$ . The NVR algorithm takes  $O(n^{5/2} \log(cn))$

time [23], where  $n$  is the number of residues in the protein and  $c$  is the maximum edge-weight in an integer-weighted bipartite graph. We note that in NVR,  $c$  is determined by the resolution of NMR data. NMR data is, in general, accurate to no more than 5-6 significant digits. Consequently, setting  $c = 10^7$  suffices. Thus,  $q$  proteins can be processed in  $O(qn^{5/2} \log(cn))$  time. Sorting the models by  $\mathcal{L}(A|H)$  takes time  $O(q \log q)$ , for a total runtime of  $O(pn + qn^{5/2} \log(cn) + q \log q)$ . In our experiments, the entire database of 4,500 structures was filtered, assigned, and the HD-score computed in about an hour. To exhaustively rank *all* of the proteins in  $M$  (i.e., without applying the filters), takes  $O(pn + pn^{5/2} \log(cn) + p \log p)$  time. Clearly, it is trivial to parallelize HD.

## 5 Results and Discussion

Our goal was to identify structural homology between proteins with less than 30% sequence identity. While there are over 18,000 protein structures deposited in the PDB to date, only a small handful of these proteins have RDC data (as required by HD) published in the BMRB. This is due, in part, to the fact that the recording of RDCs in solution has only recently been perfected. In contrast, NOE data is available for thousands of proteins. Unfortunately, simulating RDC data is difficult for two reasons. First, one needs to predict the alignment tensor for a given medium. This devolves to simulating the tumbling dynamics for the interaction of the protein with the aligning medium *in solution*. This is, in general, difficult to do. Furthermore, it is difficult to create an accurate noise model because the noise in real experimental RDC data is governed in part by such factors as the internal dynamics of the protein. We felt that

we could not reasonably simulate realistic RDC data. Thus, the number of proteins we tested was limited by the contents of the BMRB. Only 5 proteins have the necessary published data. Of these, only 3 have structural homologs that have less than 30% sequence identity; the 76-residue human ubiquitin (PDB Id 1D3Z [8]), the 56-residue streptococcal protein G (SPG) (PDB Id 3GB1 [20]), and the 128-residue  $G\alpha$  Interacting Protein ( $G\alpha$ IP) (PDB Id 1CMZ [10]). We will refer to these proteins as our *test proteins*. Experimental data for 3 different proteins is considered to be a more than adequate test suite by the NMR community [40], and many new computational protocols are tested on only one protein (e.g., [17, 39]).

Table 2 lists the homologous protein structures with low sequence similarity detected using the HD algorithm. Each of the models in Table 2 has an RMSD less than  $2.3\text{\AA}$  to the native structure of the test protein. Thus, we report no false positives from our experiments on three different proteins against a database of 4,496 protein structures. Moreover, no significant similarity was detected between the primary sequences of 2 of our test proteins (Ubiquitin and SPG) and their respective homologs using NCBI's pair-wise BLAST analysis using a threshold of 10. For the third protein ( $G\alpha$ IP), a modest similarity was observed ( $E = 5 \times 10^{-8}$ ). The only other structures identified by HD were indeed structurally homologous, but had  $> 30\%$  sequence homology to our test proteins. For example, the protein SPG was (correctly) identified as being homologous to the structure 2IGD. 2IGD has an RMSD of  $0.6\text{\AA}$  to 3GB1 (SPG's native structure) but the two proteins have 87.5% sequence identity. Hence, predicting homology between SPG and 2IGD would have been easy using sequence-alignment techniques. However, it is completely correct for HD to detect the SPG-2IGD homology, so this is not a false-positive, but rather an easy case. For comparison, HD computes a score of -5.03 for 2IGD, which is higher (i.e., fits the data better) than the scores reported in Table 2 for 1JML (-9.27) and 1HEZ (-9.65). We also report no false negatives for structures with backbone RMSDs less than  $2.3\text{\AA}$ .

