9-3-2003

An Improved Nuclear Vector Replacement Algorithm for Nuclear Magnetic Resonance Assignment

Christopher James Langmead
Dartmouth College

Bruce Randall Donald
Dartmouth College

An Improved Nuclear Vector Replacement Algorithm for Nuclear Magnetic Resonance Assignment

Christopher James Langmead[*] **Bruce Randall Donald** [†, ‡, §, ¶, ‖]

September 3, 2003

Abstract

We report an improvement to the Nuclear Vector Replacement (NVR) algorithm [24] for high-throughput Nuclear Magnetic Resonance (NMR) resonance assignment. The new algorithm improves upon our earlier result in terms of accuracy and computational complexity. In particular, the new NVR algorithm assigns backbone resonances without error (100% accuracy) on the same test suite examined in [24], and runs in $O(n^{5/2} \log(cn))$ time where n is the number of amino acids in the primary sequence of the protein, and c is the maximum edge weight in an integer-weighted bipartite graph.

Dartmouth Computer Science Technical Report TR2004-494.
http://www.cs.dartmouth.edu/reports/abstracts/TR2004-494/

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

1 Introduction

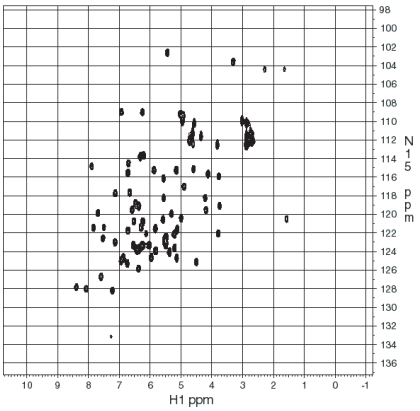
The technique of Nuclear Vector Replacement (NVR) for Nuclear Magnetic Resonance (NMR) assignment was introduced by Donald and co-workers in [25, 26], and subsequently enhanced in [24]. The algorithm discussed in [24] improves upon the one presented in [26] in terms of accuracy and the ability to handle missing data. Here, we report additional improvements to NVR which confer still higher levels of accuracy as well as an improvement in asymptotic complexity. Table 1 summarizes the accuracy and complexity of the NVR algorithm as reported in [26], [24], and the present paper.

[*]Carnegie Mellon Department 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

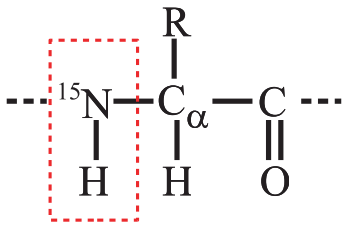
This work is supported by grants to B.R.D. from the National Institutes of Health (R01 GM-65982), and the National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, and EIA-9802068, EIA-0305444).

Reference	Accuracy	Complexity
[26]	90%	$O(nk^3 + n^3)$
[24]	99%	$O(n^3 \log(n))$
This paper	100%	$O(n^{5/2} \log(cn))$

Table 1: **Incarnations of NVR:** The evolution of NVR accuracy and computational complexity in [26], [24], and the present paper. Each incarnation was tested on the same suite of three test proteins. n is the number of amino acids in primary sequence of the protein, k is the resolution of a grid over $SO(3)$, reflecting a discrete search over 3D rotations, and c is the maximum edge weight in an integer-weighted bipartite graph



(**A**)



(**B**)

Figure 1: **HSQC**: (**A**) The *unassigned* ^{15}N-edited HSQC spectrum of the protein Ubiquitin. The peaks in this spectrum report the chemical shifts of correlated H^N (x-axis) and ^{15}N (y-axis) nuclei. (**B**) The amide group (outlined in the box) on the backbone of an amino acid. R signifies the side-chain of the protein, which differs by amino-acid type. Each backbone amide gives rise to a peak in the ^{15}N-edited HSQC spectrum.