We next set out to determine the relationship between the score computed by HD, and backbone RMSD. We searched our database and identified a subset of structures having between  $2.4$  and  $11\text{\AA}$  backbone RMSD and less than 30% sequence identity to our three test proteins. We will refer to these structures as our *comparison set*. See Table 3 for details of the comparison set. We ran NVR and computed the HD-score for the structures in the comparison set. The HD-scores computed for the comparison set are lower than the scores listed in Table 2, reflecting the fact that the structures in the comparison set are less similar to the test proteins. Note that the scores in column 5 of Table 2 are all greater than -10.0. None of the structures in the comparison set had a score higher than -15.2. The mean HD-score for the struc-

Protein	Homolog	Sequence Identity	RMSD	HD-score
Ubiquitin	1C9F:A	12.1%	$3\text{\AA}$	-17.52
	1XGM:A	6.1%	$4.6\text{\AA}$	-18.52
	1ESR:A	0%	$5.9\text{\AA}$	-15.21
$G\alpha$ IP	1VLK	8.2%	$3.6\text{\AA}$	-26.65
	1I4Y:D	4.8%	$4\text{\AA}$	-28.06
	1SWG:A	0%	$4.5\text{\AA}$	-26.38
	1B33:D	5.8%	$4.9\text{\AA}$	-22.08
	1CFC	5.1%	$6.1\text{\AA}$	-28.48
	1J95:A	0%	$7.7\text{\AA}$	-23.37
	1IDR:B	5.2%	$8\text{\AA}$	-25.61
	1E8E:A	7.1%	$9.4\text{\AA}$	-35.18
	1IDR:A	5.2%	$10.9\text{\AA}$	-33.90
SPG	1EX4:B	8.7%	$2.5\text{\AA}$	-19.04
	1EXQ:B	6.2%	$3\text{\AA}$	-31.04
	1MPG:A	7.3%	$4.4\text{\AA}$	-26.31
	1DH3:A	0%	$5\text{\AA}$	-17.27

**Table 3. Comparison Set** The structures in this table comprise our comparison set. These structures correspond to the blue filled-in diamonds in Figure 2.

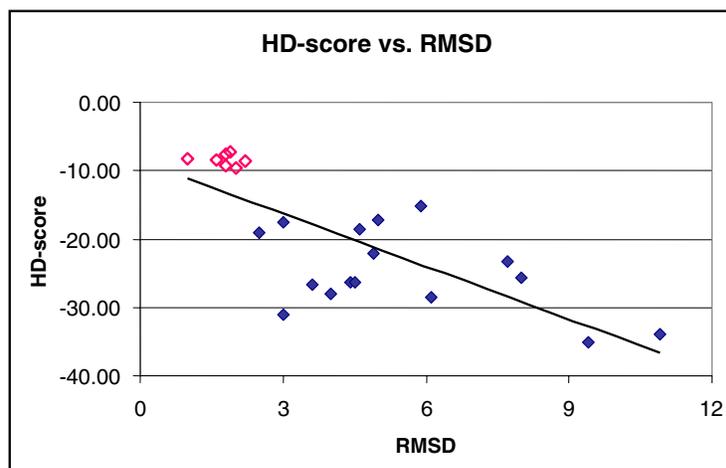
tures listed in Table 2 is -8.5, while the mean HD-score for the comparison set is -24.7. A *t*-test reports a significance of  $6.8 \times 10^{-7}$  between the means of the two distributions of HD-scores. Figure 2 shows the relationship between the HD-score and the RMSD. The score computed by HD is correlated with RMSD (correlation coefficient = -0.75).

## 5.1 Sub-domain Detection

As indicated in Table 2, one of the homologs of 1D3Z (ubiquitin) is a subdomain of a larger protein, 1EF1 (Moesin). To obtain these results, we used the program PDP [2] to predict the domain boundaries in 1EF1. PDP predicted 3 sub-domains (residues 4-84, 85-202, and 203-297). Residues 4-84 are homologous to ubiquitin and it was this sub-domain that was detected by our algorithm. In principle, a program like PDP could be used to process every entry in an initial database of structures. Any extracted sub-domains could be added to the database.