1.1 Organization of paper

We begin, in Section 2, with a review of the relevant biology and then define the problem NVR addresses. Section 3 summarizes the NVR algorithm. In Section 4, we detail the enhancements made in the current paper and analyze the computational complexity of the new algorithm. Finally, Section 5 presents the results of applying our method on real biological NMR data.

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 (Hydrogen, Nitrogen, Carbon, 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, NVR processes the 2-dimensional ^{15}N-edited Heteronuclear Single-Quantum Coherence (HSQC) spectrum, where each peak identifies an amide (bonded H^N and ^{15}N atoms) pair (Fig. 1). 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 NVR operates. Hence, we will (re)define assignment as the mapping of peaks to residues. The goal of the NVR algorithm is to assign each resonant peak in the HSQC. In general, resonance assignments are required prior to structure determination, dynamics studies, and other applications of NMR.

NVR models the assignment problem using weighted bipartite graphs. Let K be the set of peaks in the HSQC. Let R be the set of residues in the primary sequence of the protein. 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 $|K| = |R|$.

3 Related Work

Assigned RDCs have previously been employed by a variety of structure refinement [7] and structure determination methods [18, 3, 41], including: orientation and placement of secondary structure to determine protein folds [12], pruning an homologous structural database [4, 29], *de novo* structure determination [34], in combination with a sparse set of assigned NOE's to determine the global fold [30], and a method developed by Bax and co-workers for fold determination that selects heptapeptide fragments best fitting the assigned RDC data [9]. 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 [43, 9, 36].

The idea of correlating unassigned experimentally measured RDCs with bond vector orientations from a known structure was first proposed by [2] and subsequently demonstrated in [1] who considered permutations of assignments for RNA, and [19] 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 [19], 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, $^{13}\mathrm{C}'$-$^{15}\mathrm{N}$, $^{13}\mathrm{C}'$-H^{N}, $^{13}\mathrm{C}^{\alpha}$-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 [1] 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 [19] runs in time $O(In^3)$, where $O(n^3)$ is the complexity of bipartite matching [21] 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

*In reality 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 peaks.

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 [19] is $O(k^3n^3)$. The method presented in [25, 23] 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 [25] is $O(nk^3 + n^3)$. [42] 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 [42] provides an $O(nm^6)$ tensor estimation algorithm that searches over the possible assignment permutations for the m RDCs.

The algorithm presented in [24] requires neither a search over assignment permutations, nor a rotation search over $SO(3)$. Rather, the technique of Expectation/Maximization (EM) [10] 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 $\mathbf{S}$. NVR eliminates the rotation grid-search over $SO(3)$, and hence any complexity dependency on a grid or its resolution k, running in $O(n^3 \log n)$ 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., [17, 33]). This paper improves the complexity of NVR to $O(n^{5/2} \log (cn))$.

4 Nuclear Vector Replacement

The experimental inputs to NVR are detailed in Table 2. Our algorithm computes assignments by correlating topological and geometric constraints to a given model of the protein's structure. These constraints are extracted directly from the NMR data and are converted into assignment probabilities. These assignment probabilities become the edge weights described above. We will summarize these constraints here and direct the reader to [24] for a detailed explanation of the NVR algorithm. The topological constraints are obtained from an assay for measuring amide-exchange rates and serve to identify labile, solvent-accessible amide protons. NVR uses two categories of geometric constraints, H^N-H^N NOE's (d_{NN}s) and RDCs. d_{NN}s may be observed between pairs of amide protons that are within approximately 5 Å of each other. d_{NN}s are *local* measurements. In contrast, RDCs [37, 38] provide *global* orientational restraints on internuclear bond vectors. 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* [35]. $\mathbf{S}$ is a symmetric, traceless, rank 2 tensor with 5 degrees of freedom, which describes the average substructure alignment in the dilute liquid crystalline phase. If the assignments of five or more RDCs in substructures of known geometry, $\mathbf{S}$ can be determined using singular value decomposition [28].

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 HSQC spectrum is assigned

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 Predictions	Restraints on Assignment	Tensor Determination, Resonance Assignment

Table 2: **NVR Experiment Suite:** The 5 *unassigned* NMR spectra used by NVR to perform resonance assignment. 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_{NN}s can be obtained from the NOESY. These d_{NN}s provide distance constraints between spin systems which can be correlated to the structural model. Chemical shift predictions are used as a probabilistic constraint on assignment.

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).

The NVR algorithm is divided into two phases, *Tensor Determination* and *Resonance Assignment*. 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 least 5 peaks for the purpose of determining the alignment tensors directly [28]. 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).

NVR uses EM to make assignments. EM is a statistical method for computing the maximum likelihood estimates of parameters for a generative model; it has been applied to bipartite matching problems in computer vision [8]. 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 tensor. Of course, the values of the hidden variables are unknown. Specifically, Y_G is the set of edge weights of a bipartite graph, $G = \{ K \cup R, K \times R \}$, 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, 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 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

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

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^*,

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

Then the master list of assignments is 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, and thus is guaranteed to converge in at most n iterations. In practice, the algorithm converges in about 10 iterations. The interested reader is directed to [24] for algorithmic details.

5 Algorithmic Improvements

As previously stated, NVR employs weighted bipartite graphs to represent assignment probabilities. Due to a filtering step ([24], pg. 135), these graphs are guaranteed to be sparse. That is, each vertex is connected to a constant number of other vertices. NVR repeatedly calls maximum bipartite matching as a subroutine; this subroutine is the dominant term in the time complexity of NVR. In [24], we used an implementation of the Kuhn-Munkres algorithm for maximum bipartite matching [21]. On sparse graphs, such as those used in our algorithm, Kuhn-Munkres runs in time $O(n^2 \log n)$. The maximum bipartite matching subroutine is called at most n times, yielding an $O(n^3 \log n)$ time complexity.

Alternatively, it is possible to convert the real-valued edge weights to integers by first scaling each edge weight by some constant c, and then rounding to the nearest integer. When the edge weights are integers, a different class of weighted bipartite matching algorithms can be used. In particular, the cost-scaling algorithm for assignment [13] can be used which has a time complexity of $O(n^{3/2} \log (cn))$. For sufficiently large values of c, the matching computed on the integer-weighted graph will be the same as the matching computed on the real-weighted graph, with high probability. NMR data is, in general, accurate to no more than 5-6 significant digits. Thus, setting $c = 10^7$ suffices. Once again, the maximum bipartite matching subroutine is called at most n times, yielding a $O(n^{5/2} \log cn)$ time complexity. Thus, we obtain an improvement by an $O(\sqrt{n})$ factor in the time-complexity of our algorithm.

6 Results

NVR was applied to NMR data from 3 different proteins matched to 20 trial structures (Tables 3-4). The experimental data used for our experiments are identical to those used in [24], including significant amounts of missing data (Table 5). The algorithm reported in [24] achieved an average accuracy of just over 99%. In contrast, the present version of NVR is 100% accurate on the same

		Comparison to 1D3Z	
PDB ID	Exp. Method	Sequence Identity	RMSD
1G6J [6]	NMR	100%	2.4 Å
1UBI [32]	X-ray (1.8 Å)	100%	1.3 Å
1UBQ [40]	X-ray (1.8 Å)	100%	1.4 Å
1UD7 [20]	NMR	90%	2.4 Å

PDB ID	Exp. Method	RMSD vs. 3GB1
1GB1 [16]	NMR	1.3 Å
2GB1 [16]	NMR	1.3 Å
1PGB [14]	X-ray (1.92 Å)	1.2 Å