## 5.2 Missing Data

The Expectation/Maximization method, upon which NVR is based, is known to be robust to missing and corrupted data [22]. Our algorithm for homology detection inherits this same property. Table 4 summarizes the data processed in our 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,



**Figure 2.** HD-score vs. RMSD : Scatter plot of the score computed by our algorithm vs. backbone RMSD. The red open diamonds are the structures in Table 2. The blue solid diamonds are the structures in our comparison set (see text). The line is a least-squares fit to the data. The correlation coefficient is -0.75.

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%)
GαIP	122	6, (5%)	70	66	58 (45%)	62, (48%)

**Table 4. Missing Data.** The data processed on our experiments contained both missing peaks and missing RDCs. By missing, we mean that if the protein has  $n$  amino acids (excluding prolines and the  $N$ -terminus), then the ideal HSQC spectrum should have  $n$  peaks. Ideally,  $n$  RDCs should also be recorded for each medium. In reality, some data is not obtainable. Column 2 indicated the number of HSQC peaks contained in our experimental data. Column 3 indicates the number of missing HSQC peaks (number of expected peaks – number of observed peaks). Columns 4-5 indicate the number of RDCs obtained in media 1 and 2. Columns 6-7 indicate the number of missing RDCs in media 1 and 2. The modified NVR algorithm in HD processed all data as-is, and handles missing data.

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. 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 ( $P = 1/72$ ) 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. As shown in Table 4, our algorithm performed well on data sets that contained up to 5% missing HSQC peaks and up to 48% missing RDCs.

## 6 Conclusion

We have described a fast, automated procedure for structural homology detection from sparse unassigned NMR data. The relationship between structure and function is strong, thus our algorithm can be used to help characterize

the function of new proteins. Perhaps more important, homology can be detected very early based on a sparse, fast, and inexpensive set of NMR experiments, without resorting to full-blown structure determination. HD identifies the 3D structural models in a protein structural database whose geometries best fit the unassigned experimental NMR data. The algorithm runs in  $O(pn + pn^{5/2} \log(cn) + p \log p)$  time, where  $p$  is the number of proteins in the database,  $n$  is the number of residues in the target protein, and  $c$  is the maximum edge-weight in an integer-weighted bipartite graph. The NMR data required by our algorithm can be recorded in about 2 days, far less than the time required for full-blown structure determination via NMR.

Our method has been tested on NMR data from 3 test proteins against a protein structure database containing almost 4,500 models. No false negatives or positives were observed, despite i) sequence identities of less than 30% between the target and homolog and ii) significant amounts of noise and missing data. Our method was also able to correctly identify structural homology between ubiquitin and a sub-domain of the protein moesin. Thus, the method is both robust and accurate, suggesting the possibility that it may be useful in structural genomics.