Table 3: **Human Ubiquitin and Streptococcal Protein G (SPG).** The 4 structures of human ubiquitin used in the 4 separate trials of NVR. Both X-ray crystallography and NMR derived structures were tested. The structure 1D3Z (Cornilescu et al. 1998) is the only published structure of ubiquitin to have been refined against RDCs. The RDCs used to solve that structure have also been published and were used in each of the 4 NVR trials. 1G6J, 1UBI and 1UBQ have 100% sequence identity to 1D3Z. 1UD7 is a mutant form of human ubiquitin. As such, it demonstrates the effectiveness of NVR when the model is a close homolog of the target protein. The backbone-RMSD (all-atom) is reported for each protein relative to the 1D3Z structure. The 3 structures of SPG used in the 3 separate trials of NVR. Both x-ray crystallography and NMR derived structures were tested. The structure 3GB1 (Juszewski et al. 1999) is the only published structure of SPG to have been refined against RDCs. The RDCs used to solve that structure have also been published and were used in each of the 3 NVR trials. 1GB1, 2GB1 and 1PGB have 100% sequence identity to 3GB1. The backbone-RMSD (all-atom) is reported for each protein relative to the 3GB1 structure.

PDB ID	Exp. Method	RMSD vs. 1E8L
193L [39]	X-ray (1.3 Å)	2.1 Å
1AKI [5]	X-ray (1.5 Å)	2.1 Å
1AZF [27]	X-ray (1.8 Å)	2.1 Å
1BGI [31]	X-ray (1.7 Å)	2.1 Å
1H87 [15]	X-ray (1.7 Å)	2.1 Å
1LSC [22]	X-ray (1.7 Å)	2.2 Å
1LSE [22]	X-ray (1.7 Å)	2.2 Å

PDB ID	Exp. Method	RMSD vs. 1E8L
1LYZ [11]	X-ray (2.0 Å)	2.1 Å
2LYZ [11]	X-ray (2.0 Å)	2.1 Å
3LYZ [11]	X-ray (2.0 Å)	2.1 Å
4LYZ [11]	X-ray (2.0 Å)	2.1 Å
5LYZ [11]	X-ray (2.0 Å)	2.1 Å
6LYZ [11]	X-ray (2.0 Å)	2.1 Å

Table 4: **Hen Lysozyme** The 13 structures of hen lysozyme used in the 13 separate trials of NVR. Both x-ray crystallography and NMR derived structures were tested. The structure 1E8L (Schwalbe et al. 2001) is the only published structure of lysozyme to have been refined against RDCs. The RDCs used to solve that structure have also been published and were used in each of the 13 NVR trials. Each protein has 100% sequence identity to 1E8L. The backbone-RMSD (all-atom) is reported for each protein relative to the 1E8L structure.

set of 20 test cases on test proteins. The improved accuracy is due to small parameter changes when computing assignment probabilities from the experimental data. The new algorithm is also very fast; run-times ranged from seconds to a few minutes in our experiments.

7 Conclusion

We have described an improvement to the NVR algorithm [26, 24] for high-throughput NMR resonance assignment for a protein of known structure, or of an homologous structure. In particular, the new algorithm has an improved computational complexity ($O(n^{5/2}\log(cn))$ vs. $O(n^3\log(n))$) and improved accuracy. Resonance assignment accuracy is paramount in NMR because incorrect assignments can lead to incorrect structure determinations. We tested NVR on real NMR data from 3 proteins using 20 different alternative structures as input. Notwithstanding the fact that these NMR data sets were missing up to 19% of the (expected) experimental data (Table 5), NVR achieved 100% assignment accuracy.

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%)

Table 5: **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 HSQC spectrum should have n peaks. 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 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.

References

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

[2] AL-HASHIMI, H.M. AND PATEL, D.J. Residual dipolar couplings: Synergy between NMR and structural genomics. *J. Biomol. NMR 22*, 1 (2002), 1–8.

[3] ANDREC, M. AND DU, P. AND LEVY, R.M. Protein backbone structure determination using only residual dipolar couplings from one ordering medium. *J Biomol NMR 21*, 4 (2001), 335–347.

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

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

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

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

[8] 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.

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

[10] 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.

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

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

[13] GABOW, H. N. AND TARJAN,R. E. . Faster scaling algorithms for network problems. *Journal of Computing 18* (1989), 1013–1036.

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

[15] GIRARD, E. AND CHANTALAT, L. AND VICAT, J. AND KAHN, R. Gd-HPDO3A, a Complex to Obtain High-Phasing-Power Heavy Atom Derivatives for SAD and MAD Experiments: Results with Tetragonal Hen Egg-White Lysozyme. *Acta Crystallogr D Biol Crystallogr. 58* (2001), 1–9.

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

[17] HOCH, J., BURNS, M. M., AND REDFIELD, C. *Frontiers of NMR in Molecular Biology*. Alan R. Liss, Inc., N.Y, 1990, ch. Computer Assisted Assignment of Two-Dimensional NMR Spectra of Proteins, pp. 167–175.

[18] HUS, J.C. AND MARION, D. AND BLACKLEDGE, M. *De novo* Determination of Protein Structure by NMR using Orientational and Long-range Order Restraints. *J. Mol. Bio 298*, 5 (2000), 927–936.

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

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

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

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

[23] 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.

[24] LANGMEAD, C. J., AND DONALD, B. R. An Expectation/Maximization Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *J. Biomol. NMR. 29* (2004), 111–138.

[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] 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.

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

[28] 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.

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

[30] MUELLER, G.A. AND CHOY, W.Y. AND YANG, D. AND FORMAN-KAY, J.D. AND VENTERS, R.A. AND KAY, L.E. Global Folds of Proteins with Low Densities of NOEs Using Residual Dipolar Couplings: Application to the 370-Residue Maltodextrin-binding Protein. *J. Mol. Biol. 300* (2000), 197–212.

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

[32] RAMAGE, R. AND GREEN, J. AND MUIR, T. W. AND OGUNJOBI, O. M. AND LOVE, S. AND SHAW, K. Synthetic, structural and biological studies of the ubiquitin system: the total chemical synthesis of ubiquitin. *J. Biochem 299* (1994), 151–158.

[33] REDFIELD, C., HOCH, J., AND DOBSON, C. Chemical Shifts of Aromatic Protons in Protein NMR Spectra. *FEBS Lett. 159* (1983), 132–136.

[34] ROHL, C.A AND BAKER, D. De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta. *J. Am. Chem. Soc. 124*, 11 (2002), 2723–2729.

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

[36] TIAN, F. AND VALAFAR, H. AND PRESTEGARD, J. H. A Dipolar Coupling Based Strategy for Simultaneous Resonance Assignment and Structure Determination of Protein Backbones. *J. Am. Chem. Soc. 123* (2001), 11791–11796.

[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] VANEY, M. C. AND MAIGNAN, S. AND RIES-KAUTT, M. AND DUCRUIX, A. High-resolution structure (1.33 angstrom) of a HEW lysozyme tetragonal crystal grown in the APCF apparatus. Data and structural comparison with a crystal grown under microgravity from SpaceHab-01 mission. *Acta Crystallogr D Biol Crystallogr 52* (1996), 505–517.

[40] VIJAY-KUMAR, S. AND BUGG, C. E. AND COOK, W. J. Structure of ubiquitin refined at 1.8 A resolution. *J Mol. Biol. 194* (1987), 531–544.

[41] WEDEMEYER, W. J. AND ROHL, C. A. AND SCHERAGA, H. A. Exact solutions for chemical bond orientations from residual dipolar couplings. *J. Biom. NMR 22* (2002), 137–151.

[42] 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.

[43] ZWECKSTETTER, M. AND BAX, A. Single-step determination of protein substructures using dipolar couplings: aid to structural genomics. *J Am Chem Soc 123*, 38 (2001), 9490–1.