## References

- [1] MEILER, J. AND BLOMBERG, N. AND NILGES, M. AND GRIESINGER, C. A new approach for applying residual dipolar couplings as restraints in structure elucidation. *Journal of Biomolecular NMR* 16 (2000), 245–252.
- [2] ALEXANDROV, N., AND SHINDYALOV, I. PDP: protein domain parser. *Bioinformatics* 19, 3 (2003), 429–430.
- [3] ANNILA, A. AND AITIO, H. AND THULIN, E. AND DRAKENBERG, T. Recognition of protein folds via dipolar couplings. *J. Biom. NMR* 14 (1999), 223–230.
- [4] BAILEY-KELLOGG, C. AND WIDGE, A. AND KELLEY III, J.J. AND BERARDI, M.J. AND BUSHWELLER, J.H. AND DONALD, B.R. The NOESY Jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J Comput Biol* 7, 3-4 (2000), 537–58.
- [5] BERMAN, H.M. AND WESTBROOK, J. AND FENG, Z. AND GILLILAND, G. AND BHAT, T.N. AND WEISSIG, H. AND SHINDYALOV, I.N. AND BOURNE, P.E. The Protein Data Bank. *Nucl. Acids Res.* 28 (2000), 235–242.
- [6] BLUNDELL, T.L. AND SIBANDA, B.L. AND STERNBERG, M.J. AND THORNTON, J.M. Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules. *Nature* 326 (1987), 347–352.
- [7] CAVANAGH, J. AND FAIRBROTHER, W. J. AND PALMER, A. G. AND SKELTON, N. J. . *Protein NMR Spectroscopy: Principles and Practice*. Academic Press, San Diego, CA, 1995, ch. 6, pp. 455–456.
- [8] CORNILESCU, G. AND MARQUARDT, J. L. AND OTTIGER, M. AND BAX, A. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J.Am.Chem.Soc.* 120 (1998), 6836–6837.
- [9] CROSS, A. D. J. AND HANCOCK, E. R. Graph Matching With a Dual-Step EM Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1236–1253.
- [10] DE ALBA, E. AND DE VRIES, L. AND FARQUHAR, M. G. AND TJANDRA, N. Solution Structure of Gaip (Galpha Interacting Protein): A Regulator of G Protein Signaling. *J.Mol.Biol.* 291 (1999), 927.
- [11] DEMPSTER, A. AND LAIRD, N. AND RUBIN, D. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1 (1977), 1–38.
- [12] ERDMANN, M.A. AND RULE, G. S. Rapid Protein Structure Detection and Assignment using Residual Dipolar Couplings. Tech. Rep. CMU-CS-02-195, Carnegie Mellon University, Computer Science Department, School of Computer Science, 2002.
- [13] FETROW, J.S. AND BRYANT, S.H. New Programs for Protein Tertiary Structure Prediction. *Bio/Technology* 11 (1993), 479–484.
- [14] GREER, J. Comparative Modeling of Homologous Proteins. . *Meth. Enzymol.* 202 (1991), 239–252.
- [15] GUERMEUR, Y. *Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines*. PhD thesis, Univ. Paris, 1997.
- [16] HOLM, L. AND SANDER, C. Database algorithm for generating protein backbone and side-chain coordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.* 218, 1 (1991), 183–194.
- [17] HUS, J.C. AND PROPMERS, J. AND BRÜSCHWEILER, R. Assignment strategy for proteins of known structure. *J. Mag. Res* 157 (2002), 119–125.
- [18] JOHNSON, M.S. AND SRINIVASAN, N. AND SOWDHAMINI, R. AND BLUNDELL, T.L. Knowledge-Based Protein Modeling. *Mol. Biochem.* 29 (1994), 1–68.
- [19] KUHN, H.W. Hungarian method for the assignment problem. *Nav. Res. Logist. Quarterly* 2 (1955), 83–97.
- [20] KUSZEWSKI, J. AND GRONENBORN, A. M. AND CLORE, G. M. Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. *J. Am. Chem. Soc.* 121 (1999), 2337–

2338.

- [21] LANGMEAD, C. J., AND DONALD, B. R. 3D-Structural Homology Detection via Unassigned Residual Dipolar Couplings. *Proc. IEEE Computer Society Bioinformatics Conference (CSB), Stanford University, Palo Alto, CA (August 11-14) (2003)*, 209–217.
- [22] LANGMEAD, C. J., AND DONALD, B. R. An Expectation/Maximization Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *J. Biomol. NMR*. 29, 2 (2004), 111–138.
- [23] LANGMEAD, C. J., AND DONALD, B. R. An improved nuclear vector replacement algorithm for nuclear magnetic resonance assignment. Tech. Rep. TR2004-494, Dartmouth Dept. of Computer Science, 2004.
- [24] LANGMEAD, C. J., YAN, A. K., WANG, L., LILIEN, R. H., AND DONALD, B. R. A Polynomial-Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *J. Comp. Bio.* (2003). In press.
- [25] LANGMEAD, C. J., YAN, A. K., WANG, L., LILIEN, R. H., AND DONALD, B. R. A Polynomial-Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *Proc. of the 7th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB) Berlin, Germany, April 10-13 (2003)*, 176–187.
- [26] LATHROP, R.H. AND SMITH, T.F. Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions. *J. Mol. Biol.* 255 (1996), 641–665.
- [27] LOSONCZI, J.A. AND ANDREC, M. AND FISCHER, W.F. AND PRESTEGARD J.H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson* 138, 2 (1999), 334–42.
- [28] MEILER, J. AND BAKER, D. Rapid protein fold determination using unassigned NMR data. *Proc. Natl. Acad. Sci.* 100, 26 (2003), 15404–15409.
- [29] NATIONAL INSTITUTE OF GENERAL MEDICAL SCIENCES. The Protein Structure Initiative. The National Institute of General Medical Sciences, 2002. URL: <http://www.nigms.nih.gov/psi/>.
- [30] NEAL, S AND NIP, A. M. AND ZHANG, H. AND WISHART, D. S. Rapid and accurate calculation of protein <sup>1</sup>H, <sup>13</sup>C and <sup>15</sup>N chemical shifts. *J. Biomol. NMR* 26 (2003), 215–240.
- [31] PEARLMAN, D.A. AND CASE, D.A. AND CALDWELL, J.W. AND ROSS, W.S. AND CHEATHAM, T.E. AND DEBOLT, S. AND FERGUSON, D. AND SEIBEL, G. AND KOLLMAN, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structures and energies of molecules. *Comp. Phy. Comm.* 91 (1995), 1–41.
- [32] SALI, A. AND BLUNDELL, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234 (1993), 779–815.
- [33] SALI, A. AND OVERINGTON, J.P. AND JOHNSON, M.S. AND BLUNDELL, T.L. From Comparisons of Protein Sequences and Structures to Protein Modelling and Design. *Trends Biochem. Sci.* 15 (1990), 235–240.
- [34] SAUPE, A. Recent Results in the field of liquid crystals. *Angew. Chem.* 7 (1968), 97–112.
- [35] SAYLE, R., AND MILNER-WHITE, E. J. RasMol: Biomolecular graphics for all. *Trends in Biochem. Sciences (TIBS)* 20, 9 (1995), 374.
- [36] SEAVEY, B.R. AND FARR, E.A. AND WESTLER, W.M. AND MARKLEY, J.L. A Relational Database for Sequence-Specific Protein NMR Data. *J. Biom. NMR* 1 (1991), 217–236.
- [37] TJANDRA, N. AND BAX, A. Direct Measurement of Distances and Angles in Biomolecules by NMR in a Dilute Liquid Crystalline Medium. *Science* 278 (1997), 1111–1114.
- [38] TOLMAN, J. R., FLANAGAN, J. M., KENNEDY, M. A., AND PRESTEGARD, J. H. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc. Natl. Acad. Sci. USA* 92 (1995), 9279–9283.
- [39] VALAFAR, H. AND PRESTEGARD, J.H. Rapid classification of a protein fold family using a statistical analysis of dipolar couplings. *Bioinformatics* 19, 12 (2003), 1549–1555.
- [40] WÜTHRICH, K., Ed. *The Journal of Biomolecular NMR*. Kluwer Academic Publishers, Van Godewijkstraat 30, P.O. Box 17, 3300 AA Dordrecht, The Netherlands, 1997-2003.
- [41] XU, X.P AND CASE, D.A. . Automated prediction of <sup>15</sup>N, <sup>13</sup>C‘alpha’, <sup>13</sup>C‘beta’ and <sup>13</sup>C’ chemical shifts in proteins using a density functional database. *J. Biomol. NMR* 21 (2001), 321–333.
- [42] XU, Y. AND XU, D. AND CRAWFORD, O. H. AND EINSTEIN, J. R. AND SERPERSU, E. Protein Structure Determination Using Protein Threading and Sparse NMR Data. In *Proc. RECOMB (2000)*, pp. 299–307.
- [43] XU, Y. AND XU, D. AND KIM, D. AND OLMAN, V. AND RAZUMOVSKAYA, J. AND JIANG, T. . Automated Assignment of Backbone NMR Peaks using Constrained Bipartite Matching. *IEEE Computing in Science and Engineering* 4, 1 (2002), 50–62.
- [44] ZWECKSTETTER, M., AND BAX, M. Prediction of sterically induced alignment in a dilute liquid crys-

talline phase: aid to protein structure determination by NMR. *J. Am. Chem. Soc.* 122 (2000), 3791–3792.

- [45] ZWECKSTETTER, M. Determination of molecular alignment tensors without backbone resonance assignment: Aid to rapid analysis of protein-protein interactions. *J. Biomol. NMR* 27, 1 (2003), 41–56.

## A Appendix

Here, we describe, in detail, the filtering steps applied by HD prior to the application of NVR. The variable names in this section are defined in the main body of the paper. Table 3 details the contents of our comparison set.

### A.1 Filtering Steps

The first step in the HD algorithm is to apply a series of conservative filters to eliminate various structures from consideration. The first filter eliminates entries based on the length of their primary sequences. In particular we only consider models whose length is within  $\pm 30\%$  of the length of  $s(t)$ . Let  $U \subseteq M$  be the set of proteins that pass this first filter. Next, using the program HNN [15], we estimate the secondary structure of the target protein using its primary sequence,  $s(t)$ . HNN was chosen specifically because it performs secondary structure predictions *without* performing a sequence alignment to known structures. HNN makes predictions using a neural network. We note that none of our test proteins were present in the training set used to train HNN. Therefore, our algorithm does not gain an unfair advantage based on sequence similarity to known structures. The total percentages of  $\alpha$  and  $\beta$  secondary structure predicted by HNN are used in the next filter. In particular, if HNN predicts that  $s(t)$  has  $a\%$   $\alpha$ -structure and  $b\%$   $\beta$ -structure, we only consider models with  $a \pm 25\%$   $\alpha$ -structure and  $b \pm 25\%$   $\beta$ -structure. The percentages of  $\alpha$  and  $\beta$  structure for each model are determined using the program RASMOL [35]. Let  $V \subseteq U$  be the set of models that pass this filter. These first two filters, while conservative, are very effective, typically reducing the number of potential candidates to a few hundred.

Next, for each model in  $V$ , we use the homology modelling program MODELLER [32] to perform both sequence alignment between  $s(t)$  and  $s(m)$  for each  $m \in V$ , and subsequently build a backbone model for  $t$  based on the backbone structure of  $m \in V$ . In our experiments the percentage of sequence identity between  $s(t)$  and  $s(m)$  was always less than 30%. Moreover, no significant similarity was detected between the primary sequences of 2 of our test proteins (Ubiquitin and SPG) and their respective homologs using NCBI's pair-wise BLAST analysis using a threshold of 10. For the third protein (G $\alpha$ IP), a modest similarity was observed ( $E = 5 \times 10^{-8}$ ). Thus, the alignments made by MOD-

ELLER are not based on significant amounts of sequence homology. Let  $T'$  be the set of models constructed by MODELLER from  $m \in V$ . Note that each  $t' \in T'$  now has the same sequence as  $s(t)$ , and therefore the same number of amino acids. Side chains for each  $t' \in T'$  are constructed using the program MAXSPROUT [16]. MAXSPROUT considers the rotamers of each side chain and avoids steric clashes.

Next, the program RASMOL is used to compare each model  $t' \in T'$  with the secondary structure prediction made by HNN. HNN reports the prediction confidence for each amino acid position. In some cases, these confidence scores are very high. For example, for the protein ubiquitin, HNN predicts, with high confidence, that residues 24-27 are in  $\alpha$ -helix. The main  $\alpha$ -helix in ubiquitin actually spans residues 23-34. HNN's confidence in the predictions for residues 23, 28-34 are significantly lower. Indeed, all secondary structure prediction methods, have trouble predicting the exact boundaries of a given secondary structure element. Using thresholds of 83%, 78%, and 84% confidence for  $\alpha$ ,  $\beta$  and random coil, respectively, we eliminate any model  $t' \in T'$  that does not conform to these high-confidence prediction made by HNN. Let  $W \subseteq T'$  be the set of models that pass this filter.

Protons are added to each model in  $W$  using the PROTONATE module from the program AMBER [31]. Next, the protonated models are then energy-minimized using the SANDER module from the program AMBER. The models are now ready for use in the NVR assignment algorithm as described in Section 4 of the main paper.